T.C.

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Electrical and Computer Engineering

DETECT MALWARE URL USING NAIVE BAYES ALGORITHM

Fatimah AlZUBAIDI

Master Thesis

Supervisor

Prof. Osman UCAN

Istanbul, 2021

Detect malware URL using Naive Bayes algorithm

by

Fatimah ALZUBAIDI

Electrical and Computer Engineering

Submitted to the Graduate School of Science and Engineering

in partial fulfillment of the requirements for the degree of

Master of Science

ALTINBAŞ UNIVERSITY

2021

The thesis titled “DETECT MALWARE URL USING NAIVE BAYES ALGORITHM” prepared and presented by “Fatimah ALZUBAIDI” was accepted as a Master of Science Thesis in Department.

Prof. Dr. Osman Nuri UCAN

Supervisor

Thesis Defense Jury Members:

Prof. Dr. Osman Nuri UCAN	School of Engineering and Natural Sciences, Altinbas University	_____
Asst. Prof. Abdullahi Abdu IBRAHIM	School of Engineering and Natural Sciences, Altinbas University	_____
Asst. Prof. Tariq Abed MOHAMMAD	Faculty of Engineering Jafar Al sadiq University	_____

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Approval Date of Institute of Graduate Studies: ____/____/____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Fatimah alzubaidi

DEDICATION

I would like to thank Allah Almighty for the power of mind, health, strength, guidance, knowledge and skills to complete this study.

This thesis is wholeheartedly dedicated to my father. There are no words to describe what you mean to me, and dedicated to my mother, there is nothing that I can repay for what you have done to me.

Lastly, I dedicated this to my dear husband (Ali) ,my dear sisters (Moroge, Marfa, and Hala), my lovely kids (Hussain , Asal and Abbas) and My dear brother and friend (Ahmad Edan) who was encouraging and helped me a lot during this study.

ACKNOWLEDGEMENTS

I would like to thank my supervisor professor Prof. Dr. Osman UCAN for all support during my study. It’s great pleasure to express my deepest gratitude to my friends who have shared with me best moments during my study for the Master degree.

ABSTRACT

DETECT MALICIOUS URL USING NAIVE BAYES ALGORITHM

ALZUBAIDI, Fatimah Yaseen

M.Sc, Electrical and Computer Engineering, Altinbas University

Supervisor: Prof. Dr. Osman Nuri UCAN

Date: 28/6/2021

Pages: 68

Hacking and fake pages are the basis of problems and suspicious activities on the Internet, Therefore, the disadvantages of those pages are the reason for the increased request for safeguard which prevents the user from accessing, our study explains the possibility of identifying suspicious links from the URL-based features of its addresses, we demonstrate that our problem is consistent with machine learning algorithms, it also fits to the modern features of the continuously evolving distribution of malicious URLs, we have also developed the model for those predictive addresses and categorized it into safe or unsafe URL by using naive Bayes algorithm we also compare this work with the researchers' other studies.

Keywords: Malicious, Url, Machine Learning, Naive Bayes

ÖZET

NAIVE BAYES ALGORITMASINI KULLANARAK KÖTÜ AMAÇLI URL'YI TESPIT EDIN

ALZUBAIDI, Fatımah Yaseen

Yüksek Lisans, Elektrik ve Bilgisayar Mühendisliği, Altınbaş Üniversitesi

Danışman: Prof. Dr. Osman Nuri UCAN

Tarih: 28/6/2021

Sayfalar: 68

Hacking ve sahte sayfalar, İnternet'teki sorunların ve şüpheli faaliyetlerin temelidir, Bu nedenle, bu sayfaların dezavantajları, kullanıcının erişmesini engelleyen artan koruma talebinin nedenidir, çalışmamız, şüpheli bağlantıların tespit edilme olasılığını açıklamaktadır. Adreslerinin URL tabanlı özellikleri, sorunumuzun makine öğrenme algoritmaları ile tutarlı olduğunu gösteriyoruz, aynı zamanda sürekli gelişen kötü amaçlı URL dağıtımının modern özelliklerine de uyuyor, ayrıca bu tahmine dayalı adresler için bir model geliştirdik ve kategorilere ayırdık. Naive Bayes algoritması kullanarak güvenli veya güvensiz URL'yi de bu çalışmayı araştırmacıların diğer çalışmalarıyla karşılaştırıyoruz.

Anahtar Kelimeler: Kötü Amaçlı, Url, Makine Öğrenimi, Saf Bayes

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ML	Machine Learning
PC	Personal Computer
AC	Accuracy
TF-IDF	Term frequency–inverse Document Frequency
AV	Avision (Manually fed scanner)
CNN	Convolutional Neural Net
URL	Uniform Resource Locator
WWW	World Wide Web
http	Hypertext Transfer Protocol
https	Hypertext Transfer Protocol Secure

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

We learned to avoid the seriousness of personal harm long ago; In order for us to be able to distinguish sites by harm, whether they may be harmful or safe, actually, the hacker exploits commercial sites and random advertisements in order to send their suspicious links through them. An example of this is the exploitation of fake traditional watch ads and bank fraud by advertising fake loans or by promoting cheap goods where people are attracted to this type of advertisements, which leads to their falling into Fraud networks")[1]. In the end, the main purpose of the existence of these ads regardless of the motives of their existence is to push people to go to those pages and ads and visit them to find out their content. In 2016, Symantec issued a report on Internet safety, in which it explained the existence of extensive and successive attacks on companies for the purpose of stealing information and causing great losses, the report also indicated the existence of large threats to personal and bank accounts and the threat of victims through threatening messages to pay a certain ransom through multiple methods, The victim is lured to click on suspicious links through fake and fake ads[2].

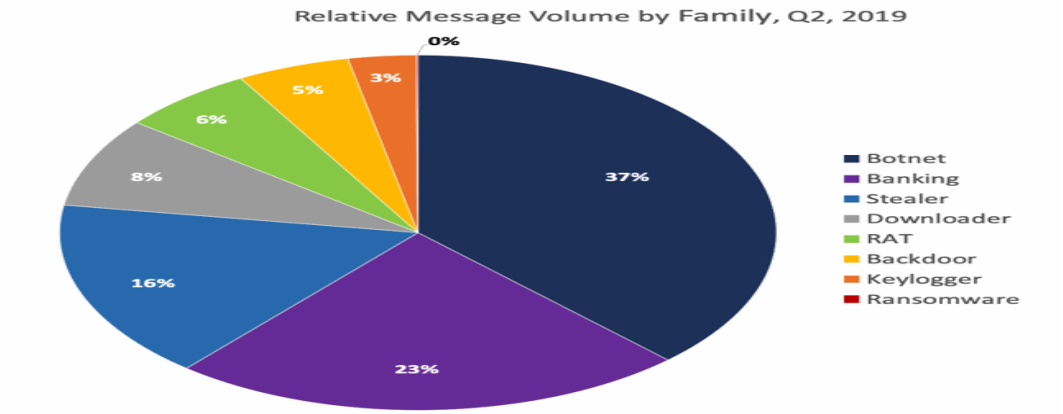


Figure 1.1: malicious attack Statistic source : zdnet.com

1.2 PROBLEM STATEMENT

Because of the large number of suspicious sites, Internet users suffer from theft of their bank accounts and passwords, which makes cyber criminals penetrate the users through the method of this malicious software, and given the large number of breaches of Internet users, information security must be strengthened and users protected from leaking their data and breaching their personal data, In this thesis, we focus on developing a model that detects harmful links and classifies those links into links to benign links or malicious links to reduce the risks that Internet users are exposed to.

1.3 MOTIVATION

The purpose of malicious programs and links is to hack polity or companies websites in order to collect the private information, or to stop their work, and in general this is done, the hacker is used to collect personal data such as names, bank accounts, personal passwords and many important details of the victim It's important to think of solutions that incorporate internet security software and technologies that can make online activities more secure, and for discover bad sites, this study examines an effectiveness of a Naive Bayes algorithm that uses URL-based features which detects links and classifies it into benign or malicious URL. We are using Naive Bayes classifiers were 80% of the dataset is used for training and 20% is used for testing.

1.4 CONTRIBUTION

It is known that combating malware is very difficult because it appears in multiple ways and spreads very quickly for that we make the following main contributions in this study:

1. We are developing a model to classify URLs into safe and unsafe using the Naive Bayes algorithm, which is a modern technology that differs from the traditional methods used previously.

2. We are conducting experiments on a large set of data and we have obtained a high accuracy of 97%.

1.5 THESIS OUTLINE

The general structure of the study consists of six chapters, including the introductory chapter:

1) In the introduction, a major view is given to the basic research idea .

(2) The theoretical dimensions of the literature review are laid out in Chapter Two, and the effectiveness of the experiment is explained.

3) In the third chapter, the methodology used for the research topic is explained.

4) The main results are presented in the fourth chapter, where the theoretical and practical information is linked in order to reach the required answer to the main question of the research topic.

(5) The fifth chapter discusses and evaluates the results in addition to examining the study limitations.

(6) Finally, The conclusion provides a summary and critique of the findings.

CHAPTER 2

LITERATURE REVIEW

The malicious links are designed to attack the victim, hack his personal account and steal his information, when the victim clicks on the suspicious link the virus is downloaded like Ransomware, The worm, and other viruses or links that cause infection of the user's device or network, the hacker locks the victim into a trap as soon as he clicks on the harmful link and thus his data and sensitive information is stolen and is exposed to various threats, so scientists called these links the term virus or suspicious link “weapon zed”, because of the severe damage it causes to the victim's apparatus or network, In fact, these bad URLs pose a lot of risk .

[Sharifi and Siadati], Blacklists are being used to overcome phishing, which is constantly increasing in size. As technology advances, innovative methods of hackers appear But these lists need constant updating in order to identify viruses and malicious links, so a new architecture has been proposed for blacklists .This new technology searches for the name of the page whose link was sent via the Internet and which claims to belong to one of the well-known companies and compares it with the names in the Google search engine. If it appears that there is a real match with the company name, the page is considered real and if the page is not found it is considered fake and classified within the lists of spam (viruses), the initial evaluation has shown good technique and accuracy in distinguishing real pages from phishing sites[3].But blacklists need constant development in order to be able to discover updated malicious links .

[Prakash et al.], The phishing network is faster to deal with basic problems when using the blacklist, although it uses broad matches but gives results when applied, the network consists of two main parts: in the first part it proposes 5 methods to enumerate a group of known fraudulent links and in the second part the algorithm splits the URL into several components and then compares it with the blacklisted entries[4].

[Hamid And Abawajy],One of the big problems that we face while using personal e-mail is the arrival of a number of spam messages, and it is possible that among those messages is a phishing message, the victim may be deceived and lured into a fake website, which leads to the planting

of malicious programs through which to spy on Personal information of the owner of the computer. To get rid of this problem, an Hybrid feature selection approach (HFS) has been proposed, whereby this algorithm predicts whether the sent e-mail is real or fictitious. The model has proven its effectiveness and an accuracy rate of 93% is obtained[5].

[Darling et al.], As a result of the increase in activities on the Internet and the increase in the types of malicious links, it was imperative to have smart systems through which malicious pages are distinguished by analyzing the URL, Previous studies have studied some lexical features, thus proposing an approach that analyzes the lexical analysis of URLs by means of a light weight approach. Which has proven its accuracy and gave appropriate results Although the workbook has proven its ability to detect URLs, it is not without obstacles, It only measures the features of long URLs and rejects the shortened URL [6]. However, the lexical analysis depends on letters, which is a process that is not fully adequate because it does not have the ability to effectively capture the semantic meaning.

[Li et al.], Malicious codes can be detected through a hybrid scheme that depends on Auto Encoder and Deep belief network (DBM) where it reduces DMB the dimensions of the data and then extracts its characteristics by DBM and the malicious code is detected and the results are verified by KDDCU99, The experience has proven to improve accuracy and reduce the time period for virus detection, but this method needs more improvement in order to give better results in the future[7].

[Gandotra et al.], Researchers focus on proposing an accurate system that detects the malware that Internet pirates seek to reach their goals, but in fact there is a very important matter which is improving the classification time in addition to the importance of accuracy in discovering links and distinguishing them into harmful and benign links, so in order to detect Zero day malware With high accuracy and taking into account the time of building the classification, a system has been proposed that uses static and dynamic features with the machine learning algorithm available in WEKA library, and the experiments have given results with high accuracy[8].

[Shibahara et al.], As technology advances, malicious attacks have increased on websites with modern piracy methods, and it has become difficult to rely on blacklists and lexical analysis.

Therefore, a modern method has been proposed through which malicious URL sequences are identified by comparing individual policy, CNN(Convolutional neural net) and EDCNN (Event de-noising) , where the latter reduces alerts False compared to CNN, The system reduces Malicious URL links, but it cannot completely remove malware[9].

[Joshi et al.], In order to prevent phishing and malware that may cause financial losses, malicious URLs are identified by blacklists, and because blacklists lack detection of modern addresses for viruses. Random forests were classified into malicious URLs through a new approach that relied on the lexical features that were used in lightweight systems, which gave quick results in real time, The model has been conveyed in the FireEye Advanced URL Identification Engine (FAUDE), bringing about a huge expansion in malignant URL detection[10].

[Khan et al.], A prototype was developed that uses machine learning methods to discover malicious URLs by using the AdaBoost (adaptive boost) algorithm , which is based on building effective compilations from many weak classifier the model gave accurate results[11].However, the use of the Adaboost algorithm is complex and costly in terms of time and computation.

[Kumi et al.], hackers take advantage of security holes in browsers in order to send their malicious links to access victims' personal or sensitive information. Extracting URL and (Java script, HTML) features from URLs using the CBA link to detect malicious links and classify them into benign and malicious is useful and gave the classification an accuracy of 95.8%[12].

From the studies and suggestions of researchers to identify and discover harmful links that can cause great harm to Internet users, whether through messages or various software applications, we always need to discover new ways to avoid viruses that are sent by internet hackers, and that these methods must be with high accuracy and appropriate speed, So we conducted a study aiming to reveal the links and classify them into harmful and benign links according to the dataset, by analyzing the URL by means of the URL-based features, and this is done using Naive Bayes prediction and prediction algorithm that have proven effective and fast in giving results with an accuracy of up to 97%.

CHAPTER 3

MATERIALS AND METHODS

3.1 MALWARE

It is an abbreviation for malicious, hostile or intrusive programs and is usually used to destroy certain data by sending malicious links that are entered through a computer network [13]. These programs use different techniques to survive for long periods, as code is added or deleted from any program in an intended way to sabotage that system or losing part of its information, these programs can appear as text ,symbol or any other form. The first intrusive programs appeared in the beginning in the form of an Internet worm as an experiment or a joke, but then it developed and became more dangerous through which personal information and bank accounts or commercial and financial websites were compromised, these programs are sometimes used against sensitive government websites to control them or to disable them[14].

The methods of distributing malware vary through:

- Social media
- Media is removable
- Websites
- Pirated software
- Email messages

Among the damages caused by these malicious programs are:

Loss of data: as Trojans and other viruses delete some files, and it will be necessary to delete these files because they will be infected files .

Account theft: The hacker can launch his attacks to steal accounts and passwords, and he has

access to the user's account files via the Internet.

Botnets: it is converting computer into a robot or zombie

Financial losses: hackers gaining access to credit card or bank information through a keyloggers program and then draining the account.

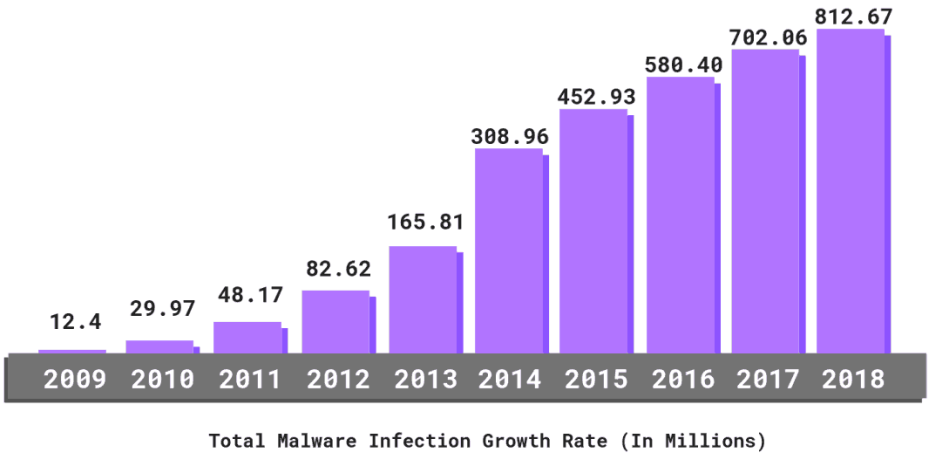


Figure 3.1: Total Malware infection growth rate

3.2 MALWARE ANALYSIS

Malware is analyzed by sample samples of viruses such as Trojans, worms and other known viruses. This analysis is considered a very important step to detect the presence of malicious links until they are removed. In the past, removal operations were traditional and performed manually, which is a tedious process and may take a long time[15].

The types of analysis for software: static analysis and dynamic analysis

- Static analysis of malware:

It is a software-only analysis process without implementation, when a specific source code is

available, the static analysis tools can find defects and validate the system[16].

The binary scan process is done without manually executing it, and the information that is compiled its source code in an executable binary file is lost and thus no further analysis can be obtained. Therefore, other techniques are used to analyze malicious programs such as:

1. File fingerprints:

It includes some operations on the file, such as an execution hash account, in order to distinguish it from the rest of the files

2. Extraction of coded strings: During which the output is printed in a binary file translated as readable text.
3. File Format: this is where important information is collected.
4. AV scan: if the detected program is malicious, it is most likely detected by AV scanners. This scanner takes a long time, but it is a necessary procedure.
5. Packet detection: malware is distributed mysteriously, for example, encrypted or compressed .
6. Disassembly: tools are used that are able to reverse the codes on the device into an assembly language such as IDAPro. This is done by debugging tools like OllyDbg.

All execution paths can be covered by comprehensive binary analysis of static malware. Since the code does not really execute it, dynamic analysis is less secure than static analysis

- Dynamic malware analysis:

This analysis is performed during runtime, where a sample of malware is executed in a controlled environment and procedures are monitored to perform dynamic analysis of malicious

links and then dismantle them[17].

One of the disadvantages of this analysis is the so-called passive code, as the analysis may not provide the entire execution path[18].

There are two main methods of analyzing malware:

First: - Analyzing the difference between certain points, where a sample of programs is executed, then the modifications are analyzed by comparison with the initial system state, and the behavior of the malicious software is clarified.

Second: - Uptime Monitor, a dedicated tool is used during runtime and malicious activities are launched by the malicious application.

3.3 TYPES OF MALWARE

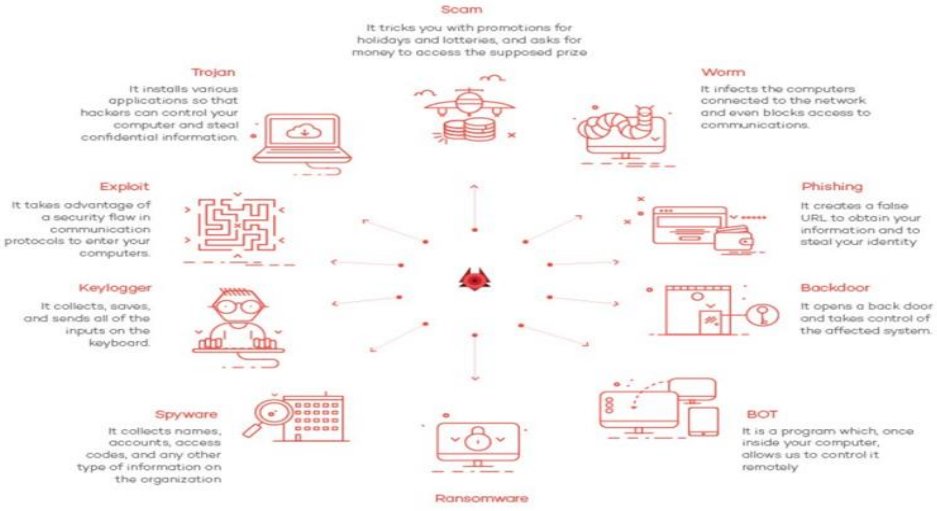


Figure 3.3: Types of malware

3.3.1 **Viruses**: It is a program that copies itself and links to files without the victim's knowledge and passes through four stages: inactivity, spread, activation, and finally the

Implementation phase. As : (Macro, Boot, Logic Bomb, Directory via Resident virus)[19].

The purpose of designing these viruses is to cause damage to the computer by destroying data or stopping the operation of the system. This type of virus is spread by downloading the Internet or by e-mail attachments, and this type of virus is used to steal information and money or to create robots networks, advertisements, etc.

The files most vulnerable to virus infection are - doc / docx, .exe, .html, .xls / .xlsx, .zip.

Virus stay inanimate till they prevalence to devices and networks.

The computer virus appeared in 1999, which is considered one of the most famous viruses, as a large number of people received an e-mail bearing the phrase (important message) and when they opened the message and clicked on the link found, this led to the infection of their computers with a virus called Melesia, where this virus stole (50) An email address from each person and then he re-sent the messages to them[20].

A computer virus is an unwanted virus that damages devices and appears in various forms through e-mail, video clips, or audio programs.

This virus differs in terms of damage, some appear in the form of an annoying program that sends random and unimportant messages, and some may be very harmful and steal information or delete files and it can also scan hard drives.In the year 2000, another virus called I Love You caused billions of dollars in loss.

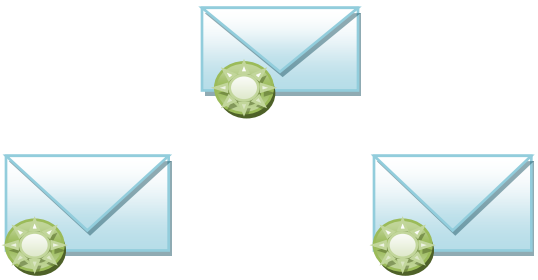


Figure 3.3.1: Computer Virus

Anti-virus programs:

When the device is infected with viruses, anti-virus programs give the alert that there is a virus, as this alert includes the names of malicious links, so the anti-virus program will isolate and remove them without the need for the user to do so[21].

After the virus is removed, the program asks the user to restart the computer and then run the anti-virus program again, and therefore in order to ensure that the virus is removed

Manual removal:

If virus removal programs cannot get rid of viruses, the following steps can be taken:

- Go to the user's anti-virus company website via the Internet.
- Giving the name of the virus that was discovered by the control software to the anti-virus software company in order for the company to discover the location of the virus on the device and delete it while keeping a copy from the operating system folder because if the non-infected files are deleted by mistake, it is impossible to operate the device.
- Finally, the computer is turned on and the anti-virus program is launched to ensure that viruses are eliminated.

3.3.2 **Trojan Horses** : The reason for its name is because it resembles the work of the legend of the Trojan horse in which the Greek soldiers hid until the city of Troy was conquered[22].

Trojan horse does not copy itself, but it is difficult to discover. It is sent by another program in the form of a file[23]. These files are usually used for logins or password.

Trojan Examples :

- Remote access Trojans (RATs)
- Backdoor Trojans (backdoors)
- IRC Trojans (IRCbots)
- Keylogging Trojans.

Trojans can be classified according to the procedure it performs on a computer:-

- Infiltration: Trojans provide the hacker with the ability to remotely control the affected computers, so the hacker can achieve a strike on the infected computer, such as sending certain data or deleting messages, operating a computer.
- Security deactivation viruses: These programs take advantage of any security vulnerabilities that exist in the applications running on the computer.
- Rootkits: They are placed in the computer in a hidden way so that the user cannot discover them.
- Banking Trojans: These programs are designed to steal online bank account data or credit cards.
- Distributed service blocking Trojans(DDos): in which multiple requests are sent in order to block the DDos service from a web address to be targeted, and thus the service is blocked.
- Trojan Downloader: These programs can download new versions of malware and install them on a computer.
- Trojan horse carrier: It is used to install malware as virus programs cannot scan all components for this type of virus.

- Simulated Trojans: In which the victim is threatened and blackmailed to pay money with threats that do not exist in reality.
- Trojans to steal games: Accounts are stolen by users' information for online games.
- Trojans for instant messaging: Information is stolen during instant messaging.
- Trojans for ransomware: The hacker blocks the use of programs on the personal computer and does not return it except by paying the ransom
- SMS Trojans: Messages are sent from the victim's device to other numbers at an expensive price.
- Spy Trojans: The user’s computer steps are tracked by taking screenshots or tracking the data entered by the user through keyboards.
- Trojans to find mail: These programs can find e-mails on a computer.

3.3.3 **Worms**: It is a computer program that copies itself and causes damage to computers,It uses the law of exponential growth, which causes many infect within a short period of time[24].

Example of Worms:

- Mail worm
- Immedate message worm
- IRC worm
- Worm files share software
- Internet worms.

Worms are spread through harmful e-mails in a manner that attracts the reader as an

invitation to view the pictures of a celebrity, or an artist, and care must be taken even in known sources because worms have the ability to send themselves from stored addresses to any e-mail, and it is necessary to update systems constantly to avoid worms[25].

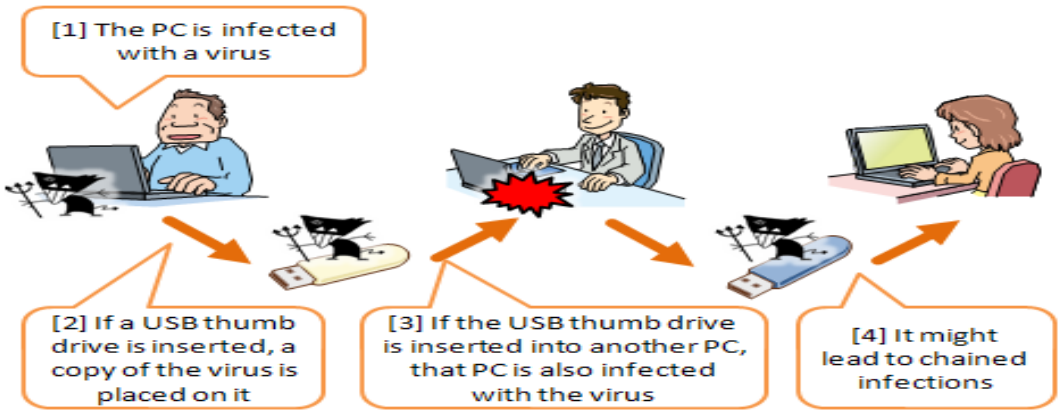


Figure 3.3.3 Computer Worms

3.3.4 **Spyware**: These are could be Troublesome programs which infiltrate User's device, stealing network use datum via sensitive inputs. Its categorize among the bad viruses because it causes damage to data and devices, Secretly In such a way that the victim does not feel the presence of the virus ,Popular spy corporation Consisting of:

Gator, Bonzi Buddy, 180 Solutions, DirectRevenue, Cydoor, CoolWebSearch, Xupiter, XXXDial and Euniverse[26].

Figure 3.3.4: Spyware

3.3.5 **Phishing**: Electronic messages, mobile messages and other means are the main gateway for the hacker to enter into unique accounts and corporate accounts, and thus the device is hacked with its data, where the hacker claims that he is the owner of an important and well-known company in order to lure the victim in order to give his sensitive information such as the password for a credit card, bank account, or The user's passwords and identification card, and thus the user's data is hacked [27].

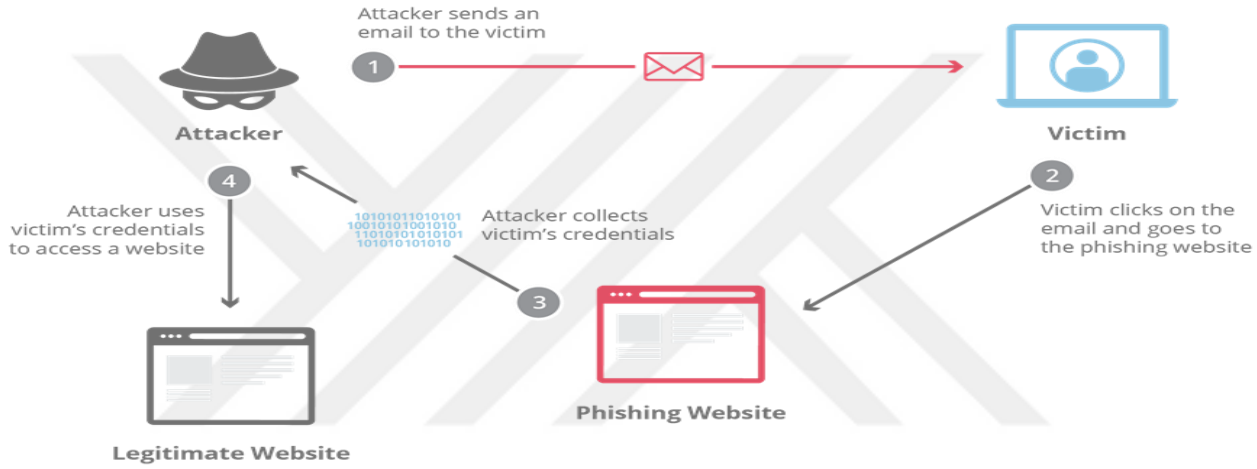


Figure 3.3.5: Phishing attacks

Types of Phishing :

a) Email phishing

b) Spear phishing

c) Whaling and CEO fraud

d) Clone phishing

e) Voice phishing

f) SMS phishing

To avoid phishing the following steps should be taken:

- It is necessary to be aware of new phishing techniques constantly because there is a continuous development in phishing methods, so without following developments it becomes very easy to fall into the trap of a hacker. Whenever there is follow-up and caution, the risk of phishing will decrease, it is also advised that security awareness training by IT officials is necessary[28].

- Do not click on links directly until after you are sure of the authenticity of their source. Think carefully before clicking on links that appear in spam emails, sometimes these links lead to sites that appear similar to the real websites, and there may be a request to fill out a form with the user's information. Most of the time, phishing messages begin with (dear customer), so caution is very important when dealing with these random links and not clicking on them until after confirming their source.

- By installing anti-phishing toolbars, a quick scan of the sites visited by the user occurs and compared with the phishing lists. If a malicious site is found, the toolbar will give the user an alert, which is a free process to increase protection against phishing attacks.

- You should take great care when disclosing financial information over the Internet. The more secure the user is on the website, the less phishing will be. One of the important things that must be focused on is to make sure that the website's URL starts with (https), in addition to the presence of a closed lock icon near the address bar, and you must ensure the safety of the site being used , If an alert is given to the user via a message indicating that the site is likely to contain harmful files, then the site should not be opened and the files should not be downloaded, and it is also necessary to not deal with low-cost purchase sites because it may lead to access to the user's credit card by phishers .

- You should visit personal accounts on the Internet regularly because leaving the account for long periods of time may expose it to hackers. It is necessary to change the passwords regularly to prevent any bank fraud or any fraud related to personal credit cards. Monthly

statements of bank accounts must be obtained to avoid any fraudulent transactions without the real user's knowledge.

- Phishers exploit vulnerabilities from time to time, so the user's browser must be updated, downloaded, and installed frequently.

- Firewalls act as a barrier between the user and the PC and intruders. There are two types of firewalls that must be used together to increase protection against hackers, namely, the desktop firewall and the network firewall.

- Sometimes phishing comes in the form of pop-up windows, so you should repeat clicking on the cancel button on those windows and it is better to click (X) in the upper corner of the window in order to close it.

- You should not share personal information or financial accounts over the Internet, and it is better to go to the website of the company in question and communicate with them through their official phone numbers or emails.

- Anti-virus programs are very necessary, but they need constant updating. These programs must be used with firewalls to prevent phishing attacks. These programs scan every file that comes through the Internet, thus protecting the system from damage.

- Adherence to these advices by setting up protection programs and continuous updating of these programs and dealing with caution with all suspicious links and random messages and not giving out personal information and bank accounts except after verifying the validity of the links, programs and messages, all of these things reduce the risk of being exposed to phishing scammers and falling into a phishing trap.

3.3.6 **Spam:** Mail attack is usually by sending many random links that belong to unreal websites that do not exist in an authorized way, so that the user clicks on the suspicious link, which leads to the virus entering the user's device and damaging or stealing his data [29].

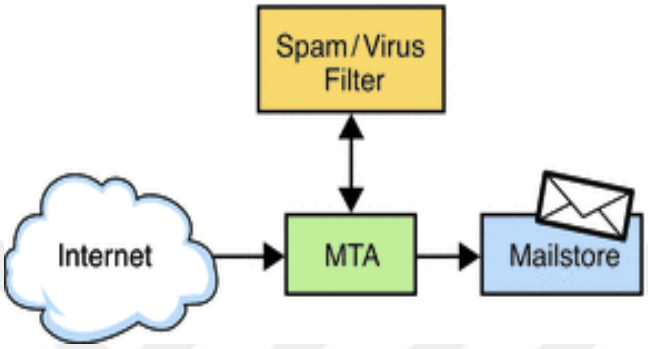


Figure 3.3.6: Spam attack

3.3.7 **Adware:** It is a malware that appears on the user's screen in the form of ads, in a frequent and annoying manner [30]. Among the malicious methods by which these harmful viruses appear is that they show that their sources are reliable in order to deceive the victim into installing those programs on his personal computer.

Figure 3.3.7: Adware Virus

Type of Adware:

- Weight loss ads where the victim is deceived into having magical ways to treat obesity and have a lean body within a very short period of time.
- Profit announcements and winning huge sums of money and making the victim under the illusion of obtaining a real sum of money, so he is drawn into giving his real information and thus falls into the trap of the hacker.
- False warning messages that make the user click on them to hack them.
- Pornographic and immoral advertisements that attract teenagers when browsing on the Internet and make them trap.
- An advertisement that makes the user enter another browser without his will.

There are two ways to make Adware can get onto devices:

- Through free software:

The adware program is installed by visiting an infected website, and here the piracy process takes place, because hacking programs are included through free programs or through advertisements.

- Infected sites:

Visit any infected website, as this may install Adware on the device, to protect against adware PUA and malware .Advertisements should be avoided as much as possible.Taking care of the personal maintenance of your computer is very important, through:

1- Continuing to update the programs

Because developers are constantly discovering and fixing security holes

2- Dealing with extreme caution with every new link

Because hackers are trying to exploit others and gain access to their personal information through programs and advertisements.

3- Do not accept any link until after confirming its source, because hackers are trying to imitate URL links.

4- Keep using legal activities because illegal and authorized links may get you into a hacker trap.

3.3.8 **Ransomware**: It consists of links or malicious programs that ask the victim to send a ransom in exchange for his system to work again and represents another criminal means to collect money from the victims, and one of the methods used by this virus is that it closes the user's screen or his important files and does not decode the code until after the ransom is paid[31].

It uses scare tactics or intimidation to trick victims into paying up. It can come in the form of fake antivirus software in which a message suddenly appears claiming your computer has various issues and an online payment is necessary to fix them!

Ransomwares types :

There are three types of ransomware :

- Scareware

There are no really scary programs. Rather, it is a fraud and fraud process in which threatening messages are sent to the victim informing him that if a certain amount of money is not paid, the device will not get rid of malware, and in the event of non-payment, the victim continues to receive threatening messages on an ongoing basis.

Therefore, having security programs installed on the devices, the user will not have to pay the ransom because he has already paid money for these protection programs in exchange for completing this task.

- Screen lockers

The computer screen is locked suddenly as the user cannot use his personal computer completely.

This means that the computer has been exposed to the ransomware virus, and the screen will be stamped with a stamp from the Department of Justice or the FBI.

Where a message appears on the screen claiming that there is illegal activity on the user's personal computer, and he must pay a fine in order to be able to use his computer again..

- Encrypting ransomware

It is considered one of the most harmful and frightening types because the hacker steals the user's personal files and encrypts them, then demands money in exchange for deciphering the file and then returns it to the user.

The reason for the danger of this type is that there are no other ways to bring files other than paying money, because the hacker does not actually return the user's encrypted files until after payment, and the user may pay the ransom, yet he cannot get his encrypted files.

Types of modern attacks using ransomware:

- Europol: Ransomware takes the first place in the IOCTA Report
- Ransomware constantly attacks cities and businesses
- In 2018 and 2019, ransomware and Trojans have especially dominated education

Ransomware for cell phones:

Ransomware spread for cell phones after the Cyptol locker malware emerged in 2014, Where the user of the mobile phone receives a message that the device has been locked due to the detection of illegal activity, and the lock will not be opened until after the fine is paid, so the mobile phone must be switched on in a safe mode and delete all malicious applications that are hacked in order to restore the use of the mobile phone[32].

Ransomware on a Mac computer

Key Ranger is the first ransom program on Mac computers that was discovered and published in 2016, and it is a malicious program that entered an application called Transmission, where this program makes many copies of harmful files in the background in a secret way and works for three days until it turns into a dangerous virus that encrypts Files.

However, after a short period of time, a built-in anti-virus update was released from Xprotect, Apple, which prevented viruses from penetrating systems and computers.

Targeted ransomware:

At the beginning of the emergence of viruses, cyber criminals and pirates were targeting individuals and personal devices, but after that they developed their goals to target companies, as malicious ransomware programs were able to achieve great success by targeting companies, which led to a decrease in production and the loss of a lot of important data.

In 2016 the percentage of malicious ransomware reached In companies, about 12.3% of malicious software, at the same time, its percentage in individual devices reached nearly 1.8%. but in 2017, companies were exposed to a massive attack of ransomware, which reached 35%.

Procedures that must be adhered to face cases of ransomware penetration:

The ransom should not be paid when personal computers and Internet networks are hacked, and

this is what the Federal Bureau of Investigation advises, because giving the ransom in exchange for decoding the codes encourages Internet hackers to continue their attacks on individuals and companies, and some damaged files can be decrypted through the free decryption program.

Ransomware, in fact, uses advanced algorithms, so it is difficult to decipher some damaged files, and therefore individuals or company owners may pay the ransom and try to decrypt their files encrypted with wrong decryption programs, which is why the problem is further complicated. Therefore, you should pay attention to this matter and seek advice from a specialist in this field before taking any step.

There are many ways to avoid malicious ransomware, including downloading protection programs that may not return the damaged files or decrypt them, but at least they detect the virus and remove it from the computer. As for encrypting the computer screen, the best solution to get rid of this problem is to use a disk drive or Full system restore.

When a computer user feels that there is a slowdown in his computer, he must stop spreading encrypted ransomware and separate it from the network, and in the event that the user runs encrypted ransomware programs once and the malware is active, this means that it will not be able to send and receive information and obtain the ransom, so It is better to download security programs and a checking program that is constantly updated.

Figure 3.3.8: Ransomware Virus

3.4 DETECT OF MALWARE

Warning signs that your computer has been infected with malware :

- Computer becomes slow
- Annoying and unwanted ads are seen
- Malfunctions
- Pop-up messages appear
- The Internet is increasing its speed in a suspicious way
- The main page of the browser changes suddenly
- Unexpected messages appear
- Disable security
- Your friends tell you that strange messages have arrived from you
- Seeing strange icons on the desktop
- Strange error messages
- Difficulty accessing the control panel
- Computer looks perfect
- There is a bug in the browser
- Suspicious shortcut files exist

3.5 SOCIAL ENGINEERING

They are methods and techniques used to deceive people and lure them into disclosing their personal and confidential information.

And it is either by phone or e-mail, where the attackers impersonate a well-known and important figure in the community or in the form of a beautiful girl on social networking sites, and some questions are asked without making the victim feel any doubts.

An example of some tricks used in social engineering. The attackers use a site they claim to be a site for registering a lottery for immigration to Canada. The site asks to enter personal data and thus the victim is hunted.

Social engineering work mechanism :

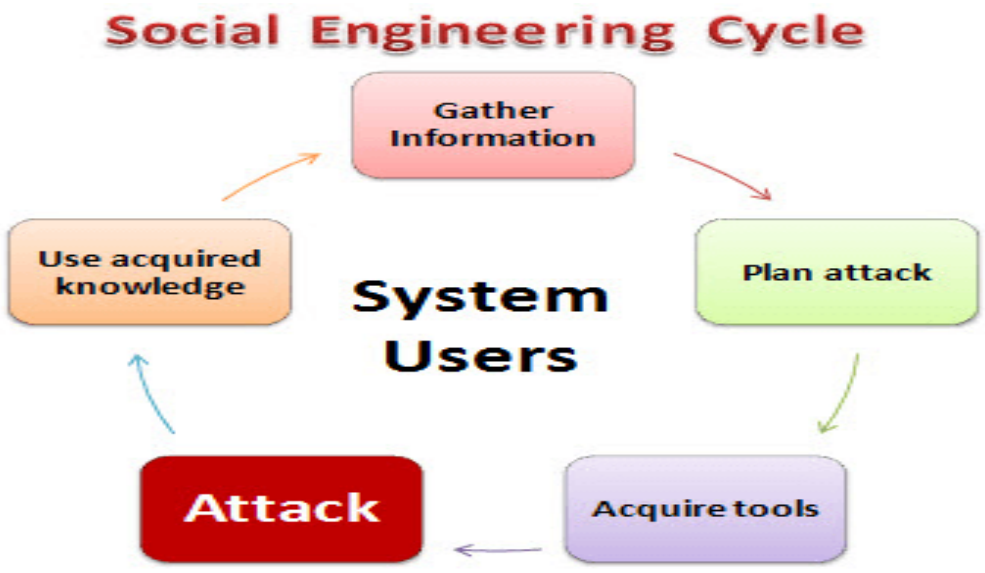


Figure 3.5: Social Engineering Cycle

1. Gather Information: In the first plan, important information about the victim is collected.
2. Plan Attack: Establish a specific plan of attack.
3. Acquire Tools: The type of program through which the victim will be attacked.
4. Attack: Knowing the weaknesses of the system that will be targeted
5. Use acquired knowledge: The password is guessed by the information gathered about the victim.

There are several forms that social engineering techniques can take, such as:

- Familiarity Exploit
- Intimidating Circumstances
- Phishing
- Tailgating
- Exploiting human curiosity
- Exploiting human greed

3.6 URL

It is an abbreviation for the word Uniform Resource Locator and it represents a basic part of the resource identifier in which websites are identified and written in the address bar[33].

before writing a URL the protocol is written like http: //

URL consists of

- Protocol available

- Domain name or page location
- Type of extension

Types of extension for web pages:

The page domain represents the name of the site and it indicates the intended address. The page domain is converted to special internationally recognized numbers such as:

- edu refers to an educational site
- org means that the page belongs to an organization
- gov is a government website
- Com stands for commercial site

Sometimes some countries put two letters to indicate that the site belongs to them, such as UK, which concerns England.

URL infrastructure:

- (Network Protocol): It is the access road to the main resource.
- (Access authorization): Network access permission.
- (host): includes the website address and the information regarding the website's domain names (IP)
- (Port) It is mandatory with (IP) address.
- (Track): gives information on how to obtain entry permits.
- (Premeter): Measures the information for the main network resource.

Figure 3.6 : Uniform Resource Locator

CHAPTER 4

PROPOSED METHOD

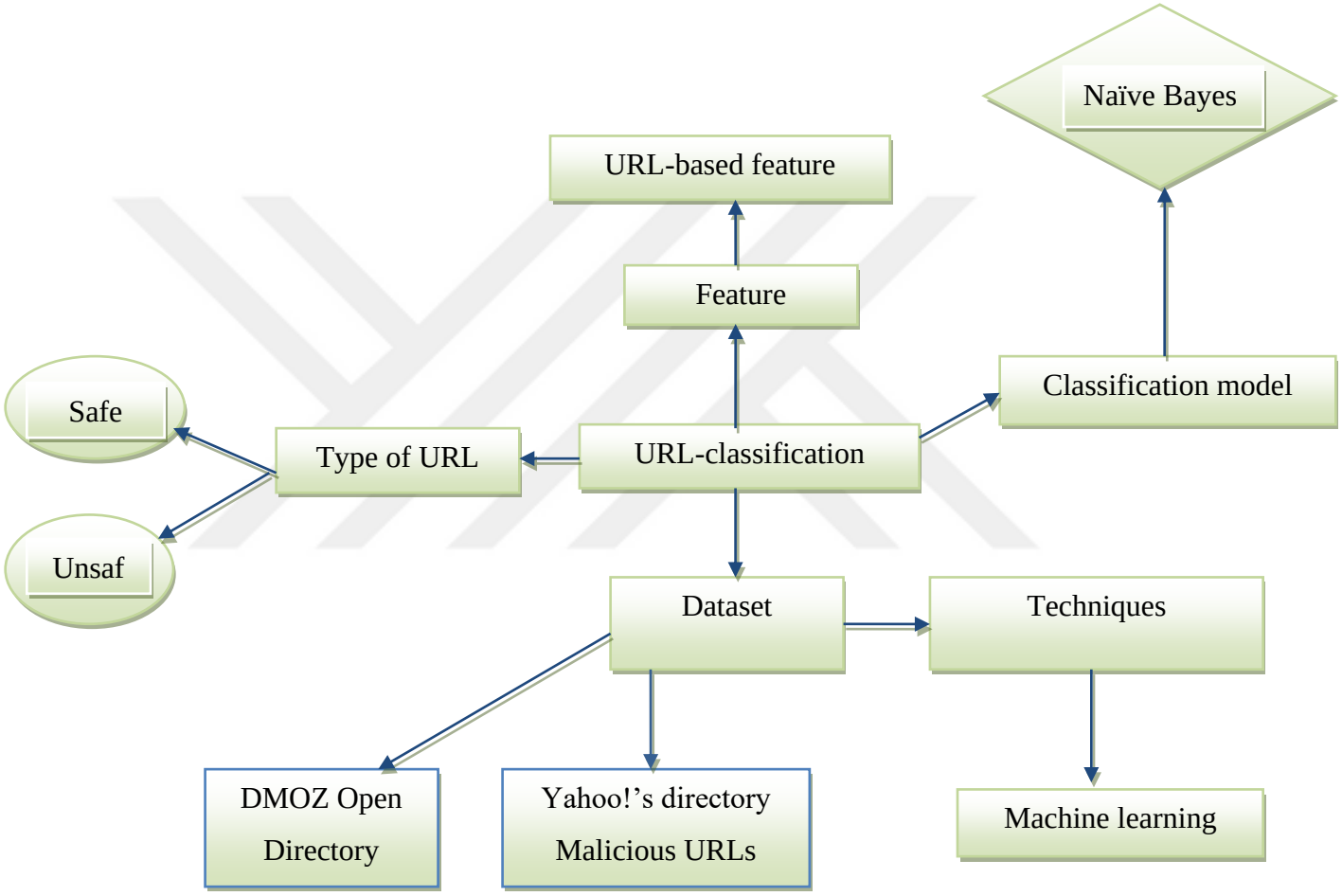


Figure (4) : URL Classification

4.1 MACHINE LEARNING

It is symbolized by the symbol (ML) and it is one of the branches of artificial intelligence (AI) that gives machines and devices the ability to learn without being programmed, and that is through algorithms that have the ability to self-learning as well as predict data because these algorithms build a model of inputs to take resolution instead of their commitment to specific codes by humans, the term machine learning appeared in 1959 in IBM labs by the pioneer of artificial intelligence(Arthur Samuel)[34].

There are two types of learning levels, one of them (inductive); this type uses big data to deduce from it the important rules and necessary provisions. And the other is called (the deductive), which uses general provisions and applies them with special examples.

In machine learning, a good algorithm gives very accurate and precise results because it deals with massive amounts of data and processes it.Researchers in this field seek to develop the capabilities of this algorithm in order to be able to give the required results.

Machine learning does not depend on human intervention except in a very small percentage, and one of the most popular practical applications of artificial intelligence based on the principle of machine learning is the robot Sophia, Artificial intelligence represents the top of the pyramid, which includes machine learning and then collects and extracts the required data.

The importance of machine learning lies in the following:

- Dealing with massive amounts of data
- Computer processing data is less than labor cost.
- Analyze data with high accuracy.
- Get accurate and fast results.
- Achieving high profits for establishments and organizations.
- Machine learning helps to make the best decision if there are many options.

To create a machine learning model, there are four important steps that have been implemented by data scientists with specialists in developing this model:

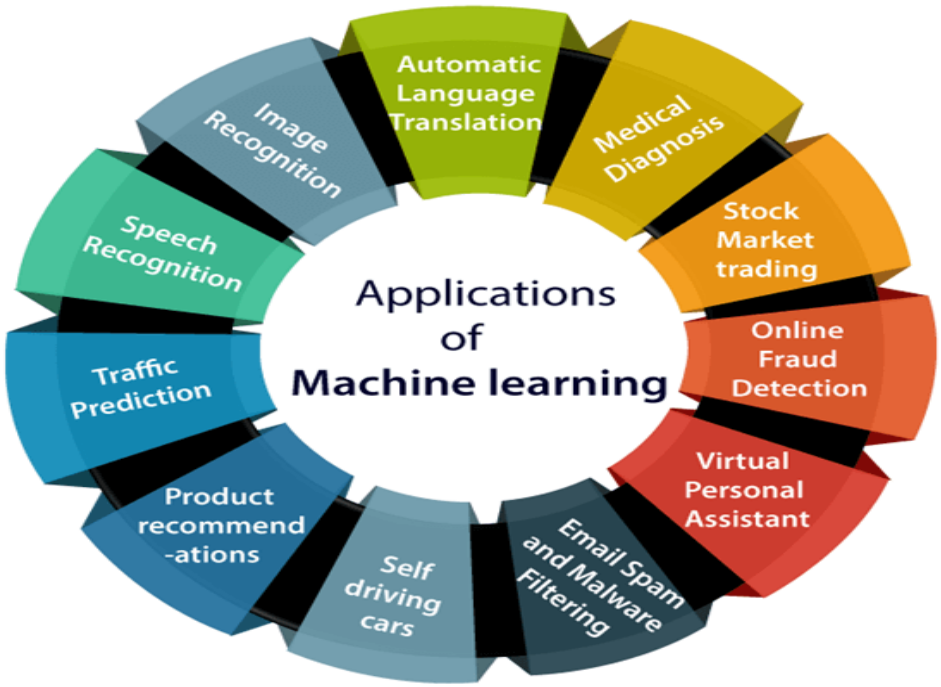


Figure 4.1: Application of Machine Learning

First, the data required for the purpose of training is determined. These data represent the inputs that the machine learning model will depend on in order to find solutions to the specific problem this model for solving.

Where in the first step, that data is marked in order to summon all the required information from the features and others that the model needs to be specified, and there are data that do not have a specific name that are classified on their own, Whether the data is specific or not, its data must be prepared in an appropriate and random manner, removing all details that would affect the progress of the process, and it must also be divided into two branches. The first is for training and the second depends on evaluation.

Second, in this step, the appropriate algorithm is chosen, which represents one of the basic steps of statistical treatment. As for the type of this algorithm, it depends on the amount of data allocated to solving the problem to which the solution is to be found, and the type of algorithm depends on the data that does not have a specific name or has no classification.

There are types of algorithms that are used in machine learning with categorized data, namely, regression algorithms, such as linear regression and logistic regression.

Where linear regression performs the prediction process, and depending on the value of the independent variable, it can determine the dependent variable, while in this case can use if variables consisting of two types, for example, the sales volume for the current year can be predicted Based on previous year's sales results , As for decision trees, they provide the required recommendations based on the type of data available

Third, training the algorithm in order to design the appropriate model is in fact an iterative process where that algorithm checks the data and runs the variables and then compares the outputs and the type of results that it was supposed to produce, and the required things are adjusted within the algorithm in order to give very accurate results. The algorithm makes it give correct results most of the time.

Finally, new data is dealt with by using the template that was designed for this purpose in order to develop its functions and accuracy, the existing problem, for which the solutions are also supposed, depends in fact on the type of data used, for example, to design a model that identifies spam messages. This model will absorb spam messages, or if this model is dedicated to finding malicious links, it will deal with viruses and specify its type and many other examples .

4.2 MACHINE LEARNING CATEGRIZE

4.2.1 Supervised Machine Learning

After determining the data that a model is to be created to classify, this model for machine learning must be subject to training on that data, and it really needs less training data than other machine learning methods, and the reason is due to making the comparison process easier with the data or Results that have already been rated.

One of the important things must be attentive to the compatibility between the model and the training data, so that the model does not lose its ability to deal effectively and accurately with the differences that may exist in the new data..

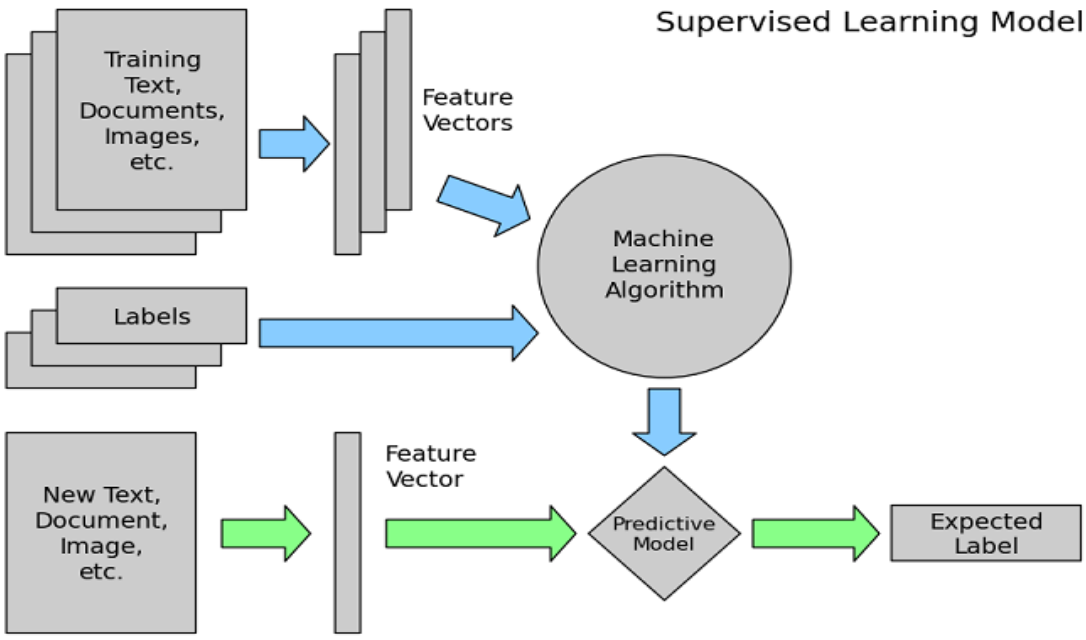


Figure 4.2.1: Supervised Learning Model

4.2.2 Unsupervised Machine Learning

This type of machine learning is not concerned with predicting or making decisions but rather determining the pattern of the data.

Unsupervised machine learning collects a lot of data that is not classified and has no specific names. Here comes the role of the algorithm to determine the characteristics and features of this data and then classify it, as this algorithm can deal with huge amounts of data.

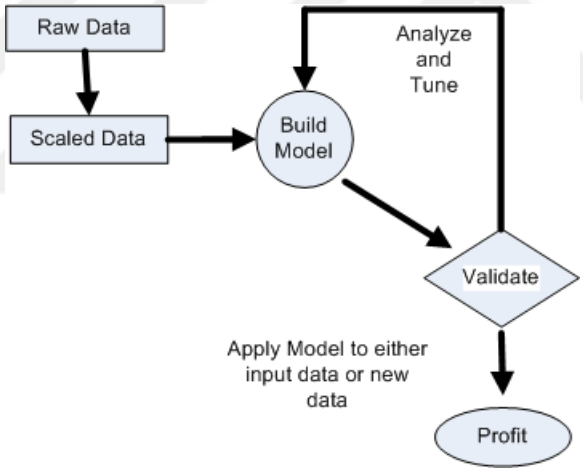


Figure (4.2.2): Unsupervised machine learning

4.2.3 Semi-Supervised Learning

It is a mediator in the midst supervised and unsupervised learning, the importance of its role being its ability to identify known and previously identified data and classifies it, and then extract features from the larger data set that has not yet been named.

This type of classification deals with the important problem of not having enough disaggregated data for the supervised algorithm to be trained.

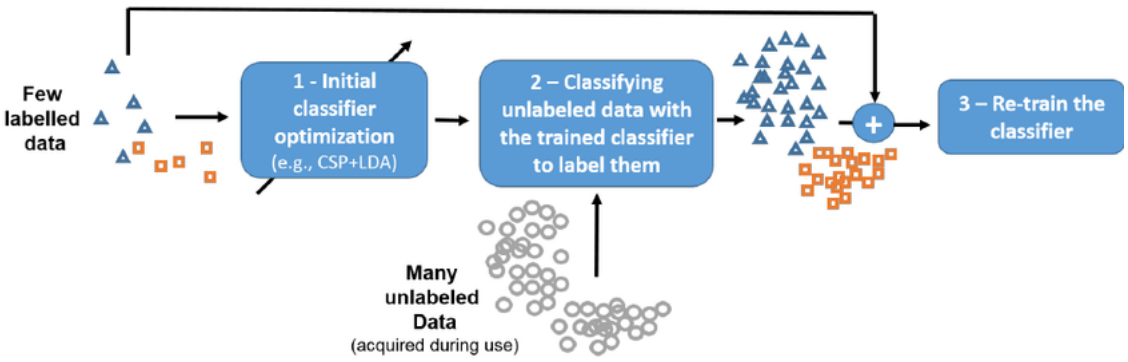


Figure (4.2.3): Semi-supervised learning

4.2.4 Reinforcement Machine Learning

This is as supervise ML, but variation between them is : algorithm in this type is not trained on data. The principle of its operation depends on trial and error.

The IBM Watson® system that won the Jeopardy is the best example of how Reinforcement machine learning works in which a wager is set to determine the answer to a question or ask a question.

Figure 4.2.4: Reinforcement machine learning

4.3 CATEGORIZE IN MACHINE LEARNING

Categorize: operation of classify certain data into groups, and that process can be carried out on organized or random data, where a category of certain data points is initially predicted. The main goal of predictive modeling of classification is to determine which category will house the new data .Classification Terminologies In Machine Learning :

- Classifier: It is an algorithm for mapping input data to a specific class.
- Classification Model: It will deduce or predict the class or category of data.
- Feature: It is an individual characteristic that measures the observed phenomenon.
- Binary Classification: It is a result: either true or false, for example.
- Multi-Class Classification: And it is multiple varieties each sample is assigned to one label only.
- Multi-label Classification: During which each sample is assigned to a group of labels.
- Initialize: It is to assign the class.
- Train the Classifier: The trapping method (X, y) is used to fit the model to train the X train and to name the y train.
- Predict the Target: The prediction method (X) returns the expected denomination (y) in the case of the unlabeled observation (X).
- Evaluate: means the evaluation of the model like accuracy , classification .

Classification Algorithms: Classification in machine learning represents the idea of supervised learning as it classifies groups into classes, and there are algorithms used for that classification, including:

4.3.1 Logistic Regression :

It is an algorithm used in machine learning that uses one or more independent variables to give the result, it can only have two results.

The goal of using this algorithm is to find a relationship between the dependent variable with the set of independent variables, This algorithm only works when the assumed variable is binary and the data has no missing values, which is a major flaw of the logistic regression algorithm[35].

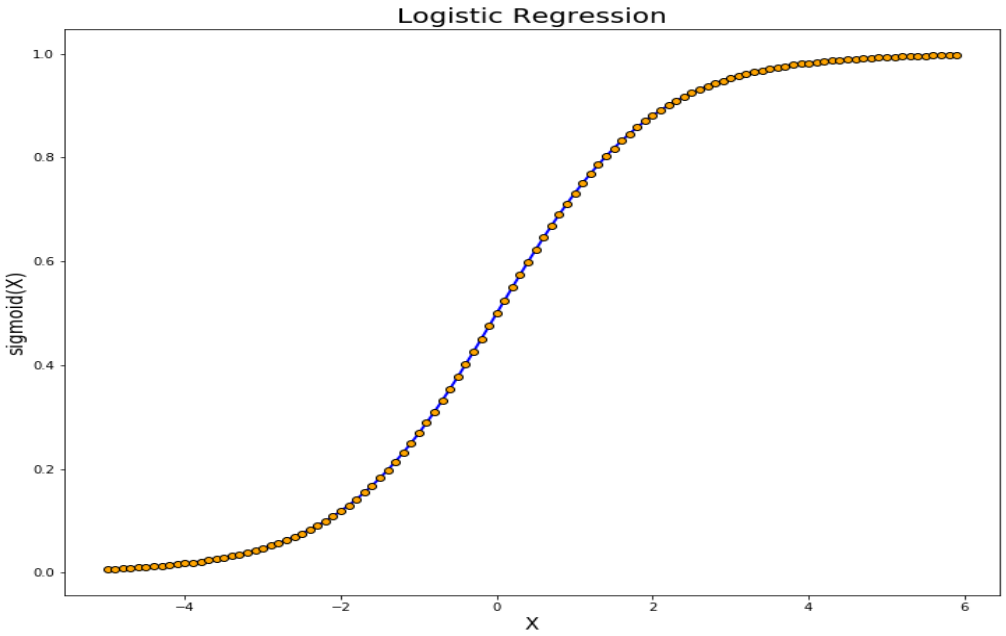


Figure 4.3.1: Logistic Regression

The three types of regressions are:

Binary regressions

Multi regressions

Ordinal regressions

4.3.2 Naive Bayes

It is very important algorithms in ML, information analysis and classification of categories, data analysis and classification of categories, as it is characterized by its speed and accuracy in predicting processes[36]. This method is based on Bayes 'theory, as this theory appeared for the first time in a paper published in 1763 by the British scientist Thomas Bayes. This algorithm calculates the probability of a given outcome and assumes that the traits and features are independent of each other.

This model (Bayes) is distinguished by its superiority over other algorithms as it is characterized by its simplicity of construction and can be developed continuously as it processes big data .

The model is trained according to the information available in the database, then the form determines the type of new records and classifies them according to the information previously available

This method is used in different systems, such as recognizing the features of a specific face in pictures or determining the content of the text (negative, positive, optimistic) through analyzing feelings for that content and other uses

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

using Bayesian probability terminology, the above equation can be written as

$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$

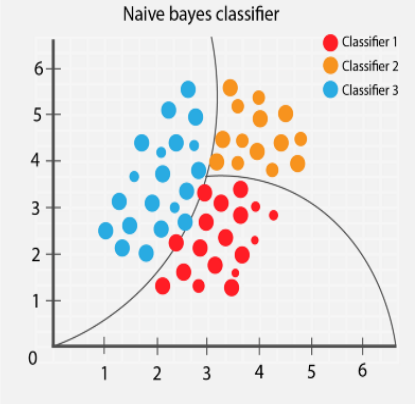


Figure (4.3.2a): Naive Bayes

There are 3 main types of Naive Bayes algorithms:

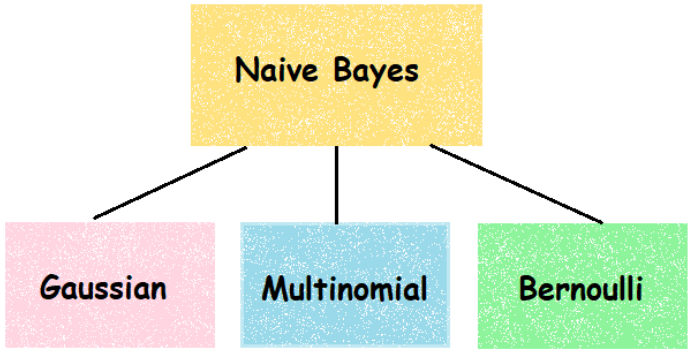


Figure (4.3.2b): types of Naive Bayes algorithms

•Multinomial Naive Bayes:

It is used to solve document problems and classify them according to the file's belonging to a specific category, and it depends on the principle of repeating words within the document[37] .

•Bernoulli Naive Bayes:

They are similar to Naive Bayes, but they are logical variables that predict a specific value, for example only (yes or no).

•Gaussian Naive Bayes:

This special type of algorithm is used in the case of continuous values and features tracking normal redistribution. This algorithm is used as the basis for complex machine learning algorithms for classification and regression[38].

The main advantages of Naive Bayes:

1. The Naive Bayes model emerged from traditional numerical hypothesis and has a steady grouping proficiency.

2. This algorithm does not need huge data, as it deals with even small data, as well as deals with the required tasks in an incremental manner.

3 This algorithm is used for the purpose of categorizing symbols and is not affected when data is lost, and Naive Bayes is a simple algorithm.

The main disadvantages of Naive Bayes:

1) The Naive Bayes model considers that the characteristics of the data it processes are mostly independent of each other for a specific class of output because this model contains the lowest error rate if compared with other classification methods.

The smaller the correlation between the characteristics of the data arrangement, the more accurate the Naive Bayes algorithm should be .

2) Naive Bayes algorithm relies on previous and default predictions, and without going back to the previous hypothetical model, the algorithm will give bad prediction results.

3) There is an error rate using the Naive Bayes algorithm due to the reliance on previous results

4) The algorithm is very sensitive to the expression of the data input.

4.4 SUGGESTION MODULE

The suggestion module are implemented as follows:

4.4.1 Training Data:

In the beginning, the training data is collected like figure (4.4.1), this datum represents URL of the recognize types meaning that their domain names of domains via class are awarded in the training group.

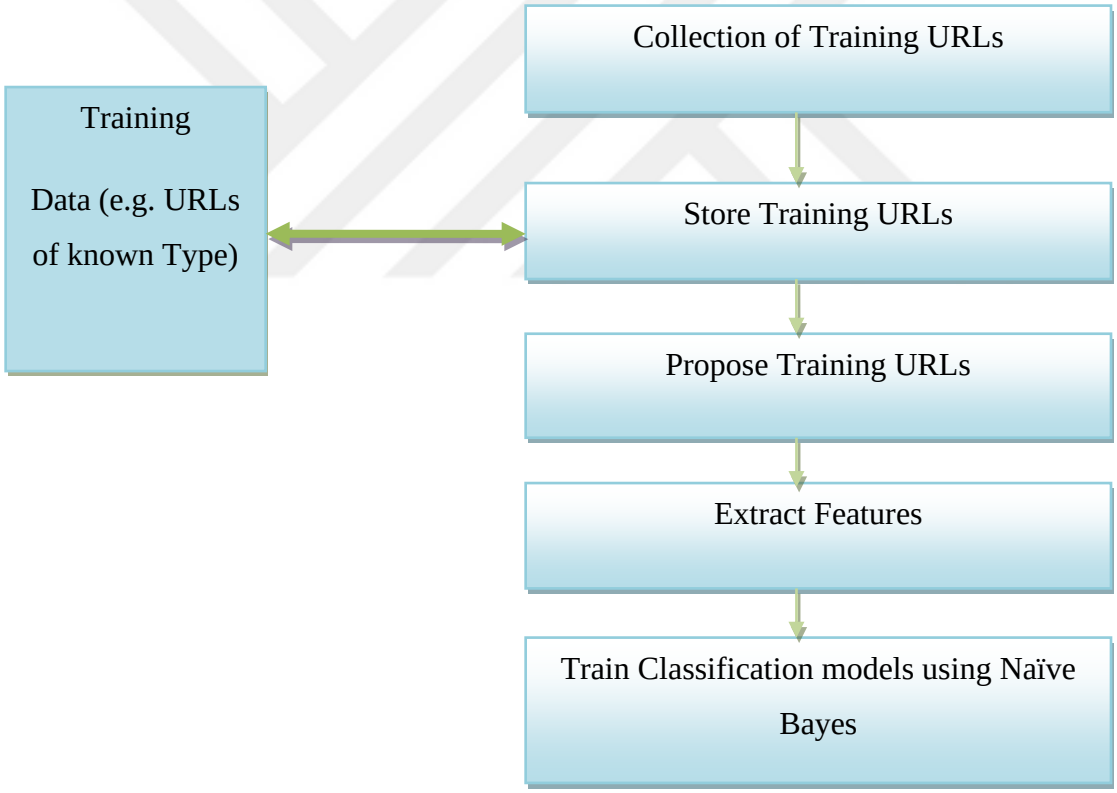


Figure 4.4.1: Training Data

4.4.2 Feature Extraction Module

Identifying features is an important step to obtain accuracy and efficiency in ML, the biggest trouble facing machine learning and the main obstacle in its path is that the algorithm deals with numbers and natural language, so these texts must be converted into numbers to be dealt with in data analysis[39].

Therefore, the model or algorithm must be chosen with high accuracy in order to obtain the desired results.

TF-IDF can be fed to algorithms such as Naive Bayes algorithm which will help improve results such as word count as the word vector represents a document of numbers with every possible word in the set, when taking a specific text a vector is created and the vector number represents what the text contains[40].

TF-IDF gives us the solution to associate every word in the document with a number that represents its relationship to that word within the document. Then these documents will contain similar words and similar vectors will be obtained, and this is the purpose of using the machine learning algorithm

4.4.3 Classification Module

Categorization Modules

Categorization Modules in Fig (4.4.3) The unidentified address is sent in order to perform the appropriate test for its type by extracting the features associated with the URL and comparing it with those features by disclosing it via the Naive Bays form to Unsafe URLs[41] .

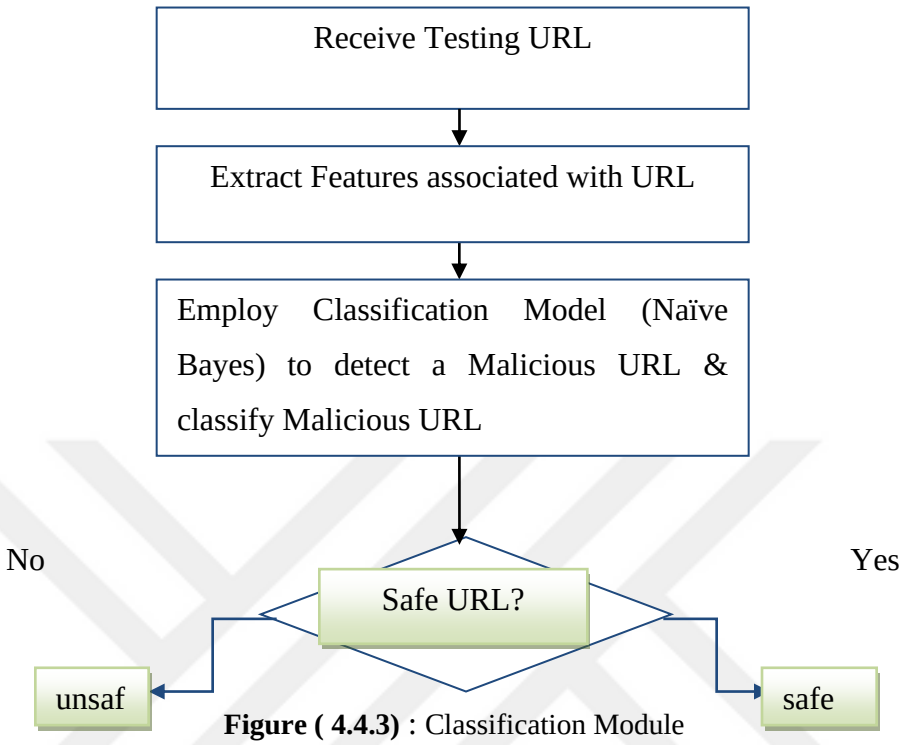


Figure (4.4.3) : Classification Module

4.5 DATA COLLECTION

Safe URL was compiled from two sources :Open Directory project (DMOZ) and Yahoo!'s directory Malicious URLs .The following sources are the approved ones, where the spam was obtained from jwSpam Spy It is called spam We have tested approximately 422466 random links, and 76,453 of them were classified as unsafe links. The number of safe links was 346013 .

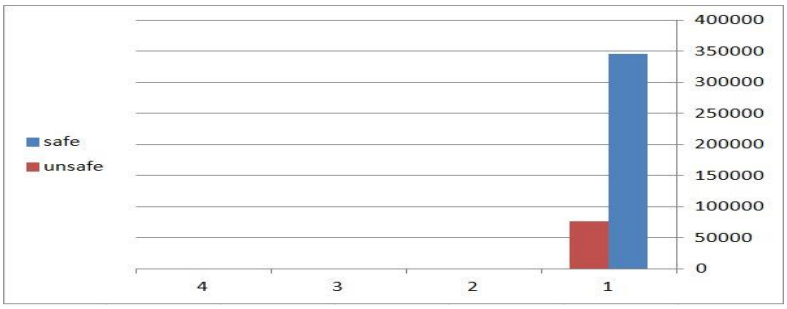


Figure 4.5: Graph Based on Dataset Features

4.6 MATHEMATICAL SAMPLE

Naive Bayes algorithm (NB) represent a cornerstone of ML and data acquisition. An algorithm is to obtain samples that have high predictability [42].

This model provides methods for understanding and exploring data. It is used when the database is large and the required output is accurate compared to the sample set. The probability model for this classifier is a model with specific conditions on a dependent class variable using Bayes' theorem[43].

•X(Z |) = probability of instance being in category Z.

•X (| Z) = prospect of descent example by category Z,

•X(Z) = prospect of appearance of category Z,

•X( ) = prospect example appearing.

We can formulate the equations as follows

posteriorly= (priority*liklyhood)/evedence

The numerator represents the common probability models, while the denominator is present in itself and the characteristic values are placed within it:

X(Z,A1,……..An), Categorize of NB the method used to identify the documents. Equipped by

Groups of suspicious training models to predict unknown samples

URLs. With low results, With conditions related to variables via.

feature (A1, A2) in URLs are separate . All features Av(1<=v<=2) purview

binary value show if it comes via URLs. Probability calculation from H (H1: safe and H2: non-

safe) like :

X (H1/A) = (X (H1)*P (A/Hv))/X (A)

The P (FvH1) and X(HV) very easy to calculate from train. A prospect

to X (h1|A), P(h2|F) are calculating via result like

follow:

X(h1|A)X(h2|A) > x (x>1), safe link.

X(h2|A)X(h1|A) > x , unsafe link.

4.7 Suggested algorithm depended on Naïve bayes (NB)

IN: Trained group, test of URLs

OUT: test the name of domain and aggression kind.

Stage one: The purpose of using sums is to calculate sub-features

Stage two: The variance and average are calculated for each sub-feature as this workbook is designed from distribution files.

Stage three : The individual likelihood is computed .

Stage four : Test features along with existing samples are taken for classification purpose

Stage five : Posteriorly to each group (Safe,

Unsafe) calculating .

Stage six : Resolve posteriorly value of all groups.

Stage seven : The largest category is determined until the next category is determined.

CHAPTER 5

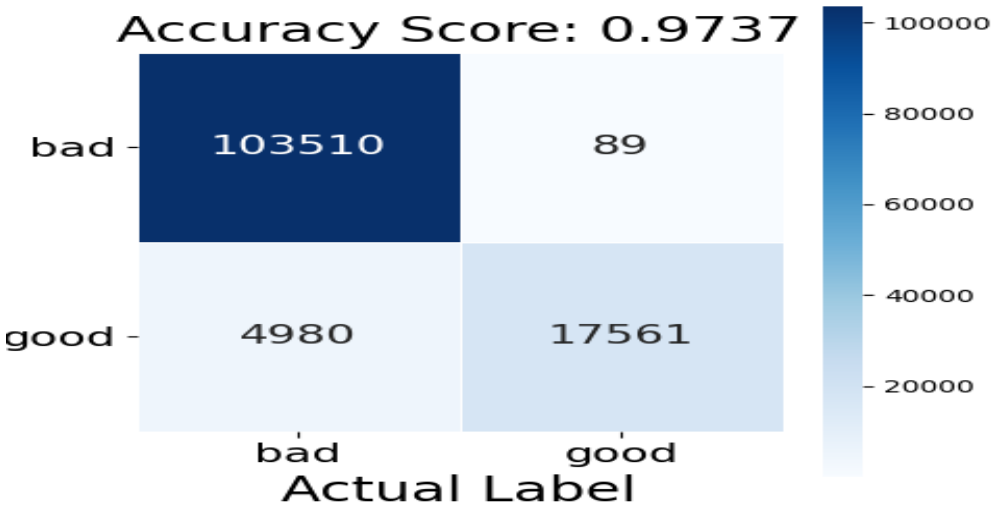
SIMULATION AND RESULTS

5.1 CONFUSION MATRIX

In Machine learning arrays are an important part, it using for testing of performance of algorithm, matrix consists of the table that includes information with detail of the current (human predicted) with categoration of predictive all columns in a matrix represents (Predictive group), and each row represents (Actual class).

And by using the data contained in the matrix, the worker's performance is evaluated. The size of the matrix depends on the number of classes .

The figure below shows the shape of the matrix and the meaning of each column and row in it :



	A	A
P	TP	FP
P	FN	TN

Table 5.1: confusion matrix

True Positive (TP): It represents the number of cases that the classifier correctly predicts of the p-class and actually belonging to the p-class. The negative false (FN): an instance number that a classifier incorrectly predicts represents class p and actually belongs to class n. The negative true (TN): cases of numbers that a classifier correctly predicts represents class n and actually belongs to class p. The Positive false (FP): instances of number that a classifier incorrectly predicts represents class n and actually belongs to class p.

Precision =TP/(TP+FP) , Precision =17561/(17561+89)

Precision = 0.99 , Accuracy = (TP+TN)/Total

Accuracy = (17561+103510)/ (103510+89+4980+17561)

Accuracy = 97%

5.2 COMPARATIVE WORK

Table (5.2) :Correlation with existed comparative works

S .no	Author	Algorithm	Accuracy
1	Garcia, Felan Carlo C., and Felix P. Muga	Random forest	0.9562
2	Hamid, I. R. A., & Abawajy, J.	Hybrid feature	0.93
3	Pektaş, A., & Acarman, T.	Hybrid feature	0.092
4	Olalere, M., Abdullah, M. T., Mahmod, R., & Abdullah, A.	Lexical feature	0.96
5	This work	Naïve Bayes	0.97

Correlation with existed comparative works viruses detect listed in the table (5.2). It appears that the classifier Naive Bayes has proven his superiority over the rest of the models

When compared with similar experiences in the same field.

CONCLUSION

Malicious links are designed to control victims' systems and steal their personal information or destroy their private files, so this phenomenon must be reduced by designing programs that detect malicious links and prevent them from spreading, Therefore, in this study we presented a model using the Naive Bayes classifier with high accuracy in which the links are detected and classified into safe and unsafe. This research, with its comparison with the research group, proved Naive Bayes algorithms are very accuracy than the results of the rest of the other classifiers.

REFERENCES

[1] G. Mahajan, B. Saini, and S. Anand, 'Malware Classification Using Machine Learning Algorithms and Tools', in *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, Gangtok, India, Feb. 2019, pp. 1–8, doi: 10.1109/ICACCP.2019.8882965.

[2] M.-S. Lin, C.-Y. Chiu, Y.-J. Lee, and H.-K. Pao, 'Malicious URL filtering — A big data application', in *2013 IEEE International Conference on Big Data*, Silicon Valley, CA, USA, Oct. 2013, pp. 589–596, doi: 10.1109/BigData.2013.6691627.

[3] M. Sharifi and S. H. Siadati, 'A phishing sites blacklist generator', in *2008 IEEE/ACS International Conference on Computer Systems and Applications*, Doha, Qatar, Mar. 2008, pp. 840–843, doi: 10.1109/AICCSA.2008.4493625.

[4] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, 'PhishNet: Predictive Blacklisting to Detect Phishing Attacks', in *2010 Proceedings IEEE INFOCOM*, San Diego, CA, USA, Mar. 2010, pp. 1–5, doi: 10.1109/INFCOM.2010.5462216.

[5] I. R. A. Hamid and J. Abawajy, 'Phishing Email Feature Selection Approach', in *2011IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications*, Changsha, China, Nov. 2011, pp. 916–921, doi: 10.1109/TrustCom.2011.126.

[6] M. Darling, G. Heileman, G. Gressel, A. Ashok, and P. Poornachandran, 'A lexical approach for classifying malicious URLs', in *2015 International Conference on High Performance Computing & Simulation (HPCS)*, Amsterdam, Netherlands, Jul. 2015, pp. 195–202, doi: 10.1109/HPCSim.2015.7237040.

[7] Y. Li, R. Ma, and R. Jiao, 'A Hybrid Malicious Code Detection Method based on Deep Learning', *IJSIA*, vol. 9, no. 5, pp. 205–216, May 2015, doi: 10.14257/ijsia.2015.9.5.21.

[8] E. Gandotra, D. Bansal, and S. Sofat, 'Zero-day malware detection', in *2016 Sixth International Symposium on Embedded Computing and System Design (ISED)*, Patna, India, Dec. 2016, pp. 171–175, doi: 10.1109/ISED.2016.7977076.

[9] T. Shibahara *et al.*, 'Malicious URL sequence detection using event de-noising convolutional neural network', in *2017 IEEE International Conference on Communications (ICC)*, Paris, France, May 2017, pp. 1–7, doi: 10.1109/ICC.2017.7996831.

[10] A. Joshi, L. Lloyd, P. Westin, and S. Seethapathy, 'Using Lexical Features for Malicious URL Detection -- A Machine Learning Approach', *arXiv:1910.06277 [cs]*, Oct. 2019, Accessed: Apr. 29, 2021. [Online]. Available: http://arxiv.org/abs/1910.06277.

[11] F. Khan, J. Ahamed, S. Kadry, and L. K. Ramasamy, 'Detecting malicious URLs using binary classification through adaboost algorithm', *IJECE*, vol. 10, no. 1, p. 997, Feb. 2020, doi: 10.11591/ijece.v10i1.pp997-1005.

[12] S. Kumi, C. Lim, and S.-G. Lee, 'Malicious URL Detection Based on Associative Classification', *Entropy*, vol. 23, no. 2, p. 182, Jan. 2021, doi: 10.3390/e23020182.

[13] M. Christodorescu, S. Jha, and C. Kruegel, 'Mining specifications of malicious behavior', in *Proceedings of the the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering - ESEC-FSE '07*, Dubrovnik, Croatia, 2007, p. 5, doi: 10.1145/1287624.1287628.

[14] M. Lelarge, ‘Economics of malware: Epidemic risks model, network externalities and incentives’, in *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Monticello, IL, USA, Sep. 2009, pp. 1353–1360, doi: 10.1109/ALLERTON.2009.5394516.

[15] M. Ijaz, M. H. Durad, and M. Ismail, ‘Static and Dynamic Malware Analysis Using Machine Learning’, in *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, Islamabad, Pakistan, Jan. 2019, pp. 687–691, doi: 10.1109/IBCAST.2019.8667136.

[16] A. Moser, C. Kruegel, and E. Kirda, ‘Limits of Static Analysis for Malware Detection’, in *Twenty-Third Annual Computer Security Applications Conference (ACSAC 2007)*, Miami Beach, FL, USA, Dec. 2007, pp. 421–430, doi: 10.1109/ACSAC.2007.21.

[17] A. Damodaran, F. D. Troia, C. A. Visaggio, T. H. Austin, and M. Stamp, ‘A comparison of static, dynamic, and hybrid analysis for malware detection’, *J Comput Virol Hack Tech*, vol. 13, no. 1, pp. 1–12, Feb. 2017, doi: 10.1007/s11416-015-0261-z.

[18] C. Willems, T. Holz, and F. Freiling, ‘Toward Automated Dynamic Malware Analysis Using CWSandbox’, *IEEE Secur. Privacy Mag.*, vol. 5, no. 2, pp. 32–39, Mar. 2007, doi: 10.1109/MSP.2007.45.

[19] J. O. Kephart and S. R. White, ‘Measuring and modeling computer virus prevalence’, in *Proceedings 1993 IEEE Computer Society Symposium on Research in Security and Privacy*, Oakland, CA, USA, 1993, pp. 2–15, doi: 10.1109/RISP.1993.287647.

[20] P. Van Mieghem, J. Omic, and R. Kooij, ‘Virus Spread in Networks’, *IEEE/ACM Trans. Networking*, vol. 17, no. 1, pp. 1–14, Feb. 2009, doi: 10.1109/TNET.2008.925623.

[21] C. Nachenberg, ‘Computer virus-antivirus coevolution’, *Commun. ACM*, vol. 40, no. 1, pp. 46–51, Jan. 1997, doi: 10.1145/242857.242869.

[22] S. M. Tabish, M. Z. Shafiq, and M. Farooq, ‘Malware detection using statistical analysis of byte-level file content’, in *Proceedings of the ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics - CSI-KDD ’09*, Paris, France, 2009, p. 23, doi: 10.1145/1599272.1599278.

[23] R. S. Chakraborty, S. Narasimhan, and S. Bhunia, ‘Hardware Trojan: Threats and emerging solutions’, in *2009 IEEE International High Level Design Validation and Test Workshop*, San Francisco, CA, USA, Nov. 2009, pp. 166–171, doi: 10.1109/HLDVT.2009.5340158.

[24] C. C. Zou, D. Towsley, Weibo Gong, and Songlin Cai, ‘Routing Worm: A Fast, Selective Attack Worm Based on IP Address Information’, in *Workshop on Principles of Advanced and Distributed Simulation (PADS’05)*, Monterey, CA, USA, 2005, pp. 199–206, doi: 10.1109/PADS.2005.24.

[25] C. Shannon and D. Moore, ‘The spread of the Witty worm’, *IEEE Secur. Privacy*, vol. 2, no. 4, pp. 46–50, Jul. 2004, doi: 10.1109/MSP.2004.59.

[26] R. K. Shahzad, S. I. Haider, and N. Lavesson, ‘Detection of Spyware by Mining Executable Files’, in *2010 International Conference on Availability, Reliability and Security*, Krakow, Feb. 2010, pp. 295–302, doi: 10.1109/ARES.2010.105.

[27] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, ‘A comparison of machine learning techniques for phishing detection’, in *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit on - eCrime ’07*, Pittsburgh, Pennsylvania, 2007, pp. 60–69, doi: 10.1145/1299015.1299021.

[28] S. F. Verkijika, '"If you know what to do, will you take action to avoid mobile phishing attacks": Self-efficacy, anticipated regret, and gender', *Computers in Human Behavior*, vol. 101, pp. 286–296, Dec. 2019, doi: 10.1016/j.chb.2019.07.034.
[29] P. Patel, D. M. Sarno, J. E. Lewis, M. Shoss, M. B. Neider, and C. J. Bohil, 'Perceptual representation of spam and phishing emails', *Appl Cognit Psychol*, vol. 33, no. 6, pp. 1296–1304, Nov. 2019, doi: 10.1002/acp.3594.
[30] J. Y. Ndagi and J. K. Alhassan, 'Machine Learning Classification Algorithms for Adware in Android Devices: A Comparative Evaluation and Analysis', in *2019 15th International Conference on Electronics, Computer and Computation (ICECCO)*, Abuja, Nigeria, Dec. 2019, pp. 1–6, doi: 10.1109/ICECCO48375.2019.9043288.
[31] R. Brewer, 'Ransomware attacks: detection, prevention and cure', *Network Security*, vol. 2016, no. 9, pp. 5–9, Sep. 2016, doi: 10.1016/S1353-4858(16)30086-1.
[32] K. Ramachandran and B. Sikdar, 'Modeling Malware Propagation in Networks of Smart Cell Phones with Spatial Dynamics', in *IEEE INFOCOM 2007 - 26th IEEE International Conference on Computer Communications*, Anchorage, AK, USA, 2007, pp. 2516–2520, doi: 10.1109/INFCOM.2007.312.
[33] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, 'Learning to detect malicious URLs', *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–24, Apr. 2011, doi: 10.1145/1961189.1961202.
[34] R. Patgiri, H. Katari, R. Kumar, and D. Sharma, 'Empirical Study on Malicious URL Detection Using Machine Learning', in *Distributed Computing and Internet Technology*, vol. 11319, G. Fahrnberger, S. Gopinathan, and L. Parida, Eds. Cham: Springer International Publishing, 2019, pp. 380–388.
[35] A. Pektaş and T. Acarman, 'Ensemble Machine Learning Approach for Android Malware Classification Using Hybrid Features', in *Proceedings of the 10th International Conference on Computer Recognition Systems CORES 2017*, vol. 578, M. Kurzynski, M. Wozniak, and R. Burduk, Eds. Cham: Springer International Publishing, 2018, pp. 191–200.
[36] F. Shang, Y. Li, X. Deng, and D. He, 'Android malware detection method based on naive Bayes and permission correlation algorithm', *Cluster Comput*, vol. 21, no. 1, pp. 955–966, Mar. 2018, doi: 10.1007/s10586-017-0981-6.
[37] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, 'Multinomial Naive Bayes for Text Categorization Revisited', in *AI 2004: Advances in Artificial Intelligence*, vol. 3339, G. I. Webb and X. Yu, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 488–499.
[38] M. Ontivero-Ortega, A. Lage-Castellanos, G. Valente, R. Goebel, and M. Valdes-Sosa, 'Fast Gaussian Naïve Bayes for searchlight classification analysis', *NeuroImage*, vol. 163, pp. 471–479, Dec. 2017, doi: 10.1016/j.neuroimage.2017.09.001.
[39] Md. S. Rana, C. Gudla, and A. H. Sung, 'Evaluating Machine Learning Models for Android Malware Detection: A Comparison Study', in *Proceedings of the 2018 VII International Conference on Network, Communication and Computing - ICNCC 2018*, Taipei City, Taiwan, 2018, pp. 17–21, doi: 10.1145/3301326.3301390.
[40] Z. Qu, X. Song, S. Zheng, X. Wang, X. Song, and Z. Li, 'Improved Bayes Method Based on TF-IDF Feature and Grade Factor Feature for Chinese Information Classification', in *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Shanghai, Jan. 2018, pp. 677–680, doi: 10.1109/BigComp.2018.00124.
[41] R. Rajalakshmi and C. Aravindan, 'A Naive Bayes approach for URL classification with

supervised feature selection and rejection framework: NB for URL Classification with FS and RF', *Computational Intelligence*, vol. 34, no. 1, pp. 363–396, Feb. 2018, doi: 10.1111/coin.12158.

[42] F. Razaque *et al.*, 'Using naïve bayes algorithm to students' bachelor academic performances analysis', in *2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, Salmabad, Nov. 2017, pp. 1–5, doi: 10.1109/ICETAS.2017.8277884.

[43] Joaquín Abellán and Javier Castellano, 'Improving the Naive Bayes Classifier via a Quick Variable Selection Method Using Maximum of Entropy', *Entropy*, vol. 19, no. 6, p. 247, May 2017, doi: 10.3390/e19060247.