Learning a Deep Neural Network for Predicting Phishing Website

by

Robat Das
13101130
Md. Mukhter Hossain
14301131
Shariful Islam
13201005
Abujarr Siddiki
13321060

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
May 2019

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Abujarr Siddiki
13321060

Mukhter Hossain
14301131

Robat Das
13101130

Shariful Islam
13201005

Approval

The thesis/project titled "Learning a Deep Neural Network for Predicting Phishing Website" submitted by

1. Robat Das (13101130)
2. Mukhter Hossain (14301131)
3. Shariful Islam (13201005)
4. Abujarr Siddiki (13321060)

Of Spring, 2019 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on April 25,2019.

Examining Committee:

Supervisor:
(Member)

Dr. Md. Ashraful Alam
Assistant Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Dr. Jia Uddin
Associate Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Dr. Md. Abdul Mottalib
Professor
Department of Computer Science and Engineering
BRAC University

Acknowledgement

First of all, praise to the Almighty for whom our thesis have been completed without any major interruption.
Secondly, to our Advisor Dr. Ashraful Alam sir for his kind support and advice in our work. He helped us whenever we needed help.

And finally to our parents without their throughout sup-port it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Abstract

In recent years, we have seen a huge paradigm shift in business because of the fast development of the Web. For this reason, consumers change their tendency from customary shopping to the electronic business. In the time of electronic and versatile trade, huge quantities of money related exchanges are directed online on regular schedule, which created opportunities for new potential scheming chances. By utilizing the unknown structure of the Web, attackers set out new procedures like phishing, to fool people with the utilization of false sites to gather their delicate data. It gather datas such as account IDs, usernames, passwords, credit card information and so on. In spite of the fact that organizations and software companies uses methodologies such as heuristics, visual and machine learning to prevent phishing attacks, still these can't keep the majority of the phishing assaults. In this paper, we evaluate the model by using LSTM technique by comparing it with previous studies and we will try to find out the best features in LSTM.

Keywords: Phishing Attack; Scheming; Machine learning; Neutral Network; LSTM

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

FN	False Negative
FP	False Positive
FPR	False Positive Rate
HTTPS	Hyper Text Transfer protocol
IFrame	Initial Frame
NV	Naive Bayes
PDENFF	Phishing Dynamic Evolving Neural Fuzzy Framework
RF	Random Forests
SFH	Server Form Handler
SMO	Sequential Minimal Optimization
TN	True Negative
TP	True Positive
TPR	True Positive Rate
UCI	University of California, Irvine
USD	United States Dollar
Weka	Waikato Environment for Knowledge Analysis
WHOIS	Domain Name Inquiries
AI	Artificial intelligence
AIWL	Automated Individual White-List
ANN	Artificial Neural Network
AOL	America Online
APT	Advanced Persistent Threat

APWG	Anti-Phishing Working Group
DNN	Deep Neural Network
DNS	Domain Name System
eBay	Echo Bay Technology Group
ICP	Internet Content Provider
IP	Internet Protocol
ISP	Internet Service Provider
LSTM	Long Short Term Memory
LUI	Login User Interfaces
PC	Personal Computer
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
URL	Uniform Resource Locator
WWW	World Wide Web

Chapter 1

Introduction

Growth of the internet is getting higher and higher day by day. With the high usage of the internet, new techniques to attack websites are applied. Hackers use various tactics to deceive a general user. They can test you by providing inappropriate picture, webpage or redirect you to inappropriate pages by source code javascript which can be malicious [1].

Phishers work by trading off a genuine space to make phishing sites or by bargaining a current site to incorporate contents to divert to a malignant server where client information can be downloaded or by tricking clients with space names, for example, pay5al.com, pay-pal.com, or paypal.sign-in.online, which resemble favorable site Paypal.com [2].

The word "phishing" comes from "fishing". The idea behind is that like someone will throw out a bait for the users and hope that any user falls for it [1] .

In our case, bait is a malicious URL or simply known as Website. The main issue is that phishing websites are also look like legitimate websites [3],[4] .

In our thesis, we are going to follow on mixed methodology. By and large, mix strategy investigate to look into that includes gathering, examining, and interpreting quantitative what's more, subjective information in a solitary report or in a progression of studies that examine the equivalent basic marvel [5]. Thus various studies illustrated on collecting data, testing etc., we have found comparing results are more worth it to fulfill the goal of our research.

1.1 Problem Statement

The security challenges nowadays we are facing due to phishing are increasing rapidly. According to a prominent Washington based cyber security company F5 Networks, Inc. stated in a report that, strategy of a phishers includes three individual missions marked as Target selection, Social engineering and Technical engineering [6]. As per the Anti-Phishing Working Group, there were 18,480 remarkable phishing assaults and 9666 interesting phishing locales detailed in March 2006 [7]. It effects billions website users and big costing barrier to businesses [7], [8]. As indicated by Microsoft, in 2018, the potential expense of digital wrongdoing to the worldwide network is a marvelous 500 billion USD and an information break will cost the normal organization about 3.8 million USD.

With the huge work exists for phishing attack detection, in this thesis we are going to use Long Short Time Memory(LSTM) algorithm on thirty features we have selected.In addition, we want to show improvement in some features. All in all, the following problems are going to carried out in this thesis:

1. Detecting the best classification algorithms for the features.
2. Determining the best features to be used for detecting phishing attack.
3. A way to improve the performance of best features selected (if possible).

1.2 Objectives of the Study

The objective of this exploration is to direct a relative appraisal between different classified algorithm systems and within the different feature selection outline. Objectives of this research are as followings:

1. Detecting the best feature to be used for detecting phishing attack in URLs both manual features applied in URL structure and automated selection technique.
2. Detecting the best classification algorithm for the features make comparison if LSTM is better than the other classified algorithms.
3. Finding the best features in LSTM.

1.3 Motivation

The quantity of phishing assaults has been developing impressively as of late and is considered as one of the most perilous present day web violations, which may lead people to lose trust in internet business. Thus, it has an enormous negative impact on online trade, promoting endeavors, associations' earnings, connections, clients, and by and large business activities. In this thesis, we are going to make comparisons and comments on various features active on the field of phishing website detection.

The harmful effects of phishing could be extent to access the users' confidential details, which could result in financial losses for users and even prevent them from access their own accounts. Therefore, in this study, we will quantify and qualify the phishing website features to prevent and mitigate the risk of phishing websites.

1.4 Scope and Limitation

The scope of this research is phishing website detection where 30 features were selected and categorized in four groups that cover important part of website components. In this study,a deep learning algorithm named LSTM classifier algorithm is used for phishing websites detection.
The limitation of this research is that it won't cover phishing email detection. Instead, this research focus on how website component performs on LSTM.

1.5 Contribution

The thesis goal is to build an efficient prevention model that uses data mining techniques. The contribution of the thesis are described below:

1. Select best sets of features for phishing detection automatically.
2. Experimentally evaluate the performance of the classification algorithms for phishing website detection techniques.
3. Propose a multi integration system for phishing detection.

1.6 Thesis Organization

The thesis is consists of five chapters organized as follows:

1. Chapter 1 : Introduction: Overview of phishing attacks, problem statement, the objective of the study, the motivation, the scope and limitation, thesis contribution and finally thesis organization.

2. Chapter 2 : Research Background: This chapter provides an overview of the related works in phishing website detection and summary of the articles that published by other researchers.

3. Chapter 3 : Methodology Dataset: This chapter explains an outline of the research methodology which is used in our thesis. Overview of the software that used for the evaluation of the proposed method and the Dataset were used in this research.

4. Chapter 4 : The implementation details of experiment and the results that were obtained for all the proposed scenarios and comparison of the results.

5. Chapter 5 : Conclusion and Future work.

Chapter 2

Research Background

2.1 Literature Review

Phishing has grown dramatically since its America online show day. Phishers began focusing on online installment frameworks. In spite of the fact that the main assault in June 2001, which was on E-Gold, was not estimated to be successful, it was the structure square to what came later. In the last quarter of 2003, phishers enrolled several spaces that have all the earmarks of being authentic locales like eBay and PayPal. Phishers utilized embraced worm projects to spread parodied messages to PayPal or eBay clients. Unfortunate casualties were diverted to caricature locales and provoked to refresh their accreditation, charge card data and other distinguishing data [9].

A report from Anti Phishing Working Group(APWG) which is published on 2016 phishing is a criminal mechanism employing both social engineering and technical subterfuge to steal internet users personal information. Social Engineering method use a website which is looks like a legitimate website to grab a internet users attention. This website is using a well known brand name. By clicking any button from that website it can access internet users personal and sensitive information[10].

An example of cyber crime can be given by the attacks made on American Online (AOL), one of the largest internet service provider in America. In early 90's a group of hacker named, 'The Warez Community' did build an algorithm and that can produce fake credit card number. After producing or generating a fake credit card number then they use that number to open a cringe AOL user account. After doing this, then they try to spam another AOL user from the AOL user pool. [11].

Though AOL solved the issue successfully but later the methodology of phishing changing everyday. Different Bank, Social Sites, Credit Card providers, Corporations have their own user database and hackers want to have their system information to access a big number of data. If they can access a single users data, from that user data they can contact with the other user and spreading phishing on that network easily.

After AOL incident world has faced many phishing attacks and phishers are targeting Banking sector. Online payment sectors are initial targets for them. Till 2003, phishers have registered hundreds of domains which are seems like paypal and eBay and these domains are so legitimate that a single individual can easily trapped by them[11].

Amid the most recent decade, the ascent of state-supported PC hacking, coined Advanced Persistent Threat (APT) for its stealthy and nonstop nature, has caused incredible worry among security experts and analysts and figured out how to draw the consideration of the overall population[12].

(Bejtlich,2007) explains the components of APT. Advanced refers to the enemy is acquainted with PC guidance instruments and procedures and is equipped for creating custom adventures. Persistent means the adversary intends to accomplish a mission. They recieve directives and work towards specific goals. Threat means the adversary is organized, funded and motivated.
In figure 1, the whole life cycle of Advanced Persistent Threat (APT)- architecture is drawn. First of all, attackers select the category of users that they will like to test. Secondly, as initial infection, attackers act as a normal user and enters into the field with other users and then try to control over the personal information of the users. Internal pivot refers to attackers command and controls over the users that remains still in the websites after the infection and wait for the next attempt. After that attackers try to extract data from users by acting like admin [13].

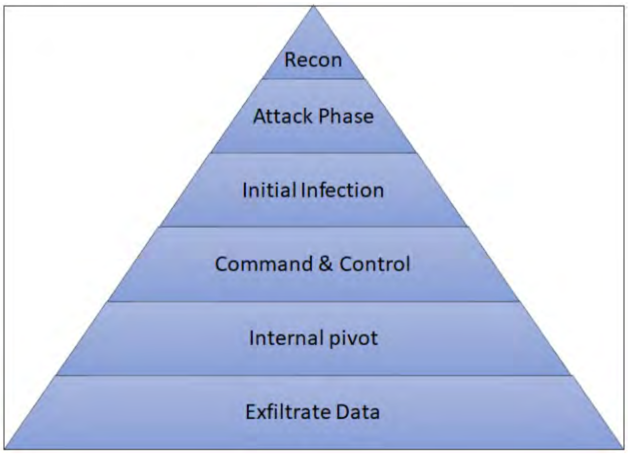

Figure 2.1: Life cycle of Advanced Persistent Threat(APT))

The achievement of phishing site identification procedures essentially relies upon perceiving phishing sites precisely also, inside a satisfactory timescale. Numerous ordinary methods dependent on fixed highly contrasting posting databases have been recommended to recognize phishing sites [9]. In any case, these procedures are not sufficiently effective since another site can be propelled inside couple of moments. Accordingly, the majority of these procedures are not ready to settle on a precise choice progressively on whether the new site is phishing or not. Thus, numerous new phishing sites might be classified genuine sites [1],[2],[4].

A framework was found by Al Momani and others that also detects unknown zero-day phishing emails relying on a the “evolving connectionist system”. The new system was named the phishing dynamic evolving neural fuzzy framework (PDENFF) and follows a hybrid learning approach (supervised/ unsupervised) and is supported by an offline learning feature to achieve the intended purpose. Using this system helped in enhancing the detection of zero-day phishing e-mails was improved between 3% and 13%. Moreover, it used rules, classes or features to enhance the learning process using ECOS which provided the system with the advantage of distinguishing phishing emails from legitimate one [8].

In 2007, a study was conducted to measure the efficiency of the existing tools for phishing detection. This study showed that even the best phishing detection toolbars missed over 20% of the phishing websites [14]. Another study in [15] and [16] phishing detections approaches based on heuristics check common properties of phishing sites such as unique keywords used in URLs or web pages to identify zero-day phishing websites. Nevertheless, these types of heuristics will be effortlessly by-passed by attackers once their mythology is exposed. Visual similarity-based detection techniques have been proposed to circumvent this limitation. Since phishing web pages must imitate victim sites, visual similarity between phishing sites and their target sites is alleged to be an inherent and not simply concealable property. However, these techniques require images of real target sites for detection.

In [7], it was proposed to use a phishing detection mechanism based on visual similarity among phishing sites that imitate the same target website. It was claimed that Just by analyzing visual similarity among web pages without a previous information, the method automatically extracts 224 different web page layouts imitated by 2,262 phishing sites. However, it achieves a detection rate around 80% while maintaining the false-positive rate to 17.5%.

Another study by [17], the URL and content material feature based approach makes a speciality of reading the traits of the URL and the content material of a goal internet site. In that study become created a classification model that could come across phishing websites by means of concerning particular area functions. The proposed version does no longer depend on previous expertise or assumptions about authentic websites.That study prioritizes URL and content material characteristic primarily based method, when you consider that it's far the most normally used technique, due to the fact it may integrate and evaluate detection capabilities on a domain. by using integrating new functions on present website with a few detection function prediction utilized in preceding research, a feature vector was created for the proposed version that includes two components: URL features and web content functions. URL feature includes the following steps:

1. If the URL contains an IP address because most of the case phishing websites contain IP addresses in the URLs. The URL of a target internet site incorporates an IP deal within preference to a domain call, then this option variable feature might be assigned 1; in any other case 0.

2. Regardless of whether a URL contains the image '@' Phishing sites frequently embed @ into a URL that takes clients to a site not the same as what Internet offenders anticipate. In the event that a URL contains the image '@', the

following feature will be allocated an estimation of 1; generally 0.

3. Regardless of whether the characters in a URL are coded in UNICODE In contrast with an honest site, a phishing site is bound to utilize UNICODE in its URL to shroud the URL of a genuinely planned site. The feature will be doled out an estimation of 1 if the space name of the URL of an objective site contains characters encoded in UNICODE; generally 0.

4. The quantity of dots ('.') in a URL Past research [5] recommends that the bigger the number of dots in a URL, the higher the likelihood the site is a phishing site.

5. The quantity of suffixes in an area name Users for the most part get a look at the initial segment of a URL in any case, likely miss the rest of the part, which really focuses to a phishing site.

6. Age of a domain name which is spoken to by the quantity of days spend since an area name was enrolled.

7. Termination time of an domain name Which is spoken to by the quantity of days staying before a domain name lapses.

8. Regardless of whether the location of a DNS (Domain Name Server) is steady with a URL DNS server locations can be gotten through WHOIS (Domain name inquiries). On the off chance that it coordinates, the estimation of the feature will be 1; generally 0.

9. Information regarding website registration. If registered, 1; otherwise 0;

10. This feature checks whether the domain is registered as individual (1) or as an enterprise(0).

11. Regardless of whether area privatized by proprietor. Following feature is utilized to speak to whether a recorded site name and genuine showed site are predictable (1) or not (0).

    Web content highlights are automatically dragged from the source code of a site and the features includes the following:

12. Regardless of whether site contains ICP (Internet Content Provider) permit number. On the off chance that the site contains ICP permit number, at that

point feature no. 12 will give the result 1; else, it will be 0.

13. The quantity of void (invalid) interfaces on a site. According to past study [18], a phishing site will in general have more void connections than a legitimate site.

14. The number of out links on a website. It is normal for a website to have some out links, but when there are too many, it may increase the probability of a website being a phishing website.

15. Last feature works on e-business certification whether it provides e-commerce certificate information;if not, the feature gives result 0 and 1 otherwise.

2.2 TYPES OF PHISHING ATTACK

Attackers are creating numerous amounts of ways for exploiting the holes of internet security of users. Various distinctive sorts of phishing assaults have now been recognized. A portion of the more pervasive are described below :

Algorithm-Based Phishing

America Online (AOL) flagged the concept of phishing in the early 1990s [11]. Amid that time, the first phishers made an algorithm to create arbitrary credit card numbers to get a unique card's match from the AOL accounts.

Deceptive Phishing

The expression ”phishing” initially alluded to account theft utilizing texting ; however the most widely recognized communicate strategy today is a Deceptive email message. This messages can be regarding verification of account information, framework failure expecting clients to re-enter their data and many other similar type scams that contain links. After clicking the link the users are redirected to hackers' websites where their personal information are licked and stored [8].

URL Phishing

In URL phishing assaults, hackers utilize the phishing page's URL to taint the objective [1]. One of the many ways to trap a person is by using a hidden link. These links are generally remain hidden behind “click here”, “Subscribe” and many other type of buttons. By clicking those buttons, users are redirected to attackers' page.

Hosts File Poisoning

Before entering into internet , users' given URL must have to be translated into an IP address. Most of small and medium sized business clients' PCs running a Microsoft Windows working framework first look up these ”host names” in their ”hosts” record before embraced a Domain Name System (DNS) query. By ”Poisoning” the hosts record, phishers have a fake location transmitted, taking the client accidentally to a phony ”resemble the other alike” website where their data can be stolen[11].

Content-Injection Phishing

It describes the circumstance where hackers supplant some portion of the substance of a genuine site with false substance intended to deceive or mislead the client into surrendering their secret data to them. For instance, hackers may embed malignant code to log client's certifications or an overlay which can subtly gather data and convey it to the phishers' phishing server [8].

Clone Phishing

In a clone phishing assault, a formerly sent email containing any connection or connection is utilized as a genuine duplicate to make a practically indistinguishable or cloned email [8]. Phishers supplant the connection or connection in the email with a noxious connection or attachment.The cloned email is sent to the contacts from the victims' inbox. The beneficiaries of the cloned email will expect it to be a genuine email and snap on the malignant link [11].

Nevertheless, there are so many other types of phishing attacks but we mentioned some general and common types.

2.3 Life Cycle of a Phishing Attack

Phishing attacks can be divided into two layers; social engineering and technical subterfuge. Social engineering layer includes attackers, victim, sending fake email, which contains spoofed webpages. This process starts by sending this email, which comes from organizations for gathering some sensitive information such as user name, id, password, credit card information etc. Second layer is about spoofed web page. Fake email directs the victim to the spoofed web page which appears visually very similar to the original page. This layer also uses cross-site scripting, session hijacking, malware phishing, DNS poisoning and key/screen loggers' techniques. These layers send the obtained information and get remote access by attackers to victim's computer or original webpage [19],[14].

In figure 2.2, the figure is showing the life cycle of a phishing attack.

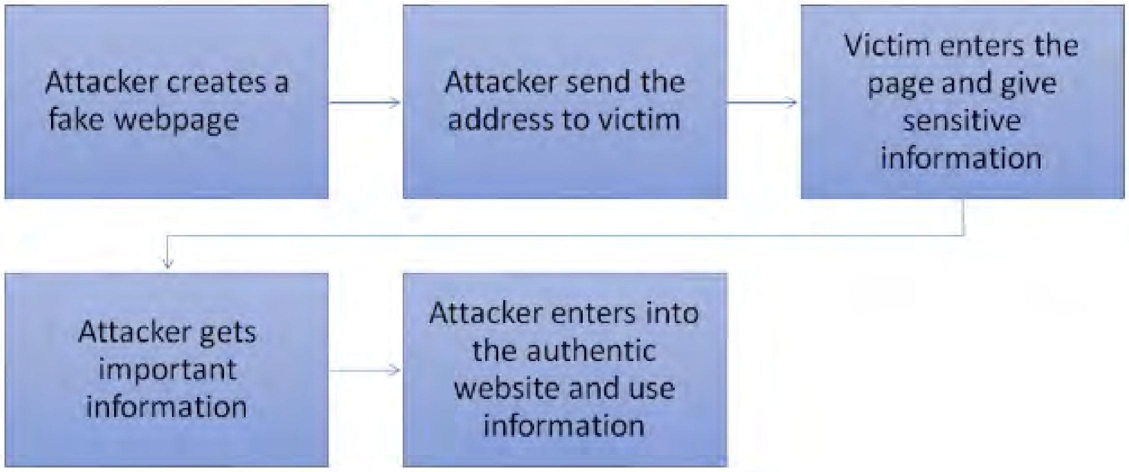
Attacker creates a fake webpage
Attacker send the address to victim
Victim enters the page and give sensitive information
Attacker gets important information
Attacker enters into the authentic website and use information

Figure 2.2: Life cycle of a Phishing Attack

2.4 Phishing Website Detection Techniques

There is versatile number of filtering process right now to predict and prevent phishing website and manage occurring threats on traditional techniques and modern techniques of machine learning or data mining which are developed by scientist and specialist [20].

2.4.1 Traditional Methods

Traditional Methods of detection categories into two major category, the network-level protection and another one is authentication protection. Category one which is the network-level protection, this category includes Blacklist Filters and Whitelist Filters. Blacklist Filters and Whitelist Filters can prevent phishing by suspected domains from accessing the network and internet protocol (IP) addresses. There is also Rule-based filters and Pattern Matching filters, these are based on manually entered and updated fixed system for phishing detection [14].

Blacklist Filter

This technique gives protection at a network level by classifying websites address, IP address or DNS address. This is one of the predominant phishing filtering techniques. Twenty or more than that phishing filtering blacklists are using now days. These technique may contain IP addresses or domains which are used by known phishers, open proxies and relays IP addresses, country and ISP netblocks that send phishing elements and virus and exploit attackers. All these information are taken out from these details and prevent the phishing [21].

Whitelist Filter

The white-list separating gives assurance at system level too, however in opposite to blacklists; this strategy contrasts the user login information, payment information and a predefined list containing static IP locations of authentic spaces and IP addresses. In such manner, just messages with information coordinating the rundown will be permitted to get to the system to the client's inbox. date. Contrary to blacklist, white-list approach maintains a list containing all legitimate websites[8].

AIWL (Automated Individual White-List) automatically attempts to keep up a white-rundown of client's everything commonplace Login User Interfaces (LUIs) of websites. When a client endeavors to present his/her private data to a LUI that isn't in the white-list, AIWL will caution the client to the conceivable assault. Next, AIWL can effectively safeguard against phishing assaults, in light of the fact that AIWL will alert the client when the genuine IP is vindictively changed. Messages with information coordinating to this rundown may be delegated real dependent on this channel, while different messages are considered phishing and kept from getting to the system for which this channel in called additionally genuine messages classifier [5],[6].

Pattern Matching Filter

Pattern matching is intended to deal with obscure assaults (Blacklist/whitelist is pointless for this situation). For this type of phishing assaults, all the data we have is the genuine connection from the hyperlink (since the visual connection does not contain DNS or IP address of the goal site), which give next to no data to further examination. So as to determine this issue, we attempt two strategies: First, we separate the sender email address from the email. Since phishers for the most part attempt to trick clients by utilizing (parodied) legitimate DNS names in the sender email address, we expect that the DNS name in the sender address will be not the same as that in the real connection [8],[18].

Second, we proactively gather DNS names that are physically contribution by the client when she surfs the Internet and store the names into a seed set, and since these names are contribution by the client by hand, we expect that these names are reliable.

If the two DNS names are similar but not identical, then it is a possible phishing attack. For instance, Pattern Matching can easily detect the difference between www.icbc.com.cn (which is a good e-commerce Web site) and www.1cbc.com.cn(which is a phishing site), which has similarity index 80% [18].

2.4.2 Automated Method

Regardless of the improvement of aversion techniques, phishing remains a fundamental hazard even after the essential countermeasures and in perspective on open URL blacklisting. This methodology is lacking a direct result of the short lifetime of phishing sites . So as to conquer this issue, building up an ongoing phishing site discovery strategy is a compelling arrangement. This examination presents the PrePhish calculation which is a robotized AI way to deal with dissect phishing and non-phishing URL to deliver dependable outcome. It speaks to that phishing URLs regularly have couple of associations between the piece of the enlisted space level and the way or on the other hand question level URL. Utilizing these associations URL is described by between relatedness and it gauges utilizing highlights mined from traits. These highlights are then utilized in AI method to recognize phishing URLs from a genuine data set. The arrangement of phishing and non-phishing site has been executed by finding the range esteem and edge an incentive for each characteristic utilizing basic leadership order. This technique is likewise assessed in tensorflow utilizing RNN and ANN to discover how it functions on the dataset evaluated [22].

Random Forest

Random Forest is one of the most popular and powerful algorithm because:

1. simplistic approach
2. recovering lacking of decision tree

3. useful for both classification and regression

Random Forest builds multiple decision tree and merges them together to get a more accurate and stable prediction. Most of the time it is trained with bagging method with bootstrapped dataset. Bagging method is the combination of all the learning models which increases the overall result and create better accuracy[23]. Bootstrapped dataset means randomly sampling datasets with replacements. It means that it can take a sample data multiple times and a sample data can be ignored in a random forest which later can be used as testing data. Moreover, Random Forest uses almost the same hyper-parameters which are used in decision tree or a bagging classifier. But in this case we don't need to combine a decision tree with a bagging classifier and we can only use the classifier class of random forest.[23] While growing the trees, Random Forest adds additional randomness to the model. It searches for the best feature among a random subset of features instead of most important features while splitting the node. For this reason, it generates an wide diversity that generally results in a better and accurate model. So in this algorithm, we consider only a random subset of the features for splitting a node.We can make trees more random by additionally using random thresholds for each feature rather than searching for the best possible threshold.[23],[24]. Feature Importance-

1. Easily we can measure the important of each features on the prediction making process by looking at how much tree nodes, which uses that features reduce impurity of the forest.
2. For each feature it calculates the score after training and scales the results so that the sum of all importance is equal to 1.
3. By looking at the feature importance we can choose which feature i should take or which one i should ignore because they do not contribute enough or may be they have no contribution to the prediction process. The more features you have, the more likely our model will suffer from fitting.

Precondition: A training set S := (x1, y1),..., (xn, yn), features F, and number of trees in forest B.

```
function RandomForest(S , F)
H ← 0
for i $\epsilon$1,..., B do
S (i) ←A bootstrap sample from S
H ←H $\bigcup\{h\_i\}$
end for
return H
```

```
function RandomizedTreeLearn(S , F)
At each node:
f ←very small subset of F
Split on best feature in f
return The learned tree
```

Precondition: A training set S := (x$_1, y_1), ..., (x_n, y_n)$, features F, and number of trees in forest B.

Gaussian Naive Bayes

Naive Bayes is a probabilistic classifier that makes classifications using the Maximum A Posteriori decision rule in a Bayesian setting. This Classification is named after Thomas Bayes (1702-1761), who proposed the Bayes Theorem [25]. In machine learning, Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem . It makes two fundamental assumption about the dataset that all the features are independent and equal.This classifier uses the Bayes theorem to classify. Equation for this algorithm is -

$$P(X|c_i) = n \prod k = 1P(xk|ci) = P(x_1|c_i)$$

·P(x$_2|c_i$)·P(x$_2|c_i$)·····P(x$_n|c_n$)(2.1)

Pseudo code for this classification algorithm -

```
Learning Phase:
For each class value C$_i$
Computer P(C$_i$)
For each attribute X$_i$
//Compute Distribution D$_{ij}$
if attribute X$_j$ is discrete
D$_{ij}$ =Categorical Distribution for C$_i$ and X$_j$
else
D$_{ij}$ =Normal Distribution for C$_i$ and X$_j$

Testing Phase:
Given unknown X = [x$_1, x_2, ..., x_n$]
estimate class = argmaxC$_i P(C_i) \Delta \prod$j D$_{ij}$
```

Deep Neural Network

A neural network, in general, is a technology built to simulate the activity of the human brain – specifically, pattern recognition and the passage of input through various layers of simulated neural connections.

Many experts define deep neural networks as networks that have an input layer, an output layer and at least one hidden layer in between. Each layer performs specific types of sorting and ordering in a process that some refer to as "feature hierarchy." One of the key uses of these sophisticated neural networks is dealing with unlabeled or unstructured data. The phrase "deep learning" is also used to describe these deep neural networks, as deep learning represents a specific form of machine learning where technologies using aspects of artificial intelligence seek to classify and order information in ways that go beyond simple input/output protocols.

LSTM (Long Short Term Memory)

Long Short Term Memory algorithm – known as ”LSTMs” – are an exceptional sort of RNN, fit for adapting long short term conditions. Their default behavior is to remember information for long periods of time, not something they struggle to learn. They were presented by Hochreiter and Schmidhuber (1997). It uses a combination of hidden units,element wise products and sums between units to implement gates that control “memory cells.” These “memory cells” are designed to retain information without modification for long periods of time. It has its own input and output gates, which are controlled by learnable weights that are a function of the current observation and the hidden units at the previous time step. Below we are describing the most popular variant of LSTM architecture which is complex but has produced state-of-the-art results on a wide variety of problems.

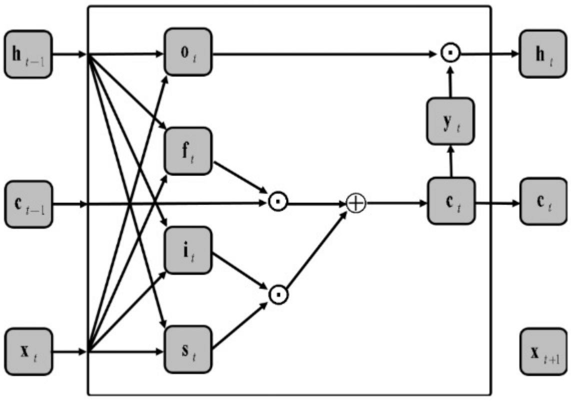

Figure 2.3: LSTM Architecture

At each time step there are three types of gates: input i, forget f and output o. Each are a function of both the underlying input x at time t as well as the hidden units at time t-1, h. Each gate multiplies x by its own gate specific W matrix , by its own U matrix and adds its own bias vector b. This is usually followed by the application of a sigmoidal element wise non-linearity. At each time step t,input gates i are used to determine when a potential input given by s is important enough to be placed into the memory unit or cell,c. Forget gates f allow memory unit content to be erased. Output gates o determine whether y, the content of the memory units are transformed by activation function, should be placed in the hidden units h. Typical gate activation functions and their dependencies are shown.

LSTM unit output,

$$h_t = o_t y_t \tag{2.2}$$

Output get Units,

$$o_t = sigmoid(W_o x_t + U_o h_{t-1} + b_o) \tag{2.3}$$

Transformed memory cell contents,

$$y_t = tanh(c_t) \tag{2.4}$$

Gated update to memory cell units,

$$c_t = f_t c_{t-1} + i_t s_t \tag{2.5}$$

Forget gate units,

$$f_t = sigmoid(W_f x_t + U_f h_{t-1} + b_f) \tag{2.6}$$

Input gate units,

$$i_t = sigmoid(W_i x_t + U_i h_{t-1} + b_i) \tag{2.7}$$

Potential input to memory cell,

$$s_t = tanh(W_c x_t + U_f h_{c-1} + b_c) \tag{2.8}$$

At the beginning of our LSTM, it has to be figured out what data are we going to pass from the beginning state. A Sigmoid layer does this decision. Than it looks at st-1 and yt, and gives a number 1 or 0 as result in the cell pt-1. Here 1 represents as yes or "keep complete of it" and on the other hand 0 says no or "get rid of it completely".

Next mission is to figure out the information we want to save in the cell state which is divided into two branches. First of all, the Sigmoid layer(input gate layer) points out the values we have to update. Then, for new data, a tanh layer makes a new vector.

It's currently time to update the previous cell state, into the new cell state. The previous steps already determined what to try and do, we tend to simply have to really pair. We multiply the previous state forgetting the items we tend to determined to forget earlier. Then the sigmoid layer tends to add it. This is often the new values, scaled by what quantity we tend to determined to update every state worth.

Finally, time for the final result. It will be dependent on the cell state. Moreover, it will be a filtered version. In the first step, a Sigmoid layer figures out which portion of the cell state will be the result. After that the cell state goes through a tanh and arranges a multiplication with the output from Sigmoid gate to give the filtered output.

Chapter 3

Methodology and Dataset

3.1 Introduction

In this chapter, we are going to present our proposed model for detecting phishing websites.Initially,we developed a Deep Neural Network(DNN) based model by classification techniques to enhance the detection accuracy. Moreover, we added comparative analysis between various classification algorithms and feature selection scenarios.

3.2 Proposed Approach

We started our proposed work by finding the correct dataset with complete features. Secondly, we investigate previous works of different authors on the dataset and several classification algorithm techniques applied on the field. Subsequently,we analyzed the accuracy of different phishing website detection techniques applied on our dataset. After that, we choose several main features which can be further improved using DNN alongside with LSTM. By doing this,we reduced time and space to apply our model. Furthermore, we analyze and verify our model and results by comparing with the results of several classification algorithms on the selected features. In the next subsections, we will explain these steps. The figure 3.1 below describes the workflow used in this research.

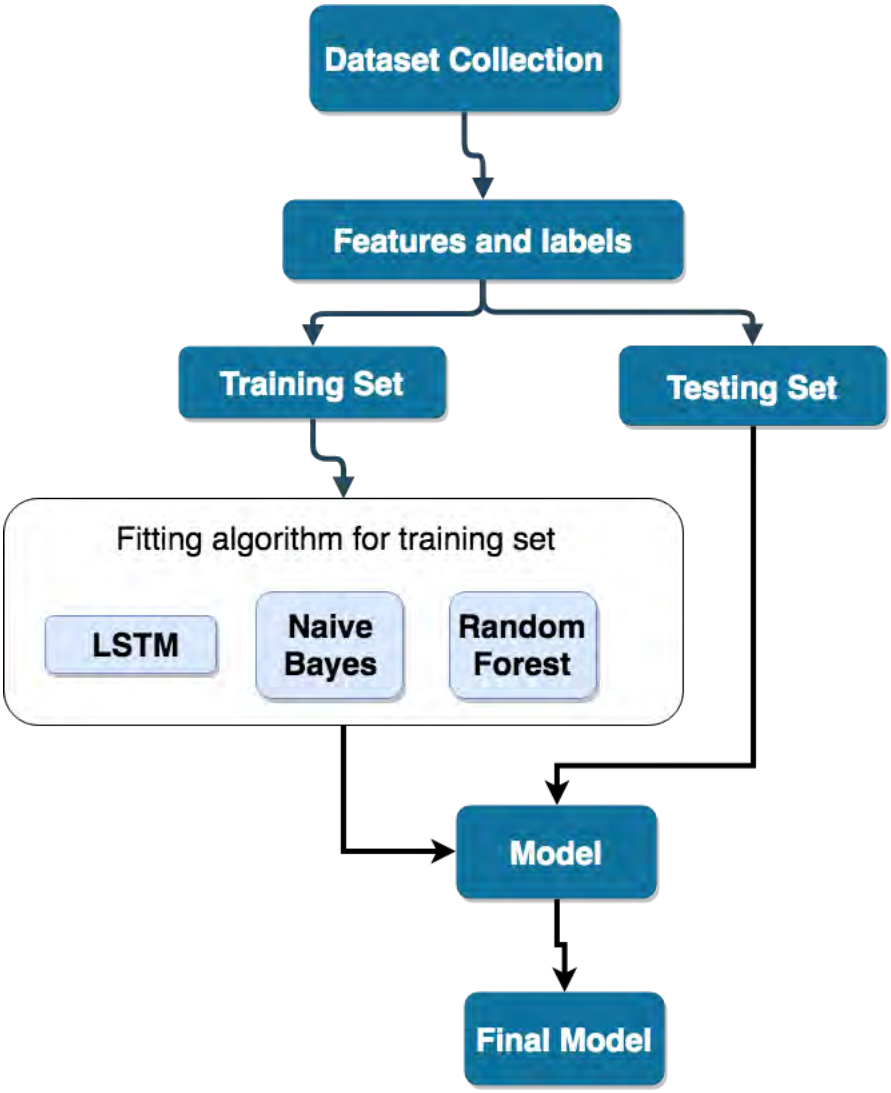

Figure 3.1: Workflow

3.2.1 Data Collection

Data collection and understanding dataset is the initial part of our proposed model.Our dataset was collected from UCI [26] and it has 11055 lists of websites where there are both legitimate and phishing websites shown in the table no. .

Table 3.1: Results after Simulation

Description	Value
Total Features	30
Total Websites	11055
Phishing Websites	4898
Legitimate Websites	6157

To understand how the data was collected, we need to understand the URLs which is described below.

Understanding URLs

If we look at the phishing attack patterns, we can see that URLs(Uniform Resource Locators) play a vital role. We need to have clear understanding of the different parts of URL. Generally, a URL consists of 5 basic components which is depicted in the figure below:

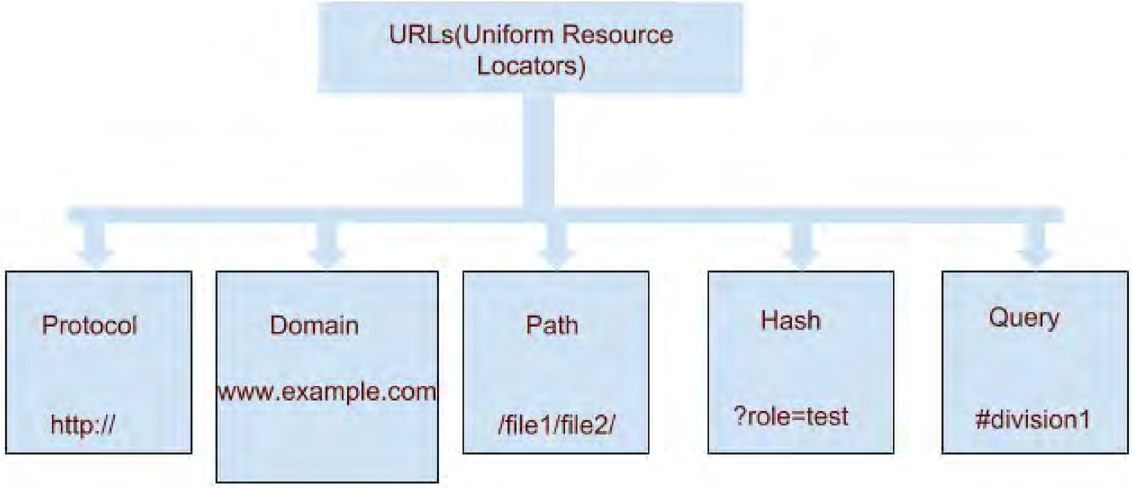

Figure 3.2: Parts Of A URL

First part of the URL is called Protocol which is a set of rules used for transferring data. Second part of the URL is called domain which has several parts. First part of domain often contains “www” prefix which referred as World Wide Web Address. Then, the name of the website is given after a dot. After a dot, we add the organization type (com,edu,gov etc) followed by the country code if added. Third part of URL contains the path details which addresses the section and page of the website. Fourth part consists of query which refers part of a page. Finally the Query part sends additional information sent with the page request.

3.2.2 Feature Explanation

Before stepping into the feature selection part, we need to analyze features and the potentiality of working with these features. Basically, there are 4 main features which has in total 30 sub-features. Based on the data, every single feature provides information about whether the website can be legit,phishing or suspicious. In this section, we are going to emphasize on the features.

Address based Features

Address based feature generally shed light on the name of the URL. It has 12 sub-features which is shown on the table 3.1 below

Abnormal Based Features

Abnormal based feature has 6 sub-features. It generally focuses on unusual activities on the website. In he table 3.2 below contains the explanation of the features.

HTML and JavaScript based features

This features has 5 subfeatures which is shown on the table 3.4 below.

Domain based features

This feature contains 7 sub-features which is described on the table 3.5 below

Table 3.2: Address Based Feature

Feature Name	Feature explanation
IP Address	If the domain contains an IP Address instead of domain name, it considers as phishing website.
URL length	Long URL name can contain malicious contents. If the length of the URL is more than the average length of an URL, it is considered as Suspicious or phishing.
TinyURL	TinyURL is used for shortening the URL length. By clicking on the shorter URL, it redirects to to main page. TinyURL links are considered as phishing website because it can redirect the user into fake website instead of legitimate website.
Having "@" symbol in URL	Browser generally skips the part attached with @ symbol so it is avoided in real addresses.
Using "//" symbol	"//" symbol is used for redirecting to another website. We consider it legit if the sign is used after HTTP or HTTPS. If the symbol is used after the initial protocol declaration, we consider it phishing.
Having "-" in domain name	Most of the real URLS don't contain "-" symbol . We considered an URL phishing if it contains "-" in its domain name.
Dots in domain	We need to add dot to append a sub-domain with the domain name. If more than 1 subdomain occurs, we consider it suspicious and greater than that will point it as a phishing site.
HTTPS	HTTPS protocol and the age of certificate is very important as most of the legit website uses HTTPS and has the trusted certificate.
Domain expiry date	A legit website generally have longer expiry date of their domain name.
Favicon	Favicon is a graphic image used in websites. If it is loaded from external domain, it can redirect user to suspicious sites.
Using unimportant ports	If a URL has some open ports which is unnecessary, phishers can take advantage.
"HTTPS" on domain	If an URL have "HTTPS" on it's domain name, it is considered as phishing website.

Table 3.3: Address Based Feature

Feature Name	Feature Explanation
Request URL	If a page contains higher amount of external URL or contents from another domain, we consider it suspicious or phishing based on the percentage.
Using <a>tags	Similar to the request URL features, the more we see <a>tags used in the website, the risk of phishing increases.
Links in <meta>,<script>and <Link>tag	If <meta>,<script>and <Link>tag contains high amount of external links, it is considered as either suspicious or phishing based on the percentage.
Server Form Handler(SFH)	If SFHs is blank or empty, it is considered as phishing. It is marked as suspicious if the user is redirected to a different domain by SFHs.
Submitting Information to Email	If the information submitted by web form directed to a personal email instead of a server, it is considered as phishing.
Abnormal URL	If the identity is not included in the URL, it considered as phishing.

Table 3.4: Address Based Feature

Feature Name	Feature Explanation
Website forwarding	If redirecting is occurred multiple times, it can be alarming.
Status Bar Customization	"onMouseOver" event can be used to change the status bar of the URL. This can hide the fake URL and show the real URL to trick users. It is considered as phishing if it is applied on the website.
Disabling Right Click	Phishers generally disable right-click function so that users can't check the source code. If the function is disabled in the website, it can be taken as a phishing website.
Pop-up Window	If a web page consists of Pop-up window with a text field, it can be marked as phishing webpage.
IFrame usage	IFrame is used to attach external contents to show in a domain. Phishers might use IFrame tag by hiding them in the website.

Table 3.5: Address Based Feature

Feature Name	Feature Explanation
Age of Domain	If the age of domain is longer than 6 months, we can consider it as a legitimate website as phishing sites tend to live for shorter period of time.
DNS record	If a website doesn't contain any DNS record, it can highly considered as phishing site.
Website Traffic	If a website is visited by huge amount of people, it would have higher ranking. This ranking can help us to identify if a site is phishing or not. A higher ranked website tends to have lower chance of being a phishing website.
PageRank	A value is assigned to a webpage based on its importance. Most of the phishing sites have no pageRank value.
Google Index	If a site has name on the Google Index, we can assume that it is a legitimate website.
Links pointing to Page	A phishing website has shorter lifetime , so it doesn't have much links pointing towards it.
Statistical Reports	If the host of the webpage belongs in any Top phishing IP's or domains, we can count it as phishing webpage.

3.2.3 Data Pre-processing

Database is an important part for this thesis. In the real world, most of the dataset are either incomplete or have some missing values. Although we don't have any missing values in the data set but it is important to standardise data for future implementation in the proposed model. For this reason we didn't have to impute in this paper.

Stratification

We have to split our dataset into two parts while training. First part is for training the data and the second part is testing. Training a portion of your data and then checking with other parts give you more accurate results though splitting dataset shrinks the number of tested values which might reduce the accuracy of the result. In this paper, we use k-fold technique to split our data. The idea behind k-fold is explained here:

1.Shuffle the dataset randomly. The reason for shuffling data randomly is reducing variance and making sure that models remain general and over fitting less.
2.Split the dataset into k groups.

3.For each group:
- Take the group as test dataset
- Take the remaining group as training dataset
- Design a model based on on training dataset
- Evaluate it on test dataset
- Retain the evaluation score and discard the model

4.Summarize the skill of the model and return scores

Chapter 4

Results and Analysis

4.1 Tools

The file was converted in .csv format first. After that, the file was tested with the selective algorithms through weka tool.
Weka is a machine learning tool;used for data-mining, data pre-processing, regression, classification, association rules, visualization and clustering.
For applying LSTM, we integrated "Deeplearning4j" deep learning library with Weka.

4.2 Experimental Results

In our study, Several experiments are implemented based on different scenarios. The experiment and the result were evaluated using several measurements, the performance of several experiments were compared and the results were highlighted.

4.2.1 Evaluation Measures

Accuracy is the rate of correct predictions that the model achieving when

compared with the actual classifications in the dataset. On the other hand, Precision and recall are two evaluation techniques, which calculated based on confusion matrix as shown in Table 4.1 and computed according to Equations 4.1, 4.2 and 4.3:

Precision =TP/(TP + FP)

Recall =TP/(TP + FN)

Accuracy =(TP + TN)/(TP + FP + TN + FN)

Where,

True Positive (TP): The number of correct detected phishing websites.

False Negative (FN): The number of phishing websites was detected as legitimate website.

False Positive (FP): The number of legitimate websites was detected as phishing websites.

True Negative (TN): The number of legitimate websites was detected as legitimate websites.

Table 4.1: Confusion Matrix

	Classified Phishing	Classified Legitimate
Actual Phishing	TP	FN
Actual legitimate	FP	TN

4.2.2 Experimental Results on Feature Selection

In this study, accuracy, recall, precision were calculated for the automated technique. This will comparably evaluate the manual feature selection and automated feature selection. Finally average results were calculated and compared results of the test on all features to design a new model.

Table 4.2: RF,NV Algorithm Accuracy And Precision

Algorithm	Accuracy		Precision	
	Previous Study	This Study	Previous Study	This Study
RF	53.8368	93.1316	53.83	91.6163
NV	53.144	61.1417	98.61	71.5918

The study (Table 4.2) clearly shows that our model give better performance in RF and NV. The reason behind the better result in our proposed model is because the dataset we have used is much bigger than the previous study. However, the best result came from the combination of SMO (Sequential Minimal Optimization), Naive Bayes, Bagging and Multilayer Perceptron algorithm as we saw in the previous paper. In this study, we developed a model in LSTM and compared with it.

The above 30 features are taken as input, that is, the number of input layer nodes in the LSTM network is 30 and the number of output layer nodes is 12. Training network to choose a strong adaptability of the three-layer LSTM network, incentive function is sigmoid function:

$$f(x) = 1/1 + ex(11) \tag{4.1}$$

LSTM neural network for classifying phishing URLs based on LSTM units. When entering a URL into an RNN, one-hot encoding is performed on each URL first. Since the characters composing the URL are all contained in ASCII characters (128 characters in total), each URL becomes one-hot encoded. An input vector with a dimension of (len_of_URL)*128 and then brings the input vector into the RNN. So each input character is translated by an 128-dimension embedding. The translated URL is fed into a LSTM layer as a 100-step sequence. Finally, the classification is performed using an output sigmoid neuron. The learning rate of LSTM neural network is 0.1 .

In order to better illustrate the accuracy of the algorithm in this paper, We compare it with previous study and the results show the results show that LSTM network are better than previous study (table 4.3).

Table 4.3: Accuracy Of LSTM

	Precision	Recall	F-Measure	Accuracy	Dataset size	Cross Validation	Time
This Study	.966	.970	.969	96.55%	11,055	K-fold 10	93.01s
Previous Study	.954	.955	.954	95.49%	466	K-fold 5	.34s

Our proposed model gave better accuracy but took more time than the previous study is because we used the tuned dataset with good amount of data and using 10-fold cross validation instead of 5-fold used in the previous study. Furthermore, according to our LSTM model, the top 10 features are shown in the table 4.4:

Table 4.4: Top 10 Features For LSTM Model

Feature Name	Percentage
SSLfinal_State	66.78%
URL_of_Anchor	45.37%
Prefix_suffix	34.86%
web_traffic	31.60%
having_Sub_Domain	26.19%
Request_URL	25.33%
Domain_registration_length	22.57%
SFH	20.14%
Links_in_tags	16.17%
Google_index	12.89%

In the histogram (figure 4.1) below, we can see the relation between the features and results.

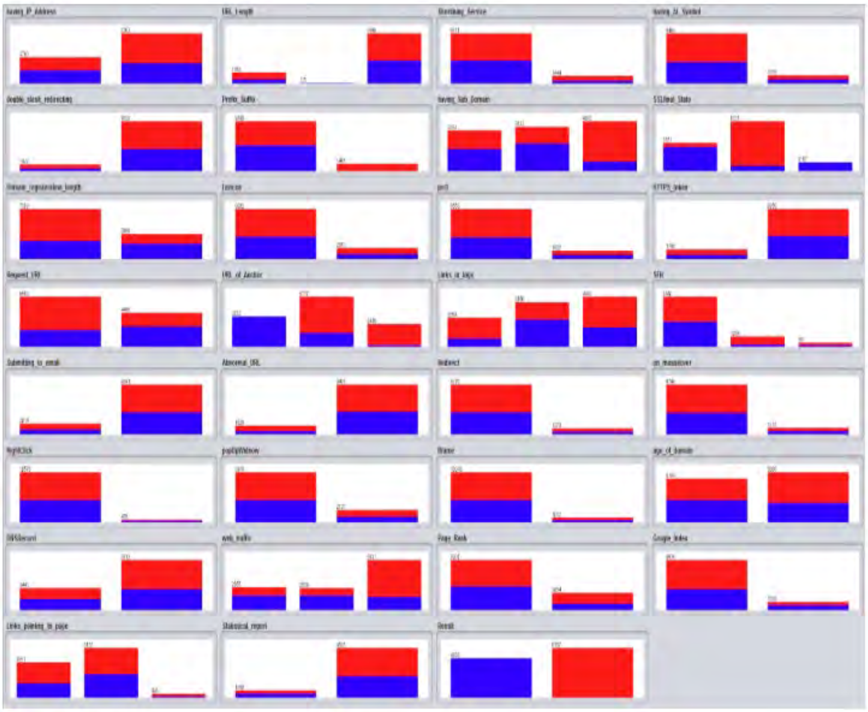

Figure 4.1: Relation Between Features And Results

For evaluating performance of our proposed model, we generate a ROC curve (Figure 4.2). The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.

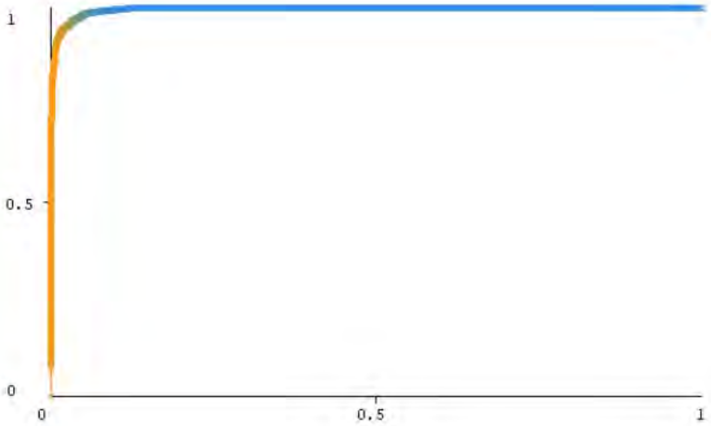

Figure 4.2: ROC Curve Model

The curve follows the left-hand border and the top border of the ROC space which means the test accuracy is quite good enough. Again, our AUC value is .995 which means our classifier is good. AUC is very useful even when there is high imbalanced dataset.

Chapter 5

Conclusion

5.1 Conclusion

Phishing website is a burning issue in recent years. Phishing is a type of attack where attackers send malicious url to the users through social media platform and email. So, detection of that website is necessary. There are many techniques to detect phishing website however there are some limitation too. Apart from those 30 features, there are lots of features that did not cover in this study.

In this study, the accuracy of phishing website detection evaluated based on manual feature selection and automated feature selection on ANN and DNN.

For manual selection, 30 features were selected and grouped in 4 groups (Url based feature, Abnormal based feature, HTML and Javascript based feature and Domain based feature) according to the url structure. The result shows that applying LSTM on those features give the accuracy of 96.55%.

5.2 Future Work

In our future works, we will test the accuracy for all features together excluding one of the four groups each time which will give us a clear picture of the importance of each feature groups and show light for further studies.
0.9

Bibliography

[1] A. Martin, N. Anutthamaa, M. Sathyavathy, M. M. S. Francois, D. V. P. Venkatesan, *et al.*, “A framework for predicting phishing websites using neural networks”, *arXiv preprint arXiv:1109.1074*, 2011.

[2] B. Namasivayam, “Categorization of phishing detection features”, PhD thesis, Arizona State University, 2017.

[3] R. M. Mohammad, F. Thabtah, and L. McCluskey, “Tutorial and critical analysis of phishing websites methods”, *Computer Science Review*, vol. 17, pp. 1–24, 2015.

[4] ——, “Predicting phishing websites based on self-structuring neural network”, *Neural Computing and Applications*, vol. 25, no. 2, pp. 443–458, 2014.

[5] N. L. LEECH and A. J. ONWUEGBUZIE, “A typology of mixed methods research designs”, 2008.

[6] *2018 phishing and fraud report: Attacks peak during the holidays*, https://www.f5.com/labs/articles/threat-intelligence/2018-phishing-and-fraud-report--attacks-peak-during-the-holidays, Accessed: 2019-03-30.

[7] R. Basnet, S. Mukkamala, and A. H. Sung, “Detection of phishing attacks: A machine learning approach”, in *Soft Computing Applications in Industry*, Springer, 2008, pp. 373–383.

[8] O. Viktorov, “Detecting phishing emails using machine learning techniques”, PhD thesis, Middle East University, 2017.

[9] O. M. Maennel and R. Matulevicius, “A new heuristic based phishing detection ap-proach utilizing selenium web-driver”,

[10] *Phishing activity trends report, 3rd quarter 2016*, https://docs.apwg.org/reports/apwg_trends_report_q3_2016.pdf, Accessed: 2019-04-02.

[11] D. Irani, S. Webb, J. Giffin, and C. Pu, “Evolutionary study of phishing”, in *2008 eCrime Researchers Summit*, IEEE, 2008, pp. 1–10.

[12] L. Herløw and S. J. Hansen, “Detection and prevention of advanced persistent threats: Evaluating and testing apt lifecycle models using real world examples and preventing attacks through the use of mitigation strategies and current best practices”, *Denmark: DTU Compute: Department of Applied Mathematics and Computer*, 2015.

[13] F. F. K. Li, “A detailed analysis of an advanced persistent threat malware”, 20111.

[14] V. Ramanathan and H. Wechsler, “Phishgillnet—phishing detection methodology using probabilistic latent semantic analysis, adaboost, and co-training”, *EURASIP Journal on Information Security*, vol. 2012, no. 1, p. 1, 2012.

[15] A. Y. Fu, L. Wenyin, and X. Deng, “Detecting phishing web pages with visual similarity assessment based on earth mover’s distance (emd)”, *IEEE transactions on dependable and secure computing*, vol. 3, no. 4, pp. 301–311, 2006.

[16] M. Hara, A. Yamada, and Y. Miyake, “Visual similarity-based phishing detection without victim site information”, in *2009 IEEE Symposium on Computational Intelligence in Cyber Security*, IEEE, 2009, pp. 30–36.

[17] D. Zhang, Z. Yan, H. Jiang, and T. Kim, “A domain-feature enhanced classification model for the detection of chinese phishing e-business websites”, *Information & Management*, vol. 51, no. 7, pp. 845–853, 2014.

[18] Y. Cao, W. Han, and Y. Le, “Anti-phishing based on automated individual white-list”, in *Proceedings of the 4th ACM workshop on Digital identity management*, ACM, 2008, pp. 51–60.

[19] P. Kumaraguru, Y. Rhee, A. Acquisti, L. F. Cranor, J. Hong, and E. Nunge, “Protecting people from phishing: The design and evaluation of an embedded training email system”, in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 2007, pp. 905–914.

[20] W. Ali, “Phishing website detection based on supervised machine learning with wrapper features selection”, *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 9, pp. 72–78, 2017.

[21] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, “An empirical analysis of phishing blacklists”, 2009.

[22] G. V.Preethi, “Automated phishing website detection using url features and machine learning technique”, 2016.

[23] *The random forest algorithm*, https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd, Accessed: 2019-03-08.

[24] *How the random forest algorithm works in machine learning*, http://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/, Accessed: 2019-02-12.

[25] D. D. Lewis, “Naive (bayes) at forty: The independence assumption in information retrieval”, in *European conference on machine learning*, Springer, 1998, pp. 4–15.

[26] *Phishing websites data set*, https://archive.ics.uci.edu/ml/datasets/phishing+websites, Accessed: 2019-01-03.