Comparison of the performance of the learning algorithms for verification of phishing uniform resource locator (URLs) using machine learning

By

Kephas Loide M
201401571

Main Supervisor:

Mr. Gabriel Tuhafeni Nhinda

Co-supervisor:

Dr. Nalina Suresh

A Research Proposal submitted in partial fulfillment of the requirements for the degree of
Bachelor of Science in Information Technology
University of Namibia

Dated (11.June. 2018)

Table of Contents

1. Introduction

In this section, we will discuss the background of the study, the problem of the statement, the objective of the study, motivation, and significance of the study.

1.1. The background of the study

Phishing is an online criminal act that occurs when a malicious webpage mimics as a legitimate webpage so as to acquire sensitive information from the user [1]. Detecting phishing websites is one of the crucial problems facing the internet community because of its high impact on the day-to-day online transactions performed. There is no doubt that phishing, as a phenomenon, is both highly successful and generally difficult to detect and prevent in a reasonable amount of time [2]. Furthermore, detection of phishing URLs has become increasingly difficult due to the evolution of phishing operations and the efforts to avoid mitigation by blacklists. The current state of cybercrime has made it possible for a phisher to host operations with short lifecycles that diminish blacklist effectiveness [3]. Moreover, a phisher uses social engineering and technical deception to fetch private information from the web user. The phishing web pages generally have alike page layouts, blocks and fonts to mimic legitimate web pages in an endeavor to influence web users to obtain personal details such as username and password. Over the few years, online baking has become very popular as more financial institutions have begun to offer free online services [4].

1.2. Problem Statement

Namibia has experienced its own share in cyber-attacks in the ream of electronic banking transactions which prompt the Namibian government to come up with a draft bill on electronic transactions and cybercrime [5]. However, due to rapidly evolving technologies, this regulation needs to be drafted with flexibility in mind, by taking into account the need for legal certainty and precision. Furthermore, despite a number of solutions to mitigate phishing by previous researchers, there is still no conclusive solution to phishing attacks particularly in the Universities environment, and University of Namibia (UNAM) is not an exception. Currently, at UNAM, there is Cyberoam in place, but little study has been done on the performance study. Hence the mail server still receives phishing URLs and Information Technology (IT) infrastructure department staff, had to warn the users not to open such emails. Therefore, there is a needs to find a solution to mitigate Phishing URLs.

1.3. Objectives of the study

The main objective of this study is to evaluate the performance of learning algorithms for verification of phishing URLs using machine learning techniques.

The following are the sub-objectives of the study:

1. To identify the lexical features of malicious URL.
2. To determine which of the three algorithms is best for determining phishing URLs.
3. To apply an appropriate standard dataset to test the three learning algorithms on.
4. To use standard metrics for measuring the performance of the learning algorithms (Naïve Bayes classifier, Decision Tree and Logistic Regression model) to benchmark the performance of the algorithms
5. To verify whether a URL is a legitimate or phishing URLs.

1.4. Motivation

The researcher herself had received spam emails claiming she won a huge amount of money and to follow some unsecured links. Furthermore, the researcher also found out that staff members at UNAM had been receiving emails from IT infrastructure department about phishing URLs. So the researcher believes that suspicious URLs in the cyberspace continue to pose a threat, which gives room to improve the existing proposed techniques in order to find a conclusive solution to phishing URLs.

1.5. The significance of the study

The study will provide a better understanding on two or more machine learning algorithms that could be employed to verify and confirm compromised and phishing URLs in the cyberspace.

1.6. Limitations of the study

In this study, only three learning algorithms will be trained and tested basically: Naïve Bayesian, Decision Tree and Logistic Regression and they will only be trained and tested based on lexical features. Furthermore, all the data will come from a single source which is UNAM computer center.

2. Literature review

This section provides an overview on some of the major studies conducted on phishing URL and the algorithms to detect phishing URLs:

Basnet and sung [6], proposed a novel approach for classifying legitimate malicious URLs using supervised learning across the features from various web services. They applied the web mining-based heuristics on logistic regression classifier, and demonstrate that Logistic Regression can detect phishing URLs with an accuracy of 99%. However, the content-based approach requires access to the phishing site. Moreover, the heuristic can still be integrated with a keyword, lexical, host and content-based features to improve phishing URLs detection.

Azeez and Oluwatosin [7], explored how malicious link in emails can be detected from lexical and host-based features of their URLs to protect users from identity theft attacks using Naïve Bayesian classifier. However, even though Naïve Bayesian was the best for their approach, more classifier or algorithms could have been used to enhance their findings.

E and K [8], proposed a system that uses lexical features WHOIS features, PageRank and Alexa Rank and PhishTank-based features on random forest and content-based algorithms to classify phishing URLs. They demonstrated that by applying web mining heuristics on random forest algorithms, a precision of more than 90% was achieved and force negative rate (FNR) and force positive rate (FPR) rates of less than 1%. However, more improvement needs to be done on the content-based algorithm as only less than 65% precision was achieved. Moreover, there is a need to work on a selection of more features for the content-based algorithm to increase the precision and decrease the FNR and FPR.

Ma et al. [9] explore an online learning approach for classifying URLs automatically as either malicious or benign, based on supervised learning across both lexical and host-based features. However, their approach is complementary to blacklisting which cannot predict the status of a previously unseen URLs.

Blum et al.[3], explore the possibility of utilizing confidence weighted classification combined with content-based phishing URL detection to produce a dynamic and extensive system for detection of present and emerging type of phishing domains based on lexical features only. However, more features could be experimented using different learning algorithms to add to the value of the models and improve the accuracy.

Despite the amount of research done by the previous researcher, there still no definitive solution to phishing problem. Hence, there is a need to improve the suggested methods by the previous researchers to find a conclusive solution.

In this study, the researcher will carry out extensive and comprehensive comparative analysis of three learning algorithms namely Naïve Bayes, Decision Tree and Logistic regression for verification of compromised, suspicious and phishing URLs, and to determine which is the best of all, based on the matrices (F-Measure, Precision and Recall) used for evaluation.

3. Methodology

3.1. Research Methodology

In this section, the researcher will discuss the research design to be used in this study, the sampling process, the procedure on how data will be collected and analyzed and the software development methodology that will be used (requirements, planning, iteration initialization, Design, Implementation, System testing and Retrospective).

3.1.1. Research Design

In order to meet the objective of this study, a quantitative research design with experimental as an approach will be used. Experimental research approach will be used during training and testing of data, on the performance of the three learning algorithms based on lexical features using machine learning techniques. According to [10], Lexical features are items of data selected from the URLs that allow us to capture this observable difference between the appearance of a legitimate URLs and that of phishing URLs. Machine learning is a set of techniques that allow implementing adaptive algorithms to make predictions and to auto-organize input data according to their common features [11].

3.1.2. Population and Sample

A standard dataset consisting of legitimate and phishing URLs will be there targeted population for this study. Furthermore, a purposive sampling will be used to select the sample of the study under discussion. Purposive sampling is a sampling technique that deliberately hand-picks the sample by choosing instances that are likely to produce variable data to meet the purpose of the research [12]. This study will be interested in URLs both phishing and legitimate URLs. Therefore, a purposive sampling will be used for this study to select the sample.

3.1.3. Procedures

A standard dataset will be collected from the University of Namibia mainly Computer center, Information Technology (IT) Infrastructure department [14]. Furthermore, the data will be sorted into a standard dataset that will contain both legitimate and phishing URLs. Additionally, the dataset will be applied to the learning algorithms such as Naïve Bayesian, Decision Tree, and Logistic Regression, to evaluate the performance of the learning algorithms by comparing the result of the experiment, to determine whether this learning algorithm can classify the URLs as a legitimate or phishing URLs. Furthermore, a comparative analysis of the performance of learning algorithms will be made for verification of vulnerable and compromised URLs after the experiment.

3.1.4. Data Analysis

In this study, the researcher will compare generated results of the three learning algorithms and interpret the evaluation of three algorithms sing a confusion matrix to provide empirical evidence to support this study. Moreover, by evaluating the performance of each algorithm, their overall accuracy will be compared in order to determine the suitable algorithm in detection phishing URLs. The following table 1 illustrates the confusion matrix to benchmark two algorithms table.

Predicted			
Actual		Negative	Positive
	Negative	TN True Negative	FP False Positive
	Positive	FN False Negative	TP True Positive

Table 1: Confusion Matrix table.

In this section, we focus on the performance of the binary classification model. Based on confusion matrix, corresponding four categories of shortcoming prediction result are described as follows:

- True Positive (TP, collect classified phishing URLs)
- True Negative (TN, collect classified phishing URLs)
- Force Positive (FP, non-phishing URLs wrongly classified as phishing), and
- Force Negative (FN, phishing URLs wrongly classified as non-phishing).

3.2. Software Development Methodology

Personal Extreme programming (PXP) development methodology will be applied in this study to design a prototype. According to [13], PXP is a software development process designed to be applied by software engineers individually. Furthermore, PXP development process is iterative and applying its practices allows the developer to be more flexible and responsive to changes. The following figure 1 gives an overview of the PXP development methodology process.

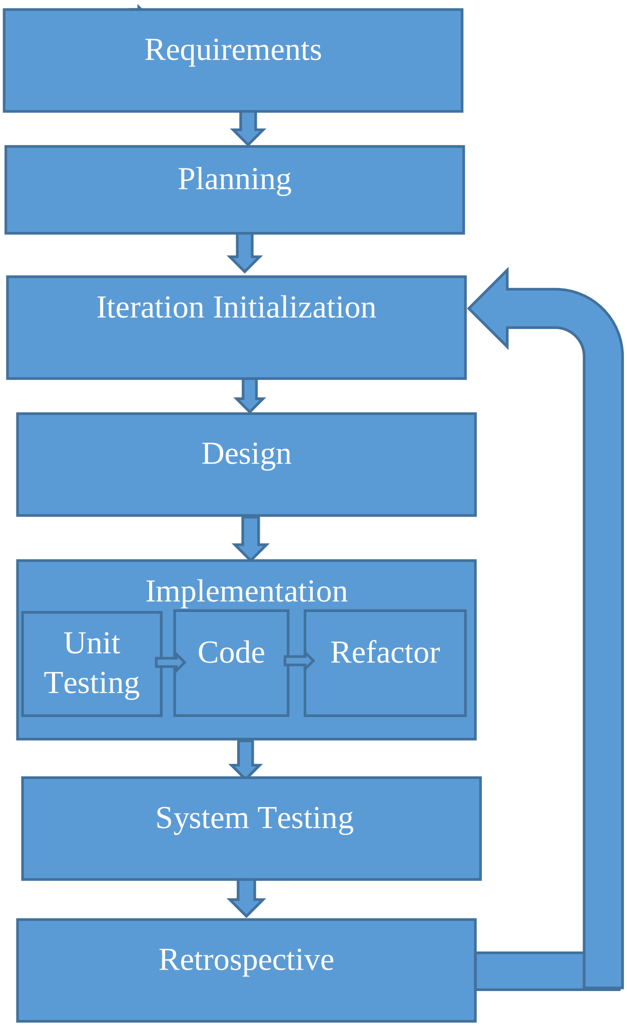


Figure 1: Overview of personal Extreme Programming (PXP) process.

3.2.1. Requirements

Hardware Requirements

- Computer

Software Requirements

- Anaconda3
- Python3
- Jupyterlab Notebook

3.2.2. Planning

In the planning phase, the researcher put together a set of tasks based on the objectives:

1. Load the dataset and prepare them for training and testing.
2. Choose the type of machine learning algorithm to use.
3. Build an analytical model based on a chosen algorithm.
4. Train the model on the dataset and revising it as needed.
5. Runs the model to generate scores and findings.

3.2.3. Iteration Initialization

Iteration initialization indicates the beginning of each iteration and starts with tasks selection, which will be the focus of the iteration. In this study, the iteration will begin by loading the dataset file.

3.2.4. Design

During the design phase, the developer will model the system modules and classes that will be implemented in the ongoing iteration. The developer will have features that will be computed and used to try and predict phishing URLs and this will be the X data and Y data which will be the URLs. In order to evaluate the three algorithms in a scientific manner, the data will be split into two section: training and testing section in order to have out of sample testing.

The developer will take the X and Y training data and run that through the machine learning algorithm which will be either Naïve Bayesian, Logistic Regression or Decision Tree to generate a model. The developer will then test the performance of that model using X testing data, and the outcome will be compared to the Y testing, which is the ground truth. The following figure 2 shows a model diagram of the system.

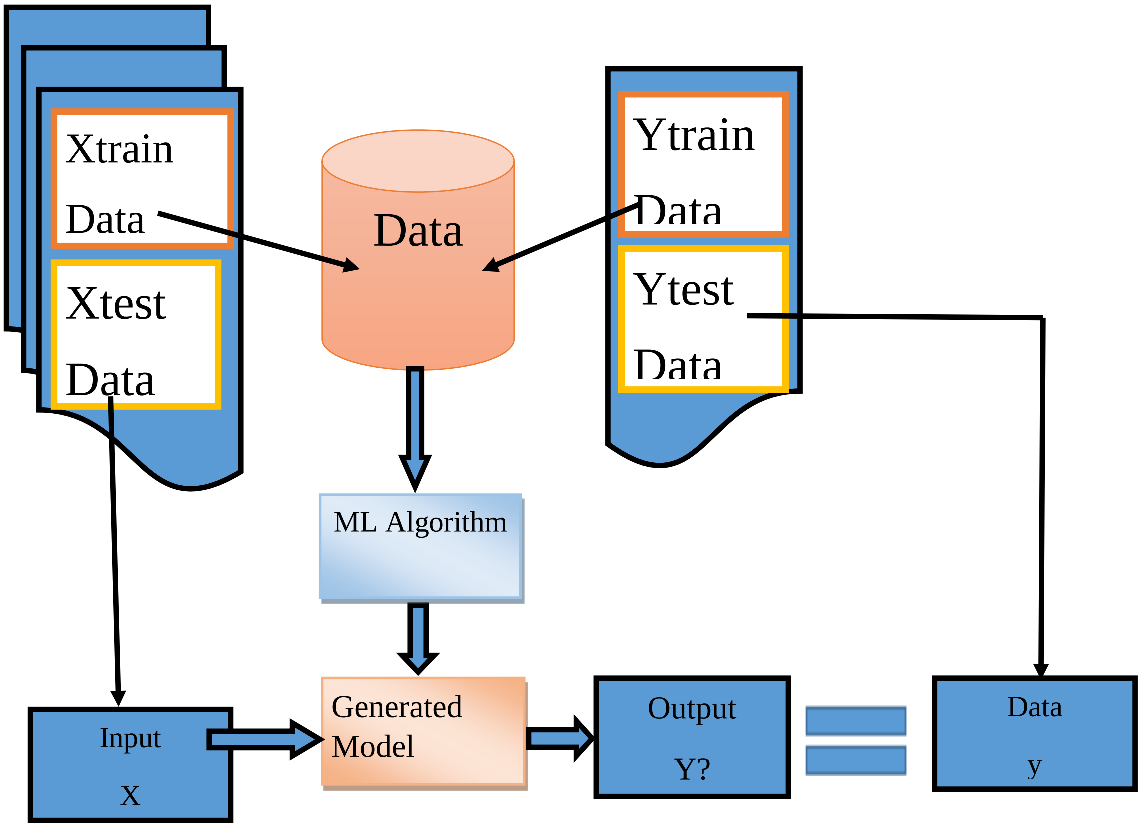


Figure 2: Shows a flow diagram for the system in the designing phase.

3.2.5. Implementation

During the implementation phase, the developer will implement all the objects defined in the previous design phase and tests them.

3.2.6. System Testing

During system testing, all features developed will be tested in this phase, and verify whether the implemented solution meets the initial project requirements. All errors that will be found will be recorded and removed.

3.2.7. Retrospective

Retrospective phase marks the end of the process iteration, which could end either in a product that could be a final product or in a final product. Furthermore, it could start a new iteration by moving to Iteration Initialization phase, or mark the end of the project development when all project requirements are met, and there are no remaining errors.

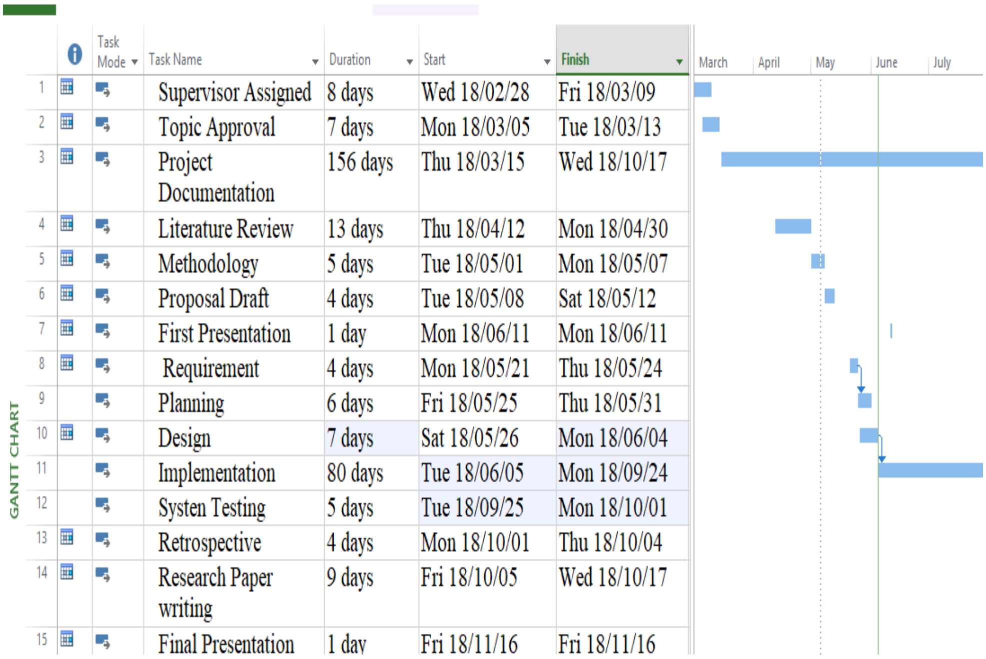
In order to use these software development methodology described in this study, the researcher will make use of the dataset collected in the research methodology during the procedure.

4. Research Ethics

The participants that will be involved in this research, will be asked first for their consent before any data will be collected for this study. Moreover, the data will not be shared with any third party and it will only be used for the purpose of this study.

5. Project Schedule

The following figure 4 shows the project schedule for this study:



	Task Mode	Task Name	Duration	Start	Finish
1		Supervisor Assigned	8 days	Wed 18/02/28	Fri 18/03/09
2		Topic Approval	7 days	Mon 18/03/05	Tue 18/03/13
3		Project Documentation	156 days	Thu 18/03/15	Wed 18/10/17
4		Literature Review	13 days	Thu 18/04/12	Mon 18/04/30
5		Methodology	5 days	Tue 18/05/01	Mon 18/05/07
6		Proposal Draft	4 days	Tue 18/05/08	Sat 18/05/12
7		First Presentation	1 day	Mon 18/06/11	Mon 18/06/11
8		Requirement	4 days	Mon 18/05/21	Thu 18/05/24
9		Planning	6 days	Fri 18/05/25	Thu 18/05/31
10		Design	7 days	Sat 18/05/26	Mon 18/06/04
11		Implementation	80 days	Tue 18/06/05	Mon 18/09/24
12		Systen Testing	5 days	Tue 18/09/25	Mon 18/10/01
13		Retrospective	4 days	Mon 18/10/01	Thu 18/10/04
14		Research Paper writing	9 days	Fri 18/10/05	Wed 18/10/17
15		Final Presentation	1 day	Fri 18/11/16	Fri 18/11/16

Figure 4: Schedule for this research.

6. References

[1] S. Marchal, J. François, R. State, and T. Engel, “PhishStorm: Detecting Phishing With Streaming Analytics,” *IEEE Trans. Netw. Serv. Manag.*, vol. 11, no. 4, pp. 458–471, Dec. 2014.

[2] P. D. Dudhe and P. L. Ramteke, “Detection of Websites Based on Phishing Websites Characteristics,” vol. 3, no. 4, p. 7, 2007.

[3] A. Blum, B. Wardman, T. Solorio, and G. Warner, “Lexical feature based phishing URL detection using online learning,” 2010, p. 54.

[4] S. C. Jeeva and E. B. Rajsingh, “Intelligent phishing URL detection using association rule mining,” *Hum.-Centric Comput. Inf. Sci.*, vol. 6, p. 10, Jul. 2016.

[5] N. E. S. Reporter, “Cybercrime a threat to national security – Shanghala,” *New Era Newspaper Namibia*, 15-May-2018.

[6] R. B. Basnet and A. H. Sung, “Mining Web to Detect Phishing URLs,” in *2012 11th International Conference on Machine Learning and Applications*, 2012, vol. 1, pp. 568–573.

[7] N. A. Azeez and A. Oluwatosin, “CyberProtector: Identifying Compromised URLs in Electronic Mails with Bayesian Classification,” in *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2016, pp. 959–965.

[8] B. E. and T. K., “Phishing URL Detection: A Machine Learning and Web Mining-based Approach,” *Int. J. Comput. Appl.*, vol. 123, no. 13, pp. 46–50, Aug. 2015.

[9] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Beyond blacklists: learning to detect malicious websites from suspicious URLs,” 2009, p. 1245.

[10] O. Youle, “Online Lexical Phishing URL Classification,” p. 118.

[11] M. Mohammed, M. B. Khan, and E. B. M. Bashier, *Machine Learning: Algorithms and Applications*. CRC Press, 2016.

[12] B. J. Oates, *Researching Information Systems, and Computing*. SAGE Publications, 2006.

[13] Y. Dzhurov, I. Krasteva, and S. Ilieva, "Personal Extreme Programming – An Agile Process for Autonomous Developers," p. 8.

[14] A. Bratha, Infrastructure department, University of Namibia Computer centre, 2018.