ANALYSIS AND DETECTION OF E-MAIL PHISHING USING PYSPARK

[1]RISHIKESH B H, [2]SHREEHARI A S, [3]SRIHARSHA G S, [4]SUNIL HUDAGE, 5REKHA K S

[5]*Department of Computer Science and Engineering*
THE NATIONAL INSTITUTE OF ENGINEERING(NIE), Mananthavady Road, Mysuru – 570008, Karnataka, India

---

Abstract - *Phishing is an act that attempts to steal information, personal data by using spoofed emails and fraudulent web sites to trick people into giving up personal information. Phishing E-mails involve malware links and is totally committed to obtain sensitive & valuable information. Phishing has become more and more complicated and sophisticated and attack can bypass the filter set by anti-phishing techniques. Phishing impact ranges from denial of access to e-mail to substantial financial loss, resulting loss of public's trust in internet. We provide robust method to detect phishing E-mails which performs some cross-validations techniques. The method includes Text Analysis, Link Analysis to encounter phishing countermeasures. Educational materials reduced user's tendency to enter information into phishing webpages.*

Key Words: **Naïve Bayes, PySpark, Big Data, Link Analysis, Machine Learning, Virus Total.**

1. INTRODUCTION

Security is a key aspect in the field of information and communication technology. Information security bears great value to personal as well as corporate sectors. different companies and organizations need to protect their customers and employee's information related to business plans, financial outcomes, product information, and the like [1]. Phishing is one of the luring techniques used by phishing artist in the intention of exploiting the personal details of unsuspected users. Phishing website is a mock website that looks similar in appearance but different in destination. The unsuspected users post their data thinking that these websites come from trusted financial institutions. Big data refers to an enormous amount of dataset that is able to expose patterns associated with human interaction through computational analysis [2]. The main purpose here is to detect the e-mails which user receives is legitimate or not. The goals of our paper as well as system are: (a) To provide security (b) Accuracy in detection.

Recently, Govt of India issued alert on spread of Locky Ransomware which is being spread through e-mail phishing. There are three fundamental attributes of email security – Confidentiality, Integrity and Availability [3].

2. SOFTWARE DESCRIPTION

2.1 PySpark

The Spark Python API (PySpark) exposes the Spark programming model to Python. To support Python with Spark, Apache Spark community released a tool, PySpark. Using PySpark, you can work with RDDs in Python programming language also.

At a high level, every Spark application consists of a driver program that runs the user's main function and executes various parallel operations on a cluster. A second abstraction in Spark is shared variables that can be used in parallel operations.

2.2 Natural Language Toolkit (NLTK)

The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language. NLTK includes graphical demonstrations and sample data.

NLTK is intended to support research and teaching in NLP or closely related areas, including empirical linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning. NLTK has been used successfully as a teaching tool, as an individual study tool, and as a platform for prototyping and building research systems.

2.3 Virus Total

The Virus Total API lets you upload and scan files or URLs, access finished scan reports and make automatic comments without the need of using the website interface. In other words, it allows you to build simple scripts to access the information generated by Virus Total.

2.4 Naïve Bayes

The Naive Bayes classifier is designed for use when features are independent of one another within each class, but it appears to work well in practice even when that independence assumption is not valid.

It classifies data in two steps (a) Using the training samples, the method estimates the parameters of a probability distribution, assuming features are conditionally independent given the class. (b) For any unseen test sample, the method computes the posterior probability of that sample belonging to each class.

3. LITERATURE SURVEY

Tarnnum et al., [4] have conducted two studies to observe the security threats for big data: (a) first study was carried out on Enron email dataset (that contains about half a

million of emails) to investigate the security challenges of big data in the field of email communication; and (b) second study was carried out on 35 undergraduate students to observe how phishing email generation based on users' intention or behavior may break the security system

Liping Ma et al., [5] has presented an approach to detect phishing e-mails using hybrid features and have presented a method to build a robust classifier to detect phishing emails using hybrid features and to select features using information gain, experimented on 10 cross-validations to build an initial classifier which performs well. The experiment also analyses the quality of each feature using information gain and best feature set is selected after a recursive learning process. Experimental result shows the selected features perform as well as the original features.

Sa'id Abdullah Al-Saaidah., [6], through this research, varied classification algorithms are discussed and compared, such as; Naïve-Bayes, Decision Tree (DT), Logistic Regression, Classification and Regression Trees and Sequential Minimal Optimization (SMO). The experiment was executed using WEKA Tool on a dataset of 4800 Email, 2400 phishing emails and 2400 legitimate emails represented the 47 features of the email structure

4.SYSTEM ANALYSIS

4.1 Existing System

Anti - Phishing using Machine Learning: This software is the normal "Network Security Filter that stops you from visiting suspicious websites with a Twist". Their software will never update. The existing system only detects spam e-mails and put into the spam folder. This is done by noticing a domain constantly sending spam messages and blacklist such sender. Phishing Domain Detection with Machine Learning: Uniform Resource Locator (URL) is created to address web pages which results in time inefficient and Phishers are intelligent to bypass the barriers.

4.2 Proposed System

PySpark is used to implement Naïve-Bayes algorithm which is fast in execution time compared to normal machine learning model. Link analysis is done using web page extraction & to check if any malicious or phishing contents are there in the link. To improve the accuracy of link analysis, we are using Virus Total API to detect any phishing sites.

5. METHODOLOGY

The two main operations performed are Text Classification and Link Analysis. Text Classification includes data cleaning, data preprocessing, bag of words model and Naïve-Bayes classifier. Link Analysis includes page extraction and using Virus Total API. The steps involved are:

a) Gathering the datasets, we have collected 48000 emails from different dataset available.

b) Data cleaning, to clean the gathered emails in the above step and convert into tsv files.

c) Data preprocessing, before text classification we are creating bag of words model and applying count vectorizer to hash the words.

d) Using Naïve-Bayes algorithm, we classify whether the texts are spam or ham.

e) We extract links present in E-mails and apply Link Analysis which comprises of 2 steps:

1) Extract the contents of web pages or check if any form is present and whether it asks for personal information.

2) Virus Total API is used to check for presence of phishing links.

5.1 SYSTEM DESIGN

Phishing attack and detection system is broken down into sub-modules like web portal, database, personal information, graphical statistics and e-mail phishing detection system

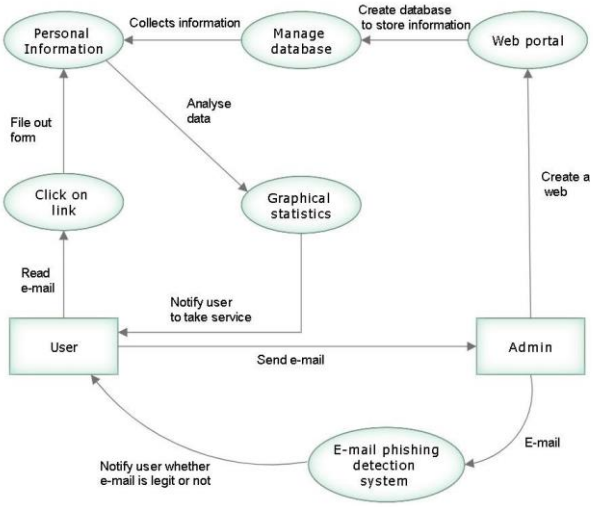


Fig. 1 Sub-modules of Phishing Detection System

The E-mail phishing detection system is broken down into several processes like link analysis, Naïve-Bayes classifier and resulting probability value is compared with the threshold probability value.

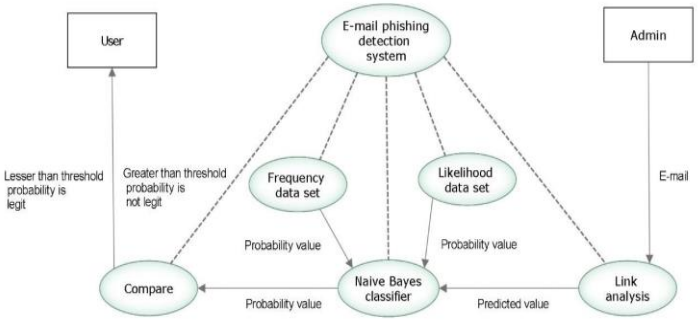


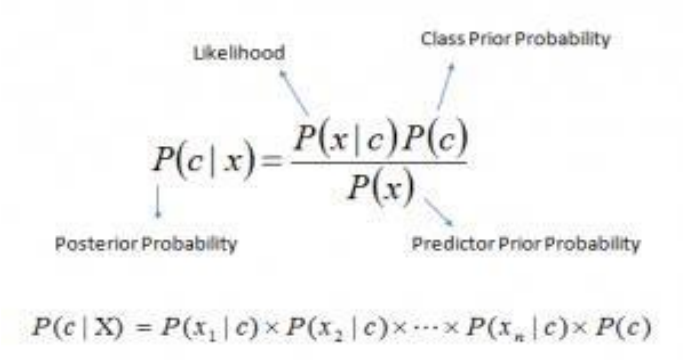
Fig. 2 E-mail Phishing Detection System

6. IMPLEMENTATION

6.1 Naïve-Bayes Algorithm

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below [9]:



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

.... [9]

Above,

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

7. CONCLUSIONS AND FUTURE ENHANCEMENTS

Phishing is a form of criminal conduct that poses increasing threats to consumers, financial institutions, and commercial enterprises. Because phishing shows no sign of abating, and indeed is likely to continue in newer and more sophisticated forms, law enforcement, other government agencies, and the private sector in both countries will need to cooperate more closely than ever in their efforts to combat phishing, through improved public education, prevention, authentication, and bi-national and national enforcement efforts.

Using Big Data analytics to detect phishing e-mails is developed in response to the increased threat posed by malicious e-mails that closely resemble legitimate ones. This methodology not only helps detect phishing messages but also makes it easier to detect such phishing messages even if they more closely mimic legitimate ones. This would, however, not be possible without knowledge of big data and previous knowledge of current threats. Detection of download links to files and improving the accuracy of detection

REFERENCES

[1] Every company needs to have a security program (2008) [Online]. Available: https://www.appliedtrust.com/resources/security/everycompany-needs-to-have-a-security-program

Oxford, "Big data," in Oxford Dictionary, Oxford University Press, 2016. [Online].Available: http://www.oxforddictionaries.com/ definition/English/bigdata

[2] P. Cocca. "Email security threats," SANS Institute, USA, Rep. Version 1.4b Option 1, pp. 1-16. Sept. 20,2004

[3] Tarnnum Zaki, Md. Sami Uddin, Md. Mahedi Hasan, "Security Threats for Big Data", IEEE-2017.

[4] Liping Ma, Paul Watters, Simon Brown, "Detecting Phishing E-mails Using Hybrid Features", workshop on Ubiquitous, Autonomic and Trusted Computing.

[5] Sa'id Abdullah Al-Saaidah, "Detecting Phishing E-mails Using Machine Learning", MEU-2017.

[6] J. Crowe. (2016). Phishing by the numbers: Must-know. Phishing statistics 2016 [Online]. Available: https://blog.barkly.com/phishingstatistics- 2016

[7] New EDRM Enron email dataset (n.d.) [Online]. Available: http://spamassassin.apache.org/old/publiccorpus/

[8] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[9] Ron Zacharski, "A Programmer's Guide to Data Mining", a book which gives detailed description of Naïve-Bayes Algorithm and unstructured text.

[10] Steven Bird, Ewan Klein and Edward Loper, "Natural Language Processing with Python", a book on NLTK toolkit for Hadoop.