International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 12, December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

Spam Detection in Email using Bayesian Classifier and support vector machine

Sumathi.k [1],Dr.Radha Damodaran[2]

[1]Research sholar, [2]Associate Professor,

Department of computer science ,CMS college of science and commerce,

Chinnavedampatty, Coimbatore, TN-49

ABSTRACT

Phishing is considered one of the crucial security challenges for the online community due to the massive numbers of online transactions performed on a daily basis. Many classification techniques have been used and devised to combat phishing threats, but none of them is able to efficiently identify web phishing attacks due to the continuous change and the short life cycle of phishing websites.

Thus, our work is to detect and filter such kinds of emails. For this purpose, we have used Bayesian classifier and support vector machine together in this work. Considering the feature: URL lexical feature, source code feature, and network feature.

Keywords: Phishing, Classification Techniques, Bayesian Classifier, Support Vector Machine, URL Lexical Feature.

I. INTRODUCTION

Phishing websites are created by dishonest individuals to imitate genuine websites. These websites have high level of visual similarities to the legitimate ones in an attempt to defraud honest internet - users. Phishing constitutes more than half of all reported security incidents on the internet. Approximately 3.6 million clients in the US alone had lost money to phishing attacks and total losses had reached approximately US $3.2 billion dollar. Phishing attacks are increasing at a rapid rate. The number of victims increased from 2.3 million in 2006 to 3.6 million in 2007, an increase of 56.5%. This become a serious problem not only because of the increased number of these websites but also the intelligent strategies used to design such websites.[1-5]

II. HEADINGS

1. EXISTING WORK

The existing work undergone an implementation on “content based spam detection in email using Bayesian classifier” by Smith B. Rathod and Tareek M. Pattewar. They used Bayesian classifier to evaluate in terms of accuracy, error, time, precision and recall for email classification and detection of spam mails.[1]

And "malicious URL detection" by Christophe Chong [Stanford], Danial Lin [Stanford] and Wonhong Lee [Neustar] using Support Vector Machine. Here , they have contributed to the creation of a realtime malware classifier blocking out the malicious URLs.

2.ALGORITHM STUDY

A. BAYESIAN CLASSIFIER

It is a statistical classifier known for its email filtering. Text classification metod is used for identifying spam mails. It uses tokens(words) with spam and ham mails to calculate the probability of mail being spam or not.

B. SUPPORT VECTOR MACHINE

It has been recently proposed by Dr. V. Vapnik as an effective statistical learning method for pattern recognition. SVM can be used to solve both linearly separable and non linearly separable problems.

3. EXPERIMENT

A. EVALUATION CRITERIA:-

We formulate the spam detection problem using Bayesian classifier and support vector machine. Here, each mail undergoes all of the three feature checking which are as follows.

- URL lexical feature,
- Source code feature,
- Network feature.

a. URL lexical features:-

The URL is approached as an NLP problem. Here, term frequency_inverse document frequency(tf_idf) to weigh the importance of a token in the URL is used. URL include both the domain and the path. Thus, it can be defined as

$$tf\_idf = tf(t; d) \_ idf(t;D)$$

where,

$$tf(t; d) = \frac{f(t; d)}{\max\{f(w; d) : w \in d\}}$$

$$idf(t;D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 12, December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

b. SOURCE CODE FEATURE:-

Javascript exploits to prevent detection by automated or manual analysis. Following is the exploitative script we found in our malicious sample.

{ k =i ; s + = String [" fro "+

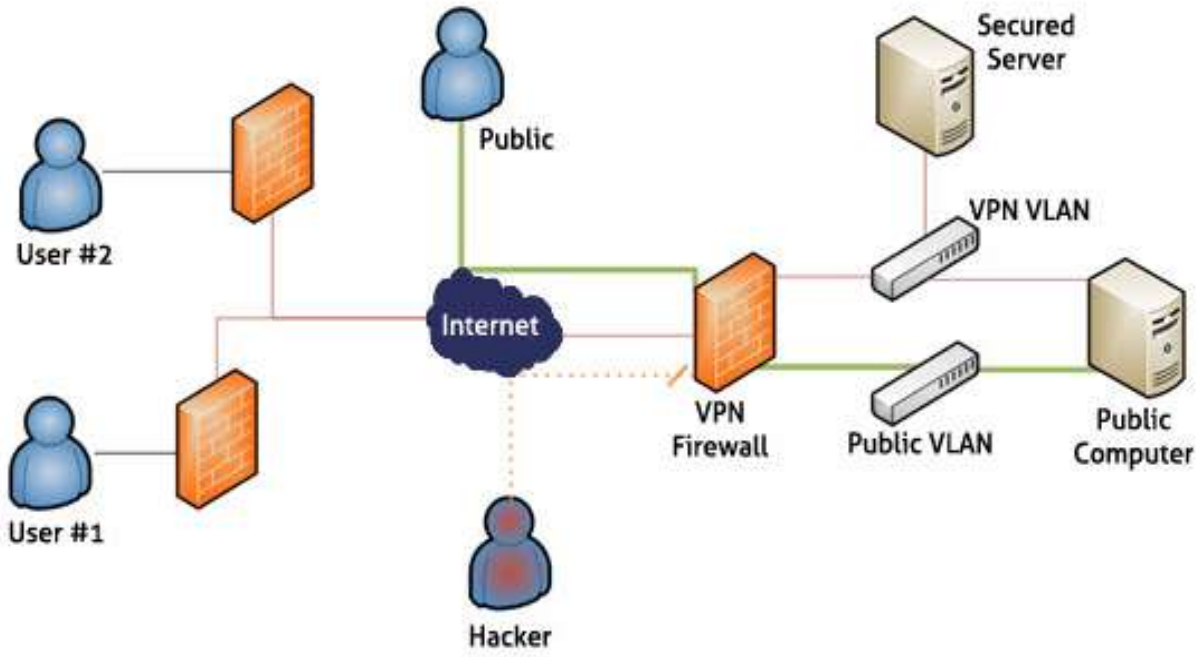
" mCh " + " arCode "] (n [k] /

(i – h * Math [f] (i / h) + 016)) ; }

If (018 – 0 x f = = = 3) eval (s) ; }

c.NETWORK FEATURE:-

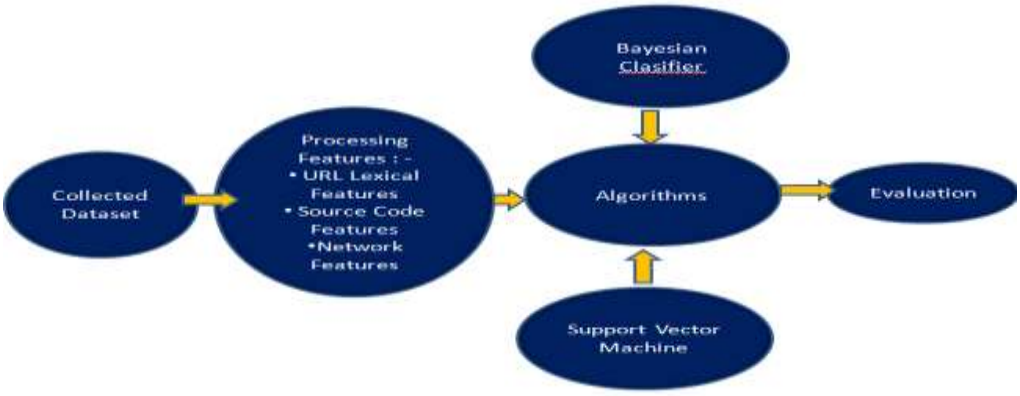
A variety of network feature such as DNS query data, latency, payload size and domain registry data are explored.



International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 12, December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

IV. FIGURES AND TABLES

EXPERIMENT :-



EVALUATION

Bayesain Classifier	Feature 1	Feature 2	Feature 3
Data set 1	93.98	95.23	97.56
Data Set 2	94.85	87.32	90.00
Data Set 3	96.45	94.21	89.12

SVM	Feature 1	Feature 2	Feature 3
Data set 1	95.21	95.23	97.56
Data Set 2	98.1	80.34	90.00
Data Set 3	96.45	94.21	80.50

V. CONCLUSION

We have emphasized on both Bayesian classifier and Support vector machine for classifying spam and legitimate mails using supervised learning across features extracted.

Applying this , we experimentally demonstrated that spam mails can be detected with an accuracy of about 98.36% with respect to real world gmail data sets.

VI. ACKNOWLEDGEMENT

We sincerely thank each and every person who help us through this work to make it successful.

REFERENCE

[I] Dhanalakshmi Ranganayakulu and Chellappan C., "Detecting malicious URLs in E-Mail - An implementation", AASRl Conference on intelligent Systems and Control, Vol. 4 ,2 013,pg. 125-131

[2] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk e-mail ... AAAiTech.Rep.WS-98-05. pp.55-62,1998.

[3] V Christina., "A study on email spam filtering techniques", international Journal of Computer Applications, Vol. 12- No.1, 2010.

[4] Sadeghian, A and Ariaeinejad, R. , "Spam detection system: A new approach based on interval type-2 fuzzy sets", iEEE CCECE -000379,2011.

[5] Congfu Xu, Yafang Chen, Kevin Chiew, "An approach to image spam filtering based on base64 encoding and N-Gram feature extraction", iEEE international Conference on Tools with Artificial intelligence, DOI I0.1109/ICTAl.2010.31,201

[6] Antonakakis, Manos. ”Kopis: Detecting Malware Domains at the Upper DNS Hierarchy.” http://static.usenix.org/ events/sec11/tech/slides/antonakakis.pdf.

[7] Bilge, Leyla, Engin Kirda, Christopher Kruegel, Marco Balduzzi. ”EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis.” http://www.syssec-project. eu/media/page-media/3/bilge-ndss11.pdf.

[8] Choi, YoungHan, TaeGhyoon Kim, SeokJin Choi . Automatic Detection for JavaScript Obfuscation Attacks in Web Pages through String Pattern Analysis. http://www.sersc. org/journals/IJSIA/vol4 no2 2010/2.pdf.

Books :

Neural network and learning machines,

Network security.