JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

EMAIL SPAM DETECTION USING MACHINE LEARNING

K.Aparna[1], P.Asmitha[2], V.Anil Kumar[3],B.Dhanumjay[4],Ch.Gowri[5]
[1,2,3,4]**Students,** Dept of ECE, Godavari Institute of Engineering and Technology (A),Rajahmundry,AP
[5]**Assistant Professor,** Dept of ECE, Godavari Institute of Engineering and Technology(A),Rajahmundry,AP

Abstract-Email is widely used for all sorts of communication nowadays, from personal to professional. Sensitive information, such as passwords, credit card numbers, and bank account details, are often sent through text message. This makes them desirable targets for cybercriminals looking to steal sensitive data. Emails that seem like they come from a legitimate business or organisation are a common tactic employed by fraudsters to trick their victims into giving over personal information. Phishing refers to the practise of sending an email that seems to come from a reputable company but is really trying to steal personal information. Results from experiments are given that set the stage for the classification challenge and investigate the use of machine learning techniques to identify fraudulent emails.

Keywords: Spam Detection, Phisihing, Email phishing,Machine Learning.

1. INTRODUCTION

Phishing is widely recognized as a complex and quickly expanding danger in today's environment. Social engineering and technological tactics are used in this illegal conduct to get sensitive information, such as user names and passwords, from unsuspecting customers (Manning & Aron 2015)[1]. Hence, according to Lungu and Tabusca, the present economic crisis is a result of the growing number of cyber attacks and data breaches that target internet users (Lungu &Tabusca, 2010)[2]. Malware, phishing emails, and malicious websites are all forms of phishing that have been identified based on their preferred method of spread (Jain &Richariya 2011)[3]. Generally speaking, phishing emails are just another kind of spam. Emails are sent to users pretending to be from well-known organisations, such as banks, and requesting that they click on a link inside the email. If you click on the link, you'll be sent to a malicious website designed to steal personal information, such as login credentials or financial details (Al-Momani and Gupta 2013)[4]. The phishing attack lifecycle is shown in Fig 1. The first step is to send an email to the recipients in an effort to convince them to click on the link within. Similar to how a fisherman uses bait and line to reel in a catch, a phisher will blast out as many emails as possible in an effort to "catch" as many potential victims as they can by getting them to click on a malicious link (Al-Momani and Gupta 2013)[5].

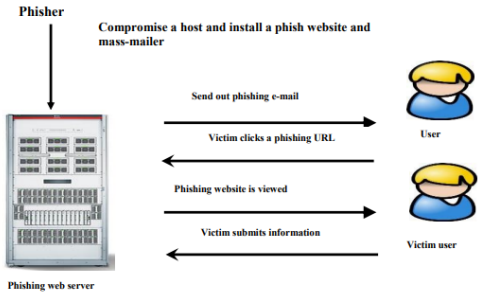

Fig1: Phishing Attack Life Cycle

Both the misleading phishing approach and malware-based phishing are used by phishers to accomplish their aims (Fig 2). The first method employs social-engineering schemes by sending emails with deceptive links; these emails can easily be mistaken for coming from a legitimate business or bank account, and they then lead the recipient to a fake website that asks for sensitive information such as login credentials, financial information, and personal details.

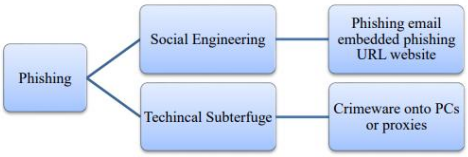

Fig2: Types of Phishing E-mails

Malware-based phishing, on the other hand, does not directly ask for details; instead, it relies on dangerous programmes or malware and technical approaches if users click on the embedded link, or looks for security weaknesses in the receivers' equipment, to get access to their online accounts. Sometimes the phisher would attempt to divert the victim to a legitimate but remotely monitored website. Al-Momani, A. (2013)[5].

In 2012, researchers released a paper online that predicted $1.5 billion in losses were caused by phishing scams. To mitigate the growing harm and danger posed by phishing emails, it is essential that more effective detection methods be developed. According to research (Akinyelu, 2014),

1.1 PROBLEM STATEMENT

Phishing is an identity theft method wherein the victim is tricked into giving over sensitive information through email that seems to have originated from a reputable company. Emails that seem to have originated from a legitimate source are often used in this way to steal sensitive data.

For financial information and personal identification, phishing emails are considered the fastest growing internet crime strategy. The safety of both the individual and their organisation is compromised when people fall for phishing scams and provide the desired personal or financial information through email, websites, or pop-up windows.

Microsoft's Consumer Safety Index found that phishing emails cost the global economy $5 billion each year. Repairing the damage they've caused, though, will set you back over $6 billion (MCSI reveals the impact of poor online safety behaviours in Singapore, 2014). No one set of characteristics has been shown to be the most useful in identifying phishing emails, despite the fact that this issue has been studied at length. The same nondeterministic setting is also used by the underlying classification mechanism. Finally, there is always room for improvement in the accuracy of detecting techniques. Last but not least, the following concerns were addressed in this investigation:

- How to choose the most effective combination of features for phishing protection.
- Choose the most effective classification algorithm for phishing detection.
- How to improve the best picked features and classifiers.
- Questions such as, "How to combine various classification algorithms for phishing detection, and how to assess such integration," and "How to integrate multiple categorization.

3. PROPOSED SYSTEM

Research conducted during the literature review indicates that numerous studies and strategies have been developed to identify or classify phishing emails; however, many of these researchers define spam and ham email filtering, and only a small number have defined proper emails filtering that is phished; and many users rely on aid from blacklists, heuristics, and visual similarities. But, machine learning has yielded the most fruitful results. When trying to safeguard their inboxes, many people unintentionally create spam filters. An example of this is the bag-of-words method, which is used extensively in the classification of emails and extracts all terms as highest occurring words, then uses the

implication of these words to categorise the messages. This approach fails miserably at email filtering but excels at stopping spam. The fundamental objective of this research is to discover methods through which automated systems can more accurately determine whether or not an incoming email is a phishing attempt. Machine learning techniques like random forest, logistic regression, Naive bayes, and Support vector machines, both supervised and unsupervised, are extensively researched and used. Indexing, content filtering, topic filtering, and content-type filtering are the four primary layers of email filtering. Messages are sorted by their text and other features in the organisation plot. Preliminary work is done to increase the yield boundary assessment, then the same provisions are deleted by simply identifying ideal credits in the list of capabilities; this impacts the consistency of the activity, and ultimately the needed data is utilised by both methods.

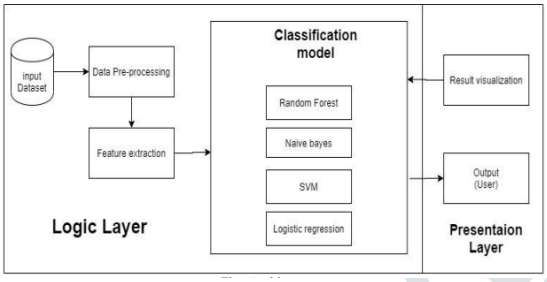

Fig3: System Architecture

4. RESULTS

Fig4: Entering the Spam message

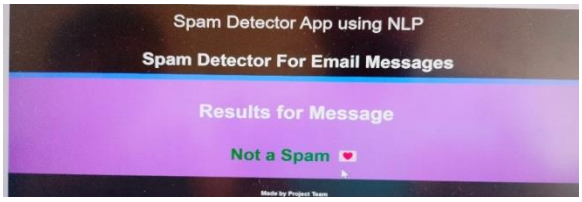

Fig5: Results for a Spam message

Fig6: Entering Spam Message

Fig7: Results for a Spam message

5. CONCLUSION & FUTURE SCOPE

Spam filtering in email is a significant challenge for both network security and machine learning. The Naive Bayes classifier plays a crucial part in this procedure for preventing spam in email. Naive Bayes's classifier's performance quality is also dependent on the dataset it was trained on. It is clear that datasets with fewer occurrences of emails and features may provide decent performance for Use of the naive Bayes classifier. The Naive Bayes classifier may also have the highest performance if the dataset was built from individual e-mail accounts, which gives largest proportion of spam message managed to block. The Naive Bayes classifier's success on the SPAM BASE dataset is clear.

REFERENCES

[1] Verizon, Data Breach Report 2016
[2] Andronicus A. Akinyelu and Aderemi O. Adewumi. Classification of Phishing Email using Random forest Machine Learning Technique 2014.
[3] Noor Ghazi M. Jameel, Loay E. George. Detection of Phishing Emails using Feed Forward Neural Network, International Journal of Computer Applications 2013.
[4] Ian Fette, Norman Sadeh, Anthony Tomasi, Learning to Detect Phishing

Emails, In Proceedings of the International World Wide Web Conference (WWW), 2006

[5] Gilchan Park, Julia M. Taylor, Using Syntactic Features for Phishing Detection 2015, https://arxiv.org/ftp/arxiv/papers/1506/1506.00037.pdf

[6] Gori Mohamed .J, M. Mohammed Mohideen, Mrs. Shahira Banu. Email Phishing - An open threat to everyone, International Journal of Scientific Research Publications, 2014

[7] C. EmilinShyni, S. Sarju, S. Swaminathan A MultiClassifier Based Prediction Model for Phishing Emails Detection Using Topic Modelling, Named Entity Recognition and Image Processing, SciRes 2016

[8] Noor Ghazi M. Jamee ,Loay E. George (2014), "Detection Phishing Emails Using Features Decisive Values",257-259

[9] Rakesh M. Verma and Nirmala Rai. Phish-IDetector: Message-Id Based Automatic Phishing Detection, International Joint Conference on e-Business and Telecommunications 2015 .

[10] Basnet R., Mukkamala S., Sung A.H. (2008) Detection of Phishing Attacks: A Machine Learning Approach. In: Prasad B. (eds) Soft Computing Applications in Industry. Studies in Fuzziness and Soft Computing, vol 226. Springer, Berlin, Heidelberg

[11] Adwan Yasin and Adbelmunem, An intelligent classification model for phishing email detection , International Journal of Network Security & Its Applications (IJNSA) Vol.8, No.4, July 2016

[12] D. J. Hand, HeikkiMannila, Padhraic Smyth. Principles of Data Mining

[13] Ron Kohavi, A study of cross validation and bootstrap for accuracy estimation and model selection, International Joint Conference on Artificial Intelligence, 1995.