ISSN 2348 – 7968

Spam Assassin E-mail Corpus Using Spectral Clustering and Random Forests

Ashutosh Thakur[1], Sangram Sathe[2], Gajanan Mahajan[3], Mangesh Pandit[4], Prof. Asmita Patil[5]

Department Of Computer Engineering,
DYPCOE, Akurdi, SPPU, Pune-411044

Abstract— **Phishing is an activity of identity theft that occurs when a malicious web site looks like a legitimate one in order to acquire sensitive information such as passwords, account details, or credit card numbers. Though there are several antiphishing software and techniques for detecting fake or phishing attempts in emails and detecting phishing information on websites, phishers come with advance techniques to avoid the available software and techniques. Phishing is a deception technique that utilizes a combination of social engineering and technology to gather sensitive and confidential information by performing as a trustworthy person or business in an electronic communication. Phishing makes use of spoofed emails that are made to look authentic and purported to be coming from legitimate sources like financial institutions, ecommerce sites etc., to lure users to visit fraudulent websites through links provided in the phishing email. The fraudulent websites are designed to mimic the look of a real company webpage. The phishing attacker's trick users.In order to handle phishing attack we use first step Spectral Clustering to analyze messages and second Random Forest based on regression.**

Keywords- Analysis, Latent Dirichlet Allocation, Link, Spear Phishing, Spectral Clustering,

I. INTRODUCTION

As the Internet started to gain popularity in the early's, it was quickly recognized as an excellent advertising tool. At practically no cost, a person can use the internet to send an email message to thousands of people. When this message contains an unsolicited Advertisement, it is commonly known as spam. Whether it is spam or not spam actually benefits the advertiser which is not really known, but to email users that receive over one hundred unwanted emails a day, the problem is serious.

Spam officially called unsolicited bulk email or unsolicited commercial email, which is rapidly becoming a major problem on the internet. It is an attempt to deliver a message, over the internet, to someone who would not otherwise to receive it. These junk emails may contain various types of message such as commercial advertising, quick rich scheme, pornography, doubtful product, illegal service or viruses. It is a direct email message, which targets individual email addresses by stealing the addresses from internet mailing list or searching the web addresses. Email spamming problem will become worse if the spam recipients reply to the spam sender, whereby caused their address available to be attacked by other spammers. As a result, the recipients will receive many spam messages in their mailbox. Spam is a big problem because of the large amount of shared resources it consumes. Spam increases the load on the servers and the bandwidth of the ISPs and the added cost to handle this load must be compensated by the customers. In addition, the time spent by people in reading and deleting the spam emails is wasted. Taking a look at the statistics 20% of the total emails are spam. These are truly large numbers.

II. RELATED WORK

Content Filtering approaches to phishing detection , which rely on text contents, where a filter either disallows delivery ofmessages recognized as phishing messages or delivers suspect messages to a special folder for careful review by the intended recipient provides a comparison of machine learning techniques for phishing detection using Term Frequency Inverse Document Frequency (TF-IDF) representation.[1] Training and test features were a function of the product of how often a term occurs within a document (TF) and the inverse of the proportion of documents containing the term (IDF).

This is known as a bag-of-words approach, because relationships between words are not directly considered in this representation. A total of 2,889 emails were used for evaluation, with 40.5% of the data set labeled as phishing messages. The phishing messages were composed of 1,171 out of 1,423 messages from the Phishing Corpus [9]. The nonphishing messages came from the researcherspersonal mail boxes. Comparing Neural Network, Logistic Regression, Bayesian Additive Regression Trees (BART), Classification And Regression Tree (CART), Random Forest, and Support Vector Machine (SVM) models, the results showed that the Random Forest outperformed the other methods when misclassification errors had equal cost.

III. PROPOSED SYSTEM

An attacker can achieve this by sending an email to a user which contains a content such as link to malicious site.Phishing detection systems typically depends on content filtering techniques, such as Latent Dirichlet Allocation (LDA), to identify phishing messages. In the case of spear phishing, however, this may be ineffective because messages from a trusted source may contain little content. In order to handle such emerging spear phishing behaviour, we use as a first step of Spectral Clustering to analyze messages based on traffic behaviour. In particular, Spectral Clustering analyzes the links between URL substrings for web sites found in the message contents. Cluster membership is then used to construct a Random Forest classifier for phishing. Data from the Phishing Email Corpus and the Spam Assassin Email Corpus are used to evaluate this approach. Phishing is a form of identity theft.

A typical phishing attempt consists of a phisher sending an email to a user. The email appears to come from a legitimate service, such as a financial institution, a social networking website, or an electronic message service provider. The email contains a link to a website that mimics the web site of the legitimate service provider. In fact, the graphics and layout of the website may be identical to the website of the legitimate service provider. The only difference may be the use of an intermediate link for grabbing form related data, such as account login information. This information is then passed to the legitimate service provider's website making it difficult for the user to know their account information has just been compromised.

Mail is transferred from SMTP protocol through internet. Then the system check it's IP whether the mail is from proper IP address or not. Next the system checks its domain such as whether the mail came from proper domain or not. The next process is Integrity Check.

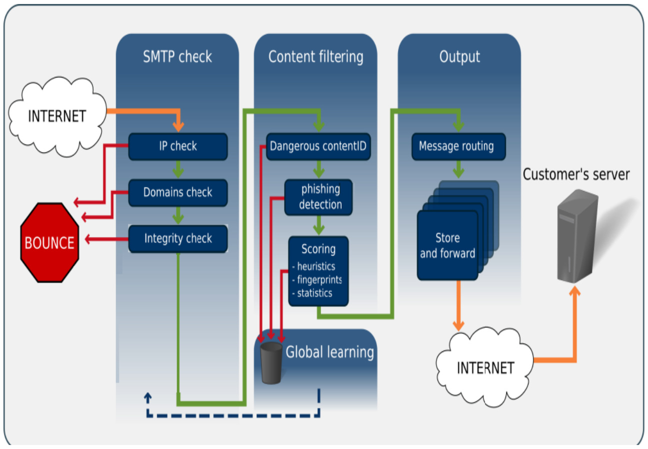

Fig1.1 System Architecture

In this process message body is checked by the system. Suppose if some malicious content is present in the mail then the system detects the mail as spam and according to scoring technique such as heuristics, fingerprints, and statistics. The system gives some degree to the mail if it is 0 then it is spam, if it is 1 then it is not spam and by message routing the message is sent to the customer's server.

Flowchart of the system is represented as follows

Module Description:

1) Data Pre-Processing:- Pre-processing is considered as an important step in text mining. There are three steps in preprocessing task for email classification, where we can use tokenization, stop word removal and stemming. First step is used as tokenization. In tokenizing process, all symbols (@, #, %,$), punctuations and numbers will be removed. The remaining strings will be split up into tokens. Second step is stopword removal. Many of the most frequently used words in English are useless in Information Retrieval (IR) and text mining. These words are called 'Stop words' .Stop-words, which are language-specific functional words, are frequent words that carry no information (i.e., pronouns, prepositions, conjunctions). In this step, the common words, which are the most frequent words that exist in a document like 'we','are', and 'is' and etc are removed. In English language, there are about 400-500 Stop words. Stop word list is based on word frequency.

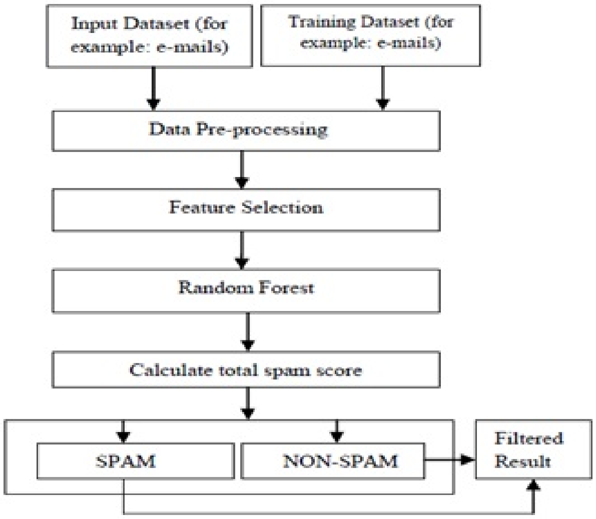
Input Dataset (for example: e-mails)
Training Dataset (for example: e-mails)
Data Pre-processing
Feature Selection
Random Forest
Calculate total spam score
SPAM
NON-SPAM
Filtered Result

Fig.1.2 Flowchart of System

This process will be identified by matching the words with the stop word lists and by comparing both of them. Removing these words will save spaces for storing document contents and reduce time taken during the searching process. Third step is stemming, 'Stemming' means finding the origin of the words and removing prefixes and postfixes. By using Stemming, forms of a word, like adjectives, nouns and, verbs, are converted to homological-like word. For instance, both ‘capturing’ and ‘captured’ are converted to a same word, ‘capture’.

2) Feature Selection: - Feature selection involves analyzing data (such as a bunch of average emails) and determines which features (words) will help the most in classification, which can then be used to train a classifier. One of Feature selection method is TF. Term frequency of each word in a document (TF) is a weight which depends on the distribution of each word in documents. It expresses the importance of the word in the document.

There are essentially 4 steps in **Spectral Clustering**:

1. Compute the Laplacian matrix L to represent the set of messages.

2. Derive the spectral decomposition of the Laplacian matrix: Eigen values (Λ) and Eigenvectors (Q).

3. Form the new spectral representation S from the Eigenvectors of Q.

4. Cluster the rows of the matrix S, normalized Laplacian matrix L is simply the affinity matrix A normalized by the degree matrix D:

$L=D^{-1/2} *A*D^{-1/2}$

3) Random Forests Algorithm: - The Random forest is a meta-learner which consists of many individual trees. Each tree votes on an overall classification for the given set of data and the random forest algorithm chooses the individual classification with the most votes. Each decision tree is built from a random subset of the training dataset, using what is called replacement, in performing this sampling. That is, some entities will be included more than once in the sample, and others won't appear at all. In building each decision tree, a model based on a different random subset of the training dataset and a random subset of the available variables is used to choose how best to partition the dataset at each node. Each decision tree is built to its maximum size, with no pruning performed. Together, the resulting decision tree models of the Random forest represent the final ensemble model where each decision tree votes for the result and the majority wins.

4) Calculate Total Spam Score: - Calculating the total spam score which is the minimum score required to mark a message as spam. If category is zero then Email is classified as Non-Spam & if category is one Email classified as Spam.

IV. UML MAPPING

It’s the graphical description of interaction between elements of system. In this use case diagram there are elements like user, system which are interacting with the system, Which consists of modules like sends mail, validate mail contents, store on Database, Malicious mail detection, domain analysis, responds to user request.

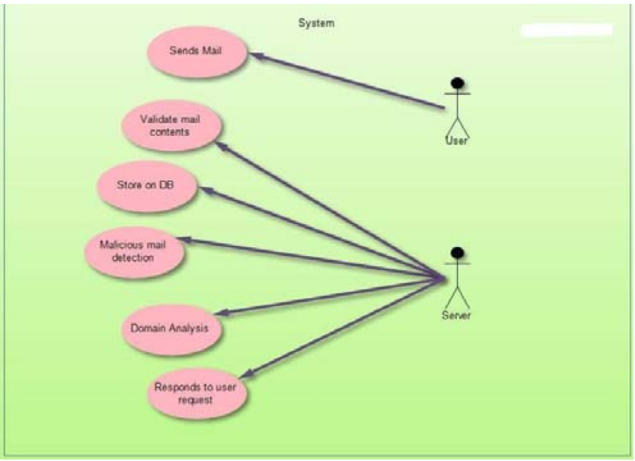

Fig 1.3.Use Case of System

V. RESULTS

Blacklisting and identify conventional blacklists constitute lists of IP addresses of likely spammers and are intended to help spam filters make better decisions about whether to block a piece of email based on the sender. Some Blacklists are policy- based (e.g., they list all IP addresses that belong to a certain class, such as dialup addresses). Other IP based blacklist are “reactive”: they attempt to keep track of whether an IP addresses is a spammer, bot, phisher. And keep this list up to date as hosts. These blacklists essentially maintain list of IP addresses and must be vigilantly maintained so as to not going out of date.

Content blocking like spam-tracker, to characterize and classify spam with analysis of network email contents. Clustering for spam classification previous studies have attempt to cluster spammer based on an email contents such as the URLS contained in the bodies of the emails. Focus on clustering spam senders to predict whether known spammer will send in the future. Cluster spam according to URLS to better understand the relationship between the sender spam messages that advertise phishing and scam sites and the web servers that host the scam themselves.

VI. CONCLUSION

This paper tells aboutAutomatic Email spam classification contains more challenges because of unstructured information, more number of features and large number of documents. As the usage increases all of these features may adversely affect performance in terms of quality and speed. The Random forest is a meta-learner which consists of many individual trees. Each tree votes on an overall classification for the given set of data and the Random Forest algorithm chooses the individual classification with the most votes. If identified category is 0 then E-mail is marked as Non-Spam E-mail otherwise if identified category is 1 then E-mail is marked as Spam E-mail.

When the system detects spam it will send notification on the mobile for verification of the mail. As this system is on receiver's side. In future it will be implemented on sender's side also.

REFERENCES

[1] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A Comparison of Machine Learning Techniques for Phishing Detection," Proceedings of the Anti-Phishing Working Group's 2nd Annual eCrime Researchers Summit, ACM, New York, 2007, pp. 60-69.

[2] D.M. Blei, A.Y. Ng, and M.I. Jordan, "*Latent DirichletAllocation",Journal of Machine Learning Research*, *vol. 3,* Jan 2003, pp. 993-1022.

[3] L. Breiman, "*Bagging Predictors*", Machine Learning, *vol. 24,* Aug 1996, pp. 123-140.

[4] L. Breiman, "*Random Forests*", Machine Learning, *vol. 45*, Oct 2001,pp. 5-32.

[5] L. Kaufman and P.J. Rousseeuw, "*Partitioning Around Medoids",*Finding Groups in Data: an Introduction to Cluster Analysis, Wiley,2005, pp. 68-125.

[6] SpamAssassin_EmailCorpus,http://spamassassin.apache.org/publiccorpus/, last accessed 2013-01-03.

[7] A.Y. Ng, M.I. Jordan, and Y. Weiss, "*On Spectral Clustering: Analysis and an Algorithm*", Advances in Neural Information Processing Systems, MIT Press, 2001, pp. 849-856.

[8] T. Fawcett, "*An Introduction to ROC Analysis*", Pattern RecognitionLetters, *vol. 27*, Jun 2006, pp. 861-874.