# MAXIMUM PHISH BAIT: TOWARDS FEATURE BASED DETECTION OF PHISING USING MAXIMUM ENTROPY CLASSIFICATION TECHNIQUE

**ASANI EMMANUEL OLUWATOBI**
Computer Science Programme
Landmark University
Omu-aran, Nigeria
asani.emmanuel@lmu.edu.ng
2348025717404

**ADEGUN ADEKANMI A.**
Computer Science Programme
Landmark University
Omu-aran, Nigeria
adegun.adekanmi@lmu.edu.ng

## ABSTRACT

Several antiphishing methods have been employed with the primary task of automatically apprehending and ruling out or preventing phishing e-mail from users' mail stream. Phishing attacks pose great threat to internet users and the extent can be enormous if unchecked. Two major category techniques that have been shown to be useful for classifying e-mail messages automatically include the rule based method which classifies email by using a set of heuristic rules and the statistical based approach which model e-mails statistically usually under a machine learning framework. The statistical based methods have been found in literature to outperform the rule based method.
This study proposes the use of the Maximum Entropy Model, a generative model and show how it can be used in anti-phishing tasks. The model based feature proposed by Bergholz et al (2008) will also be adopted. This has been found to outperform basic features proposed in previous studies. An experimental comparison of our approach with other generative and non-generative classifiers is also proposed. This approach is expected to perform comparably better than others method especially in the elimination of false positives.

**Keywords:** Antiphishing, Rule-based, Statistical-based, Machine learning, Maximum Entropy Model, generative classifiers, non-generative classifier

## 1.0 BACKGROUND TO THE STUDY

The internet with its ubiquity and simplicity continues to holds a lot of potentials for transactions between business people, researchers, co-workers etc. hence explaining its increasing proliferation as an integral part of our day to day life in this digital age. People depend on internet enabled computers, tablets and mobile devices to perform tasks on the go; and with its usefulness in research, communication, remote work etc., the internet has become an inalienable tool for faster and better performance among users. The web browsers through which internet access takes place have become one of the most popular applications among users. Unfortunately according to Hicks (2010), all of the increased benefits associated with computer technology came at the cost of personal security. Consequently, increased web based transactions has come with attendant risk to personal security. One of such risk to personal security is online identity theft through phishing attacks on internet users.

Phishing is a social engineering means of surreptitiously acquiring vital data from someone else using a clone of legitimate websites for criminal benefit. Guang (2013) describes phishing as a form of identity theft, in which criminals build replicas of target websites and lure unsuspecting victims to disclose sensitive information like passwords, etc. Phishers continue to device newer means such that even knowledgeable, security conscious persons can fall victims of phishing traps (Lucky, 2011). Phishes are done with the intent of acquiring personal data from users which can in turn be used illegally for financial transaction on behalf of the user. Attackers send email to unsuspecting victim with a link to a fake but identical version of an original website. Fig 1.0 is an example of what a typical phishing email looks like. Fig 2.0 also shows an example of a phishing website targeting paypal users. Haven taken the bait, the unsuspecting user is then required to fill in sensitive personal information that can be used by the attacker to perpetrate fraudulent activities (such information may include name, social security number, bank account number, or any other account number). Phishing is increasingly constituting threat to people's confidence in using the Web for online finance related transactions, information management and information exchange. It is important to implement an effective anti-phishing measure to protect users.

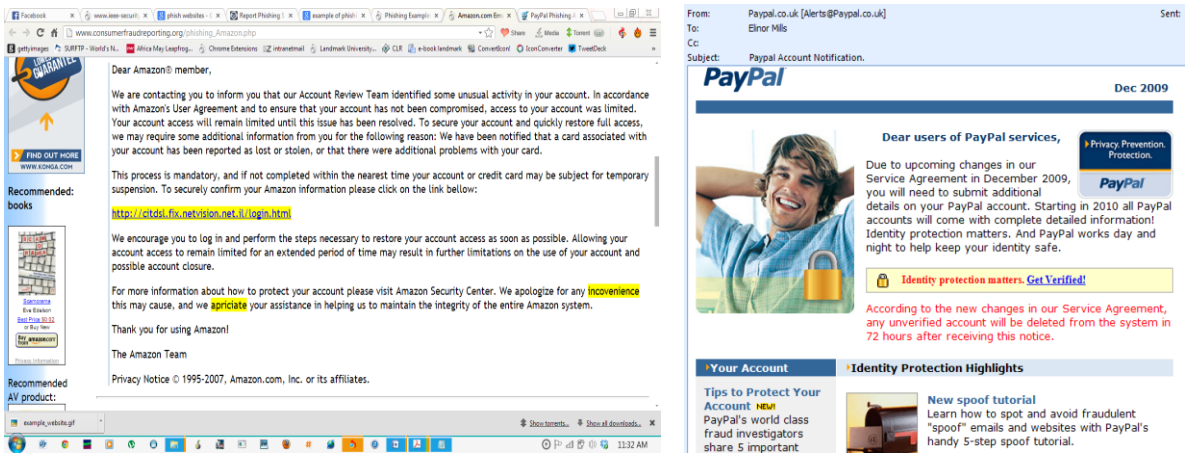


Fig1.0: Example of phishing email


Fig 2.0: Example of phishing website targeting paypal users

## 2.0 PROBLEM STATEMENT

Security problems occasioned by phishing attacks are increasing at an alarming rate thus posing serious risks to users. According to APWG (2011, 2013), more than 26,000 unique phishing attacks were reported in the first half of 2011 alone with the phishing attack up to a new high of 36,983 in March 2013.

In spite of lot of works that have been done on implementing better and efficient tools on phishing detection and prevention, it is still very hard to completely eradicate the problem and to estimate the number of users that are actually caught in bait of phishing as victim Guarav et al (2012). Literature suggests that phishers have been able to escape content filters by making the emails look legitimate Dhamija et al (2006). Most classifiers used in combating spam mails are able to classify text into one of several predefined categories. Bergholz et al. (2008) have used such classifiers as the Support Vector Machine (SVM classifier implemented in the libSVM-library) on email features generated by adaptively trained Dynamic Markov Chains and by novel latent Class-Topic Models. They have also run experiments using maximum entropy but didn't report the results in literature.

Maximum Entropy filtering has a well established history as an anti-spam weapon; Maximum Entropy was used effectively by Zhang and Yao (2003) for filtering junk mail with 99.83% precision rate. According to Zhang et al (2004), maximum entropy has been found to be non sensitive to feature selection strategy, easily scalable to very high feature dimension and good performances across different datasets. Therefore, though there are several solutions proposed and implemented for detection and prevention of phishing attacks, with the rate at which cyber scammers are able to come up with novel schemes constantly to continually circumvent phishing filters and with the unacceptable levels of false positives or missed detection

by available solutions, there is need to deploy the use of a classification model such as the Maximum Entropy which is non sensitive to feature selection strategy, easily scalable to very high feature dimension, performs excellently across different datasets and able to freely incorporate overlapping and interdependent features (Zhang and Yao, 2003).

## 3.0 RESEARCH OBJECTIVES

The main objective of this study is to develop a Maximum Entropy Model of classification to improve phish detection
The specific objectives are

- To identify feature related to phishing emails that will serve as input for classification
- Model Maximum Entropy classification scheme
- Train and validate the Maximum entropy model in real time environment
- To evaluate the performance of the Maximum Entropy model by calculating the, precision, recall, and error rate.
- Give a comparative study to demonstrate the merits and capabilities of the proposed approach.

## 4.0 RESEARCH QUESTIONS

1. What are the feature related to phishing emails that should be included to serve as input for the classifiers
2. What is the performance of the Maximum Entropy Model of classification on phishing email using precision, recall and error rate.
3. What is the performance of the Maximum Entropy Model compared to other classifiers.

## 5.0 LITERATURE REVIEW

Online transaction and correspondence via email has become an integral activity of this digital age. According to Ram and Andrew (2010), there are 3 major email categories: Ham, Spam and Phishing. Ham is solicited and legitimate email; although Spam is not necessarily illegitimate, it is unsolicited and therefore unwanted; Phishing, on the other hand is unsolicited, deceitful, and potentially harmful email.

Phishing is the general term used to refer to the creation and surreptitious use by cyber criminals of e-mails and websites designed to look like they are from legitimate, trusted and especially well known businesses, financial institutions or government agencies as a means of illegally gathering personal, financial and sensitive information that can be in turn used to perpetrate fraud. Phishes are primarily aimed at defrauding. According to Christine et al (2004) the term phishing was coined because fraudsters are fishing for personal information.

The term phishing is often used interchangeable with identity theft or identity fraud, it must however be noted that they are not exactly the same. Phishing is just one of a number of other methods employed by identity thieves to perpetrate identity fraud or identity theft.

Phishers continually evolve means of defrauding unsuspecting users of their personal information. Below are some tricks employed by phishers in phishing emails to unsuspecting users to appear as though they originate from legitimate sources as discussed by Christine et al (2004).

1. **Spoofing Reputable Companies:** In order to gain the trust of their email recipients phishers "spoof," or mimic, a reputable company. Financial companies such as Citibank, eBay, and PayPal are often spoofed apparently because of their vast customer base and reputation. Phishing emails are mass mailed without prior knowledge of recipients' profile (whether they are connected to the company or not) thus recipients that are not customers of the spoofed companies quickly realize that the email is fraudulent, or may believe that the email was sent in error to them and ignore the email. Phishers rely on the responses from those recipients who are customers of the spoofed company and who fall victim to the scam.
2. **Reply address differs from the claimed sender:** In some fraudulent emails, the email claims to be from a credible company, but is set to reply to a fraudulent reply address. The following are some examples from fraudulent emails:

   From: EarthLink Security Dept. From: Citibank
   Reply To: earthlink8770@1-base.com Reply. To: citibank3741@collegeclub.com

3. **Creating a Plausible Premise:** haven convinced the recipient of the authenticity of the email origin, phishers create premises strong enough to persuade the recipient to divulge personal information. Claims that the recipient's account information is outdated, or that their credit cards have expired, or that their accounts have been randomly selected for verification will normally produce the phishers' desired result them divulging personal information. Ironically, phishers often play on people's fear of fraud to defraud them: they create a scenario that convince the recipients that they must provide the requested information.
4. **Requires a Quick Response:** In the short time that fraudsters have to collect information before their sites are shut down, they must convince the recipients to respond quickly. The following are examples of urgent requests sent in fraudulent emails:

*"If you don't respond within 24h after receiving this Mail Information your account will be deactivated and removed from our server (your account suspension will be made due to several discrepancies in your registration information as explained in Section 9 of the eBay User Agreement.*
*"Please, give us the following information so that we could fully verify your identity. Otherwise your access to Earthlink services will be closed."*

5. **Security Promises:** Phishing emails also try to assure the recipient that the transaction is secure in hopes of gaining the recipient's trust. See fig 3.0

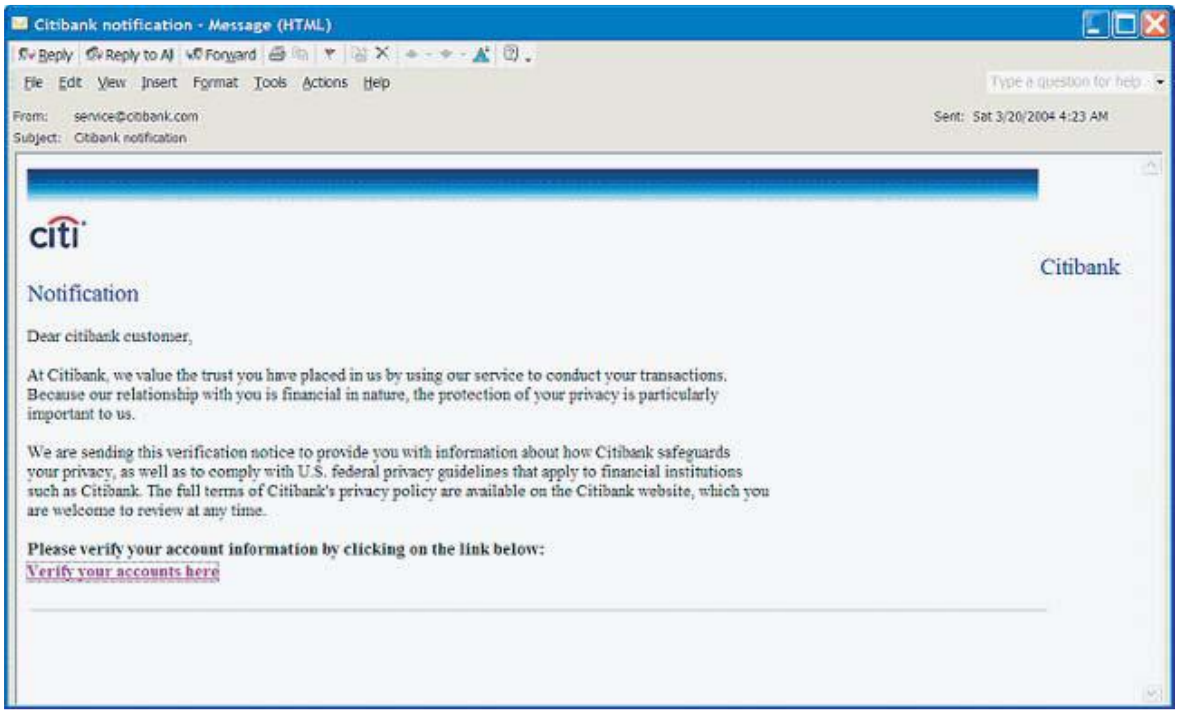

Fig 3.0: Security promises to phishing email recipient

6. **Collecting Information in the Email:** The earliest phishing emails used HTML forms within the email to gather information. This method of phishing is still used in some of today's scams. Once the information has been entered, the email must provide a method of sending the information to the fraudster. Generally the "Submit" button at the bottom of the form causes the information to be sent to the fraudster's specified location
7. **Links to Web Sites That Gather Information:** Now most phishing emails provide a link that takes the recipient to a Web site instead of using forms within the email. Some fraudsters register domain names that are very similar to those owned by a reputable company. For example, one fraudulent eBay email message used the following link:
*http://ebay-securitycheck.easy.dk3.com/Ebayupdatessl.html*
The real eBay site is located at www.eBay.com. The fraudster registered the domain name "http://ebay-securitycheck.easy.dk3.com" in the hopes of fooling the recipient into believing that this URL is owned by eBay. Other fraudsters try to conceal the true destination of the link by using HTML coding trick
8. Link Text in Email Differs From Link Destination In fraudulent email messages, the link text seen in the email is usually different than the actual link destination. In the following example, the email looks as though it is going to send the user to ***"http://account.earthlink.com"*** but instead sends the user to ***"http://www.memberupdating.com".***
"<a class="m1" target="_blank" title="Update"
href="http://www.memberupdating.com">http://account.earthlink.com</a>
9. **Using onMouseOver to Hide the Link:** Some fraudsters use the JavaScript event handler "onMouseOver" to show a false URL in the status bar of the user's email application. The following code was taken from the fraudulent PayPal email below.
***<A onmouseover="window.status='https://www.paypal.com/cgi-bin/webscr?cmd=_login-run';return true"onmouseout="window.status='https://www.paypal.com/cgi-bin/webscr?cmd=_login run'"href="http://leasurelandscapes.com/snow/webscr.dll">https://www.paypal.com/cgi-bin/webscr?cmd=_login-run</A>***
When the user puts the mouse over the link, the status bar shows: ***"https://www.paypal.com/cgi-bin/webscr?cmd=_login-run."***
However the link actually takes the user to ***"http://leasurelandscapes.com/snow/webscr.dll."***
10. **Using the IP Address:** Frequently, fraudsters attempt to conceal the destination Web site by obscuring the URL. One method of concealing the destination is to use the IP address of the Web site, rather than the hostname. An example of an IP address used in a fraudulent email message is "http://210.14.228.66/sr/." An IP address can be obscured further by expressing it in Dword, Octal, or Hexadecimal format.

11. **Using the @ Symbol to Confuse:** When the at symbol (@) is used in an "http://" or "https://" URL, all text before the @ symbol is ignored and the browser references only the information following the @ symbol. In other words, if the format <*userinfo*>@<*host*> is used, the browser is directed to the <*host*> site and the <*userinfo*> is ignored. This trick is used by scammers in hopes of fooling the person viewing the email code into thinking the link is going to the site listed before the @ symbol, while it actually links to the fraudulent site after the @ symbol. In the following link; ***http://cgi1.ebay.com.aw-cgiebayISAPI.dll%00@210.93.131.250/my/index.htm,*** the link may appear to be sending the user to eBay at ***"http://cgi1.ebay.com.aw-cgiebayISAPI.dll."*** However, the text before the @ symbol is ignored and the link sends the user to ***"210.93.131.250/my/index.htm,"*** which is the fraudulent Web site's IP address. To further conceal the URL, the @ symbol can be represented by its hexadecimal character code.
12. **Using Hexadecimal Character Codes:** Fraudsters can also hide URLs by using hexadecimal character codes to represent the numbers in the IP address. Each hexadecimal character code begins with "%."
13. **Switching Ports:** Web pages are accessed on servers through ports. A port can be specified by following the URL with a colon and the port number. If no port is specified, the browser uses port 80, the default port for Web pages. Scammers occasionally use other ports to hide their location

Several antiphishing measures have been deployed in the past as shown in literature to combat the phishing menace. Such measures include user trainings, blacklists and filters as well as the use of machine learning approaches via different classification methods.

Toolbars represent the first attempt to filter phishing attacks (Fette et al, 2007). Although Cranor et al (2007) reported most toolbars recorded 85% accuracy in identifying phishing websites, Wu et al (2006), after performing experiments on three security toolbars using 30 subjects to determine the effectiveness of these security toolbars reported that all three were ineffective in preventing phishing attacks owing chiefly to users' negligence of warning display by the toolbar and many companies' poorly designed websites.

Machine learning techniques using different classification algorithms have also been employed in literature to combat phishing. The classifiers classify email to either legitimate or phish by learning predefined features of the email. These features serve as inputs to statistical classification techniques, which are then trained to identify phishing emails.
Guarav et al (2012) have identified attribute or feature based antiphishing techniques as the best as it considers almost all area vulnerable to phishing.

Fette et al (2007), proposed machine learning approach with 10-fold cross validation training data using random forest as classifier and a total of thirteen features as input; they were able to classify emails achieving an overall accuracy of 99.5%, with a false positive rate fp of approximately 0.0013 and false negative rate fn on the dataset of approximately 0.035.

Ami and Sahani (2013) proposed the use of fuzzy classification method to tackle phish. They proposed four direct rule generation method based on fuzzy logic. Since it was a work in progress, they were unable to provide the input features to be used opining that fuzzy classification will provide good result due to its pedigree in the classification of spam mail detection.
Sonicwall (2008) reported the application of Bayesian classification to phishing email based on the premise of its well established history as an effective antispam weapon. Aside text analysis, Bayesian classification was able to filter using eight predefined indicators as input. An overall effectiveness rate of over 98% was reported with 0% false positive rate.

Zhang and Yao (2003) were able to filter junk mail using maximum entropy model. Two kind of feature (term and domain specific features) were used employing a 10-fold cross-validation approach in which the corpus is partitioned ten times into 90% training material, and 10% testing material in order to minimize random variation. The classification method recorded a precision rate of 99.83%. In addition, ME models was reported to achieve a higher recall, a lower error-rate and a better overall F1 when compared with the Naïve Bayes model with a 1.20% enhancement in recall, 35.71% reduction in error-rate and 0.69% enhancement in F1 measure, respectively. Zhang and Yao (2003) argued that it is the ME model's ability of freely incorporating overlapping and interdependent features that makes it superior to Naive Bayes.

Zhang et al (2004) identified Maximum entropy as a top performer being a very effective classifier for spam filtering. According to Zhang et al 2004 maximum entropy has been found to be non sensitive to feature selection strategy, easily scalable to very high feature dimension and good performances across different datasets.

## 6.0 JUSTIFICATION OF STUDY

Although there are several available solutions proposed and implemented in literature for improved detection and prevention of phishing attacks, particularly with the machine learning approach, the unacceptable levels of false positives or missed detection, makes the evolving of a new optimized approach becomes inevitable. Zhang et al (2004) described the Maximum Entropy Model of classification as a top performer because of its non sensitivity to feature selection strategy, scalability to very high feature dimension and good performances across different datasets.

This study is therefore important step forward in the combating of the phishing menace.

The proposed model, if implemented is expected to improve phishing detection with high precision rate and very minimal false positives.

## 7.0 METHODOLOGY

The proposed approach is a machine learning approach based on Maximum Entropy Model of classification. A model for machine learning is shown in fig 4.0 below:
Two classes are identified, namely the class of phishing emails, and the class of good ("ham") emails. Next, there is a notion of features. Features which are the properties of the emails to be classified being classified are then extracted. A learning algorithm in this case Maximum Entropy is then used to create a model, they accepts input as instance of the known features, and returns a class label as an output. The model trained by supplying a "training set" of data, comprising the feature assignments and class labels. A separate set of "test data" is then supplied to the model, and the predicted class of the data (phishing or ham) is compared to the actual class of the data to check for false positives and false negatives.

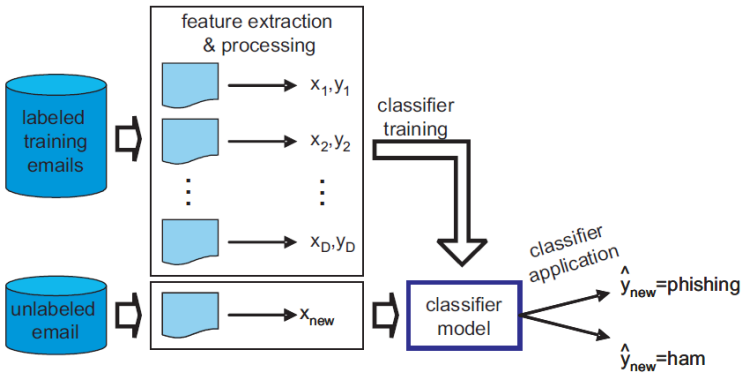


Fig 4.0: Machine Learning Approach (Bergholz et al, 2008)

The Maximim Entropy approach was chosen because of its well established history as an effective antispam weapon as reported by Zhang and Yao (2003) with 99.83% precision rate. Zhang et al (2004) also reported that maximum entropy is non sensitive to feature selection strategy, easily scalable to very high feature dimension and of good performances across different datasets.

**Maximum Entropy (ME) Model**
Maximum Entropy (ME) models have been successfully applied in classifying and filtering junk mail with state-of-the-art accuracies, relatively low error rate and high recall (Zhang and Yao, 2003).

The maximum entropy framework estimated probabilities based on the principle of making parsimoniously few assumptions such that, given a set of features, a set of functions $f_1 \ldots f_m$ (measuring the contribution of each feature to the model) and a set of constrains, we have to find the probability distribution that satisfies the constrains and minimizes the relative entropy (Divergence of Kullback-Leibler) $D(p||p_0)$, with respect to the distribution $p_0$. In general, a conditional Maximum Entropy model is an exponential (log-linear) with the form:

$$\boldsymbol{p(o|h) = \frac{1}{Z(h)} \prod_{j=1}^{k} a_j^{f_j(h,o)}}$$

o refers to the outcome
h refers to the history (or context)
Z(h) is the normalization function
*p(o|h)* denotes the probability of predicting an outcome *o* in the given context h with constraint or feature functions fj (o,h).

ME model represents evidence with binary functions known as contextual predicates in the form:

$$f_{cp,h'}(o,h) = \begin{cases} 1 & if\ o = o' and\ cp(h) = true \\ 0 & otherwise \end{cases}$$

cp is the contextual predicate which maps a pair of outcome o and context h to {true; false}

By pooling evidence into a bag of features, ME framework allows a virtually unrestricted ability to represent problem specific knowledge in the form of contextual predicates (Zhang and Yao, 2003).

**Building Corpora**
The labeled corpora provided by Spam Assasin will be used

**Feature Selection**
The proposed feature that serve as input for our classifier are adopted from Bergholz et al (2008) with a total of 27 basic features. Zhang et al (2004) noted that the Maximum Entropy Model benefits from large feature sets. The basic features are subdivided as follows:

- Structural Features (4) that is the body part structure of the email such as the ***total number of body parts, the number of discrete and composite body parts and the number of alternative body parts***
- Link Features (8): they include the ***total number of links, the number of internal and external links, the number of links with IP-numbers, the number of deceptive links (links where the URL visible to the user is different from the URL the link is pointing to), the number of links behind an image, the maximum number of dots in a link, and a Boolean indicating whether there is a link whose text contains one of the following words: click, here, login, update.***
- Element Features (4) this reflect the kinds of web technologies used in the email. Boolean of whether it is ***HTML, scripting, JavaScript, forms.***
- Spam Filter Features (2): ***the score and a Boolean of whether or not an email is considered as spam.***
- Word List Features (9): These are list of words hinting at the possibility of phishing. For each word in the list a Boolean feature of whether or not the word occurs in the email will be recorded. The list contains a total of nine word stems: account, update, confirm, verify, secur, notif, log, click, inconvenien.

**Training and Classification**
For the purpose of objectivity, 10 fold cross validation will be done. Nine of the ten parts will be chosen to serve as the training corpus, while the tenth part will be used for validation and performance measurements.

**REFERENCES**

[1] **Ami k. Trivedi and Sahani G.J. 2013.** Proposed Phishing mail detection using fuzzy classification methods The International Journal of Engineering And Science (IJES) Volume 2, Issue 1 pp 62-64
[2] **APWG. 2011.** Phishing Activities Trends Report, 1st Half 2011., (internet) http://www.antiphishing.org/reports/apwg_trends_report_h1_2011.pdf (accessed Aug 20, 2013).
[3] **APWG. 2013.** Phishing Activities Trends Report, 1st quartet / 2013. (internet) http://docs.apwg.org/reports/apwg_trends_report_q1_2013.pdf (accessed Aug 20 2013).
[4] **Bergholz A., Chang J.H., Paaß G., Reichartz F., and Strobel S 2008.** Improved phishing detection using model-based features. Proceedings of the Conference on Email and Anti-Spam (CEAS), 2008 (internet) http://www.ceas.cc/2008/papers/ceas2008-paper-44.pdf (accessed April 25, 2013).
[5] **Christine E. D., Jonathan J. O., and Eugene J. K., 2004.** Anatomy of a Phishing Email, Proceedings of Conference on Email and Anti-Spam (CEAS), 2004.
[6] **Cranor L, Egelman S, Hong J and Y. Zhang 2006.** Phinding phish: An evaluation of anti-phishing toolbars. Technical report, Carnegie Mellon University, Nov. 2006
[7] **Dhamija R, Tygar J. D. and Hearst M. 2006.** Why phishing works. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems,2006. Pages 581–590.
[8] **Fette I, Norman S, Anthony T 2007.** Learning to detect phishing emails. Proceedings of the 16th International World Wide Web Conference, ACM Press, Banff, Alberta, Canada, pp: 649-656. DOI: 10.1145/1242572.1242660.
[9] **Guang X. 2013.** Toward a Phish Free World: A Feature-type-aware Cascaded Learning Framework for Phish Detection (Doctoral Desertation) Language Technologies Institute School of Computer Science Carnegie Mellon University, Pittsburgh (internet) http://www.lti.cs.cmu.edu/research/thesis/2013/cmulti13001.pdf (accessed April 25, 2013).
[10] **Gaurav, Madhuresh Mishra, Anurag Jain, 2012.** Anti-Phishing Techniques: A Review. International Journal of Engineering Research and Applications. Vol. 2, Issue 2, pp.350-355.
[11] **Hicks D. 2010.** "Phising and pharming: Helping consumers avoid Internet Fraud" Public and reserve affairs, federal reserve bank training, 2010.
[12] **Lucky R. 2011 "**Clickphobia" IEEE Spectrum Magazine, 2011. [Online]. Available: http://www.boblucky.com/reflect/jan11.html
[13] **Ram B. Basnet and Andrew H. Sung, 2010.** Classifying Phishing Emails Using Confidence-Weighted Linear Classifiers. In Proceeding of International Conference on Information Security and Artificial Intelligence (ISAI 2010).
[14] **Sonicwall 2008.** Bayesian Spam Classification Applied to Phishing E-Mail. [online] Available: http://www.sonicwall.com/downloads/WP-ENG-025_Phishing-Bayesian-Classification.pdf (accessed April 25, 2013)
[15] **Wu M, Miller R. C. and Gar_nkel S. L. 2006.** Do security toolbars actually prevent phishing attacks? In Proceedings of the SIGCHI conference on Human Factors in computing systems, 2006.
[16] **Zhang Le and Yao Tian-shun, 2003.** Filtering junk mail with a maximum entropy model. In Proceeding of 20th International Conference on Computer Processing of Oriental Languages (ICCPOL, 2003), pp 446-453.
[17] **Zhang Le, Jingbo Zhu, Tians-hun Yao 2004.** An Evaluation of Statistical Spam Filtering Techniques. ACM Transactions on Asian Language Information Processing, Vol. 3, No. 4, pp 243-269.