Email Scam Detection with the Aid of NLP and Machine Learning

Ripsa P Khadir
College of Engineering, Cherthala
Managed by IHRD
Cherthala, India
ripsapk@gmail.com

Sony P
College of Engineering, Cherthala
Managed by IHRD
Cherthala, India
spsony@gmail.com

Abstract*—Email is a widespread used valuable communication medium among internet users in the world. Email scams are spreading like a contagious disease nowadays. Email scam is an unsought email that demands the prospect of a deal or something for nothing. Most of the scam emails incorporates exciting offers for the users and some invite victims to a website. Many greedy individuals have lost all their life savings because of these kinds of scams. Email scam is a form of email fraud which results in economic losses and wastage of time of email users. Different types of scams have been reported since this time. Many approaches have been developed in order to counteract this problem. Most of the existing approaches are static and do not consider the misspelled and conjoined words. Machine learning approaches are the most promising approaches in this field. We propose a model for email scam detection which is based on the combination of linguistic techniques along with the machine learning techniques. It helps in summarizing the email content. Misspelled words are also taken into account in the proposed approach. Content analysis is incorporated in order to identify the type of scam and it is based on the meaning interpretation of the email content. Word sense disambiguation is applied to overcome the lexical ambiguity. A trained classifier is used to classify the emails. SVM is used in order to accomplish this machine learning process. It is followed by the email summarization and content analysis phase accompanied with the word sense disambiguation. The proposed methodology yielded 94% accuracy for the machine learning phase. NLP phase yielded more accuracy in classifying phishing types of scams and yielded the least for nigerian type of scams.

Keywords*—Content analysis, Machine learning, Phishing, Word sense disambiguation

I. INTRODUCTION

The Internet is a global network of computers, much of which is unprotected against various electronic dangers. Email is a widespread critical communication medium nowadays among internet users. From the time an email composed to the time it is read and even after reaching safely at the user end, the email travels through this unprotected Internet, exposed to various malicious attacks. Along with the widespread usage of email, the Email-born attacks are also increasing day by day leading to economic losses. Scam is the type of fraud which is intended to deceive victims for personal gain or to damage through email. Both scam and spam has negative aspect. A scam is fraudulent business scheme which is often illegal. Spam is termed as junk or unwanted email and it is really annoying to legitimate users. Phishers who orchestrate scams use spam to accomplish their goals that is economic gain or satisfaction of damaging others. Phishing is a general term associated with emails and it is a major problem for internet users. Phishing which is also termed as brand spoofing is a kind of semantic attack in which victims are sent emails that deceive them into providing sensitive and personal information such as account numbers, passwords or other personal to phisher or attacker. Phishing emails are associated with a number of characteristics and some of them are jotted below:

- The body of the mail could make the user respond too quickly.
- It could upset or excite, for example, “your bank card is going to expire”, “winning a million dollar lottery” etc.
- Phishing email contents are normally not personalized, for example, “Dear customer”, “Dear user”, “Dear winner” etc.
- It will ask you to make a phone call or to update, validate, verify account information. This type of phishing emails are usually related to online payment services and financial institutions.
- The message or website of the hyperlink embedded within the email may include official-looking logos and other identifying information taken directly from legitimate websites which will make the user believe that it is legitimate.
- Phishing commonly targets Government, financial institutions (online banking websites) and online payment services like PayPal.

Phishing is actually a criminal process which steals and mistreats the customer’s personal information and financial account credentials. It can be termed as email fraud. It takes of different forms like spoofing, bogus offers, requests for

help, romance scam, Nigerian scam (419 scam) and lottery scam. The top five current email scams are those of bank, Nigerian scam, phishing email scam, virus emails and lottery email scam.

Stephen Hinde, the IS audit editor has reported that the Nigerian scams, famously known by the name 419 scam has led to an economic loss of around $5 billion in the year 2002 [1]. It has been reported that the Nigerian scam, also termed as advance free fraud scam leads to loss of hundreds of millions of dollars yearly [2]. Such 419 scams also include fake bank websites which will make the victims believe that it is legitimate. Many approaches like Ad-hoc information warfare have been adopted to overcome these kinds of scams. Such approaches hijack these fake websites as well as scammer's email accounts. Scam normally operates when a target receives an unsolicited email concerning a remunerative 'business proposal'.

Anti-Phishing Work Group (APWG) use to release reports on phishing attacks. According to the phishing trends report released by APWG for Q2, 2013, in addition to phishing websites, it studied on unique phishing emails, also termed as email campaigns that are sent out to multiple users and thereby directing the users to specific phishing websites [4]. According to the APWG phishing trends report Q2, 2013, phishing attack numbers declined notably from Q4 2012 to Q1 2013, with a 20 percent decrease between January and March 2013. They reported that February's 35,024 was the lowest number of attacks detected since the 36,733 seen in October 2011. They claim that decline in the virtual server phishing attacks was the main reason behind the decline in the phishing reports. APWG members also reported that sophisticated targeted content continues to make email a highly effective attack vector for phishing, malware, and spam and hence phishing will continue to be a major problem for the internet users ahead.

In order to overcome phishing, several approaches have been proposed in the last few years. They can be categorized into two major sections:

- Exhaustive anti-phishing efforts and
- Methodologies and algorithms to detect and filter phishing.

The former comprises the involvement of humans which includes user awareness, social and psychological studies and the education of users and providers. The latter present different approaches and algorithms used to protect the user against phishing.

This paper is organized as follows: Section 2 details the related work. Section 3 elaborates the system model used for scam email detection and filtering. Section 4 elaborates the experimental results and finally, Section 5 presents the conclusion and future work.

II. Related Work

Machine learning based approaches is the major category of phishing detection approaches. Different methods are developed in this area. These are termed as feature based approaches too. There exist a number of different structural features that allow for the detection of phishing emails. Based on that features, different approaches are designed for detection and filtering of phishing emails. They include URL-based, script-based, keyword-based, behavioral-based and content-based. The different features include URL features, script features, keywords used, text block features, image block features and style features. Machine learning based approaches can be further subdivided into different sub-sections. Many approaches has been developed for phishing detection by using bag of words model, based on different features, testing the performance of different classifiers, clustering methods, hybrid methods etc.

In the bag of words model based approach [11], the input email dataset is represented as an unordered collection of words. In this kind of method, grammar and word order are not considered. It is based on classifiers and the classifiers include SVM, k-Nearest neighbor, Naive Bayes, Adaboost etc. The major drawback is that this approach cannot deal with zero day phishing attacks. Many studies are conducted on comparing the performance of different classifiers. Abu Nimeh et al[12] has compared six classifiers namely Logistic Regression, CART, SVM, Neural Networks, BART and Random Forests and no standard classifier was found. Miyamoto et al.[13]also made comparative study of machine learning algorithms for phishing detection. Ram Basnet [14] conducted the same using 16 features, but it gained low accuracy. Ganster et al.[15] conducted binary and ternary classifications by establishing 15 new online and offline features. Comparisons were performed between the two types of classification in this methodology. But it took high cost because of online features. Isredza Rahmi A Hamid et al.[16] proposed a behavior-based approach. The features were extracted by observing sender behavior. Message-id field was also incorporated as a feature in this scheme. In this attempt, performance of different classifiers like Bayes Net, support vector machine (SVM), AdaBoost and Random Tree and it was found that Bayes Net and Random Tree achieved the highest accuracy.

An approach based on the relative probability of occurrence of the features was proposed by [17]. It used 18 features. The probability measure is calculated for both the phishing and ham training and test set of emails. In this methodology, at first the emails were pre-processed, then features were analyzed and thereafter subjected to phishing

detection using Feature Existence and Feature Decisive Value criteria (FEFDV).

A multi-stage methodology was proposed by [18]. With the aid of Conditional Random Field (CRF) and Latent Dirichlet Allocation (LDA), named entities and hidden topics were discovered in this approach. This approach employs natural language processing and machine learning. It is followed by the classification of each message as phishing or non-phishing using AdaBoost. From the emails classified as phishing emails, the impersonated entities are found using CRF. This approach gained 100% F-measure.

A multilayer approach proposed by [19] is based on three classifiers. The textual content of emails are subjected to the Bayesian classifier. Whereas the non-grammatical content of e-mails are subjected to a rule based classifier. The phishing emails usually contains fake websites. The third classifier is an emulator based classifier which uses the responses from the websites embedded in the emails. Multi-tier classifications [20] is a serial approach in which every tier uses a different machine learning algorithm. This approach suggested three types of arrangements for the classifiers c1(SVM),c2(Adaboost) and c3(Naive Bayes). It got about 97% accuracy for the arrangements c1-c2-c3 and c1-c3-c2. It got the least accuracy of about 93.33% in the arrangement c2-c3-c1. However, this technique suffers from lengthy time consumption and complexity of analysis since this technique requires many stages before arriving at the final decision.

PHONEY: was a novel approach proposed by [21]. This technique adopts an offense centric technique to detect phishing attacks by using fake responses which mimic the real users, basically, reversing the character of the victim and the enemy. This approach exists between a user's mail transfer agent (MTA) and mail user agent (MUA). The arriving emails are processed and embedded links are analyzed and then subjected to a content scanner which analyzes the corresponding webpage. The result returned by the content scanner is then compared against the hashDB entries. The comparison will tell whether the email is phishing or not.

PDENF (Phishing Dynamic Evolving Neural Fuzzy Network) is a novel approach proposed by [22] which adapts Evolving Connectionist System (ECOS) which is based on knowledge discovery. It is a supervised/unsupervised learning approach. It is designed to detect the unknown zero-day phishing email attacks. This approach has achieved high accuracy, true positive and true negative results. It used adaptive four algorithms from ECOS, namely ECM, ECMc, DENFIS and DyNFIS. But this approach needs continuous feeding.

III. SYSTEM MODEL

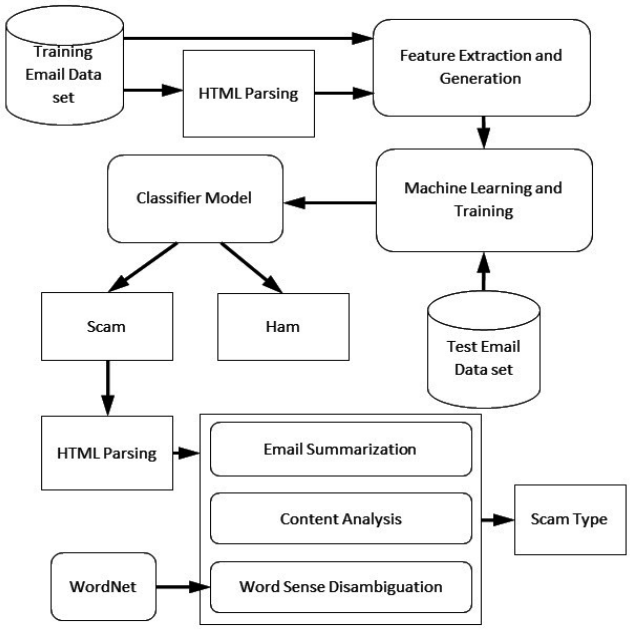

Figure 3.1. System Model

In the proposed system, as shown in Figure 3.1, the email dataset is first subjected to HTML parsing. Before that, the URL features are extracted from the emails. HTML parsing is the process of removing HTML tags from the email documents. This process is accomplished with the aid of regular expressions and NLTK toolkit. From the parsed emails, all other features are extracted.

In the pre-processing stage, the process of tokenization is done first. This results in separating the text into individual words. It is followed by the process of stop word removal, which is the process of removing the common words that are usually not useful for text classification. For example: to remove words such as "a", "the", "I", "he", "she", "is", "are", etc. Feature extraction is the next step in which different features are extracted and ranked. The best features are selected with the aid of Information Gain Ratio(IGR). Features like keywords and bigrams are extracted with the aid of document frequency measure. In case of unigram features, each word is used as a feature. In that case, the absence or presence of most frequently occurring unigrams is used as feature values.

While using, bigram features, absence or presence of frequently occurring two consecutive words are considered as a feature. URL features are extracted before the process of HTML parsing. Thus all the features extracted from pre-processing stage are passed to the feature generation phase. As a result, the feature vectors are generated. These

feature vectors are then used to train the classifier. Finally, the trained classifier is applied to the new email datasets in order to classify them into ham emails and scam emails. In this approach, SVM is adopted, for the purpose. It contributes a high level of accuracy. Machine learning approaches are like a deep ocean in the field of phishing detection. They are one of the most promising approaches.

As a result of the machine learning process, the emails are classified into scam and ham. The emails classified as scam are then subjected to the summarization phase after HTML parsing. This process will result in email summarization. The summaries are then further subjected to content analysis. In the proposed model, dictionary based approach is adopted for the accomplishment of content analysis. Several categories incorporating the most frequent keywords are used to build the dictionary. The categories correspond to the type of scam. Word sense disambiguation is incorporated along with this step in order to overcome lexical ambiguity. Lesk algorithm is used for attaining word sense disambiguation. In this phase, at first all sense definitions of the words to be disambiguated are retrieved from MRD (Machine Readable Dictionary). In the proposed system, the MRD used is WordNet. Determining the definition overlap for all possible sense combinations is the next step. As a final step, senses that lead to highest overlap is chosen. As a result, the type of scam is obtained.

A. Steps Involved

The different steps in implementing the proposed model are:

- Pre-processing: In this stage, the content is subjected to tokenization and stopword removal. Unwanted punctuation symbols are also removed in this stage.
- Feature selection and generation: The features are selected and ranked with the aid of information gain ratio and document frequency. The features vectors are generated in this stage by taking the results of previous stages into consideration.
- Training and learning: The features vectors are learned to train a model for classification.
- Classification: The test email data-set is classified using the trained model. SVM is used for this purpose in order to improve the accuracy.
- Content Extraction and parsing of emails: In this stage, the training email dataset is subjected to HTML parsing, resulting in the removal of tags and thereby extracting the content.
- Summarization: The process of summarization helps in obtaining the email summary.
- Content analysis: The process of Content analysis helps in determining the type of email scam. Dictionary based approach is used for this purpose.
- Word sense disambiguation: This stage helps to avoid the lexical ambiguity in the extracted content if there any.

B. Feature selection and ranking

The algorithms that are most used to find the most effective features and ranking them are information gain ratio (IGR) and document frequency[30]. IGR provides the information necessary to specify a feature value and it gives the highest weight to the most effective features and it can be explained by the following equations:

$$\text{gain\_ratio}(X,C)=\frac{\text{gain}(X,C)}{\text{split\_info}(C)} \quad (1)$$

Where, gain(X,C) represents the gain ratio of the feature X frequency in class C.

$$\text{split\_info}(C)=-\sum_i \frac{|C_i|}{|C|}\log\frac{|C_i|}{|C|} \quad (2)$$

Where, Ci and |Ci| denote the frequency of features X in class C, the i^{th} sub-class of C and the number of features in Ci, respectively. The highest value in an information gain will contribute a most useful feature. The ranking and IGR values of the features selected in the proposed model is given in the table below:

TABLE 3.1. RANK AND IGR VALUES OF FEATURES

Rank	Features	IGR
1	HTML format	0.58401
2	Presence of greetings	0.41725
3	IP based URL	0.31785
4	Age of domain name	0.31625
5	No: of domains	0.23585
6	No: of subdomains	0.22428
7	Presence of script	0.21105
8	Presence of misspelled words	0.14983
9	Presence of form tag	0.13329
10	No: of links	0.10658
11	No: of dots	0.08175
12	URL based image source	0.07158
13	Matching domains	0.02701

TABLE 3.2. KEYWORDS AND BIGRAMS

Keywords	lottery, winner, suspend, verify, account
Bigrams	verify online, update confirm, lucky winner, click here

Features like keywords and set of bigrams are selected with the aid of document frequency. It contributed the set of most important and frequent keywords and bigrams in the email corpus. Document frequency is given by the equation given below:

$$D_f = |\{ m_j | m_i \in M \text{ and } f_i \text{ occurs in } m_j\}| \quad (3)$$

Where M is the set of all training email dataset, f_i is a binary feature (such as "the word verify is present in the message"), and $\neg f_i$ is the negation of the feature f_i (such as "the word verify is NOT present in the message").

C. Features used

- HTML format: Hyperlinks are active and clickable only in html formatted emails. If an email is provided with only a link without formatting, the user is less likely to click on it. Thus, an HTML-formatted email is flagged and is used as a binary feature.
- Presence of greetings: Scam email contents are normally not personalized, for example, "Dear customer", "Dear user" etc. Presence or absence of greetings is considered as a feature. It is a binary feature.
- IP-based URL: A ham email will usually contain a domain name. Whereas a scam email may use ip addresses in place of domain name. Use of an IP address makes it difficult for users to know exactly where they are being directed to when they click the link. For example, http: //81.215.214.238/pp/. It is used as a binary feature.
- Age of domain name: In scam emails, the scammers may create new domains for their current purpose only. They are usually used for a limited time to avoid being caught. A WHOIS query can be used to achieve the age of a domain. It provides the creation date, expiration date, the name or person to which the domain is registered to etc. If the age of domain name is less than 30 days, it can be termed as a scam email. It is used as a binary feature.
- Number of domains: The total number of domains used in a URL is used as a feature. In scam emails, two or more domain names are used in an URL address to forward address from one domain to the other. For example, http://www.google. com/url?sa=t&ct=res&cd=3&url=http3A2Fwww.antiphishing.org2Fei=0qHRbWHK4z6oQLTmBMusg=uIZX3aJvESkMveh4uItI5DDUzM=sig2=AVrQFpFvi. In the given example, there are two domain names, google.com and antiphishing.org.
- Number of sub-domains: Most of the scam emails contain URLs with more than one subdomain. It is used as a feature. For example, https: //login.personal.wamu.com/verification.asp?d =1 has two sub domains, namely login and personal.
- Presence of Script: Scripts can be used to hide information and activate changes in the browser like on click events. Emails with keyword "Script" can be flagged as scam and it is a binary feature.
- Presence of misspelled words: Presence of misspelled words is considered as a feature. Whenever an email contains misspelled words, we flag it as a phishing email and use it as a binary feature.
- Presence of Form Tag: Presence of form tag is used as a binary feature. For example, consider <FORM action=http: //www.paypal−site.com/profile.php method= post>. The email might be addressed from www.paypal.com. But in the form tag, it is directed towards the profile.php. The victims may believe that to be from legitimate Paypal site.
- Number of Links: Most of the scam emails use links embedded within them. The number of links in email is used as a feature. A link in an email is one that makes use of the "href" attribute of the anchor tag.
- Number of dots: The scam emails will exploit the use of links for redirection and such links may use a number of dots in them. The number of dots in those links in email is used as a feature.
- URL based image source: If an email contains a single link without any images and formatting, the user will be less likely to click on it. Hence scam emails will be usually associated with URL based image source which looks similar to that of legitimate ones. Such images are usually linked from the legitimate company's web pages. The presence of such URL based image sources are used as a feature and it is a binary feature.
- Matching domains-From and Body: The domain name in "From" address and the domain name in the body is checked for matching. If the matching ratio is greater than 0.5, they cannot be scam. It is used as a binary feature.
- Keywords: Phishing emails contain number of frequently repeated keywords such as lottery,

winner, suspend, verify, account, etc. The absence or presence of words is used as the feature.

- Set of bigrams: Some handful set of bigrams like verify online, update confirm, lucky winner etc. are considered. If they are present in emails, it is used as a feature.

D. Data Used

To implement and test the proposed approach, two publicly available datasets i.e., the ham corpora from the SpamAssassin project as legitimate emails and the emails from PhishingCorpus as phishing emails (Phishing 2006, Spam 2006) are used. Additionally, lottery scams and Nigerian scams are added to the phishing mail corpus in order to make the scam corpus.

The total number of emails used in the approach is 6550, out of which 3533 are used as scam emails and 3017 as legitimate (ham) emails. The entire dataset is divided into two parts for testing and training purpose. A total of 3275 emails are considered as training samples and the remaining are considered for testing purpose. The different samples used are shown in table 3.3. Python is used to parse the scam and legitimate (ham) emails and extract the features mentioned above.

TABLE 3.3. DATA USED

Data	No:
Total samples	6550
Total scam emails	3533
Total legitimate emails	3017
Total training samples	3275
Total testing samples	3275

IV. EXPERIMENTAL RESULTS

A. ROC Curve

ROC is a graphical plot between the fraction of true positives (TP) and the fraction of false positives (FP). The point (0,1) corresponds to the perfect classifier. The perfect classifier classifies all positive cases and negative cases correctly. Thus the ROC curve is plotted by using the detection rates and false rates obtained from the phishing dataset [31].

False positive rate is the percentage of normal emails considered as scam emails which corresponds to the x-axis and true positive rate is the percentage of scam emails detected which corresponds to the y-axis in the ROC curve.

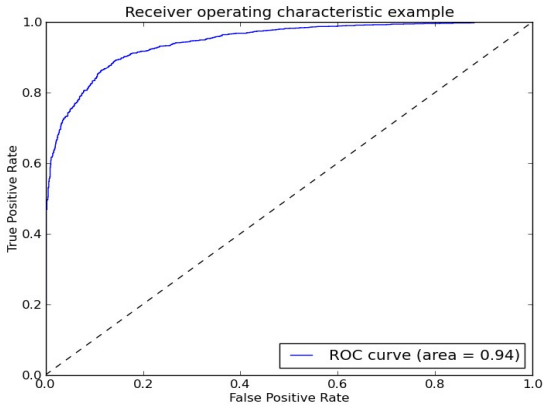

Figure 4.1. ROC Curve

The accuracy depends on how well the test classifies the emails correctly into the corresponding class. The area under the ROC curve (AUC) gives the accuracy of the classification test. An AUC of 1 represents a perfect test and an AUC of .5 represents a worthless test [31]. In the proposed experiment, an AUC of 0.9436 is got as shown in Figure. 4.1.

B. Confusion Matrix

Confusion matrix is a specific table layout or a matrix in the field of machine learning which is used for representing the performance of an algorithm, mostly supervised ones. It is termed as matching matrix in case of an unsupervised learning. The confusion matrix of the proposed experiment is shown in Figure. 4.2.

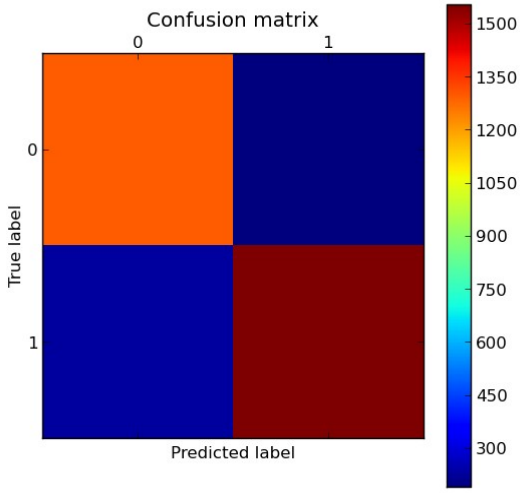

Figure 4.2. Confusion Matrix

The instances in a predicted class are represented by each column, while instances in an actual class are represented by each row of the confusion matrix [32].

C. *Types of Scam Classification*

A sample output as a result of applying NLP methodologies is shown in Figure 4.3.

```
SUMMARY:
If this is not completed by December 10, 2004 , we will be forced to suspend you
r account Indefinately, as it may have been used for fraudulent purposes.
d=1      Note: If you choose to ignore our request, you leave us no choice but t
o temporaly suspend your account.
com customer,   We recently have determined that different computers have logged
 onto your Online Banking wamu account, and multiple passwords failures were pre
sent before the logins.
[0, 0, 0, 39]
39
0
0
0
39
phish
Context: If this is not completed by December 10, 2004 , we will be forced to su
spend your account Indefinately, as it may have been used for fraudulent purpose
s
Sense: Synset('bank.n.07')
```

Figure 4.3: Sample Output

In the given sample output, the email considered is a phishing type and the content is related to suspending bank account of the customer. The summary is generated by the summarization process. The content of the email is analysed and categorised into “phish” category based on the most frequent tokens in the content. With the aid of Lesk algorithm, the sense of the summary and content is again tested inorder to improve accuracy.

As a result of the machine learning phase, 1654 emails are classified correctly as scam emails and 1297 emails are correctly classified as ham emails out of the total 3275 testing samples. In the NLP phase, the correctly classified scam emails are subjected to various NLP methodologies in the proposed system in order to classify them into different types of scams. The results are illustrated below:

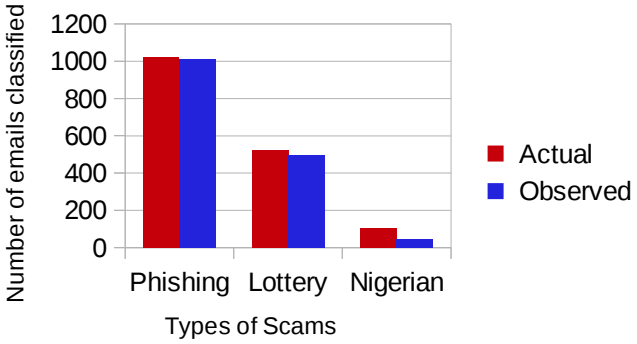

Figure 4.4: Types of scam

From the graph, it is clear that the most accuracy is found in classifying phishing type of scams and the least is that of the nigerian scams. It is mainly because of the changing and varied contents embedded in nigerian scams.

V. *Conclusion*

Scam emails are getting enormous day by day and they are annoying and dangerous to legitimate internet users. Economic losses due to phishing emails are rising in a non-stop manner. Different kinds of scam attacks are emerging per year. Scammers are developing new types of attacking methodologies overcoming the prevention and filtering mechanisms employed currently. The existing methodologies are static. It has the limitations in terms of accuracy and performance. A methodology incorporating linguistic features along with machine learning for phishing email detection and filtering is proposed. It can lead to even more promising results. Email summarization and content analysis methods are incorporated to identify the type of scam and word sense disambiguation method, in order to overcome the lexical ambiguity. SVM is used in this approach, which will lead to improved accuracy.

As a future work, a better feature selection on the same data set is planned by incorporating the set of features that produces the best accuracy. The scams based on images must be considered. The consideration of attachments on emails is also incorporated in the future plan.

References

[1] Stephen Hinde: ”Spam, scams, chains, hoaxes and other junk mail ”, Elsevier Science Ltd, 2002.

[2] Eve Edelson :”The 419 scam; information warfare on the spam front and a proposal for local filtering”, Elsevier Ltd, 2003.

[3] Anthony Elledge: ”Phishing: An Analysis of a Growing Problem”,SANS Institute InfoSec Reading Room, 2007.

[4] Anti-Phishing Work Group: Phishing Activity Trends Report 1st Quarter(2013), http://www.apwg.org, 22 November 2013. Natural Language Learning- Volume 7, Article No. 19, 2001.

[5] Blanzieri, EnricoBryl, Anton, ”A survey of learning-based techniques of email spam filtering”, Artificial Intelligence Review, Springer Netherlands, vol. 29,no.1, pp. 63-9, 2008.

[6] R. Dazeley, et al., ”Consensus Clustering and Supervised Classi cation for Profiling Phishing Emails in Internet Commerce Security”, Knowledge Management and Acquisition for Smart Systems and Services, Berlin Heidelberg, pp. 235-246, 2010.

[7] Yearwood, JMammadov, MBanerjee, A, ”Profiling Phishing Emails Based on Hyperlink Information,” 2010 International Conference on Advances in Social Networks Analysis and Mining, Odense, IEEE, Denmark 2010, pp. 120-127.

[8] Wilfried N. Gansterer David P et al., ”E-Mail Classification for Phishing Defense”, Springer-Verlag, presented at the Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, Toulouse, France,PP. 449-460, 2009.

[9] S. Abu-Nimeh, et al. : ”Distributed phishing detection by applying variable selection using Bayesian additive regression trees”, IEEE International Conference on Communications,vol.1, pp. 1-5, 2009.

[10] D. Miyamoto, et al. : ”An evaluation of machine learning-based methods for detection of phishing sites”, Advances in Neuro-Information Processing,vol.1, pp. 539-546, 2009.

[11] S. M. Ram Basnet, and Andrew H. Sung : ”Detection of Phishing Attacks: A Machine Learning Approach”, Studies in Fuzziness and Soft Computing, springer, vol. 226, pp. 373-383, 2008.

[12] W. N. Gansterer, et al. : ”E-Mail Classification for Phishing Defense”, Proc. 31th European Conference on IR Research on Advances in Information Retrieval, Springer Conf, Toulouse, France, pp.449-460, 2009.

[13] Isredza Rahmi A Hamid, Jemal Abawajy, Tai-hoon Kim : ”Using Feature Selection and Classification Scheme for Automating Phishing Email Detection”, Studies in Informatics and Control, Vol. 22, No. 1, March 2013.

[14] M Dolores del Castillo, Angel Iglesias, and J Ignacio Serrano, H Yin et al. : ”Detecting Phishing E-mails by Heterogeneous Classification”, IDEAL 2007, LNCS 4881, pp. 296305(2007).

[15] Rafiqul Islam , Jemal Abawajy : ”A multi-tier phishing detection and filtering approach”, Journal of Network and Computer Applications 36, pp. 32433, 2013.

[16] L. Ma, et al., ”Establishing phishing provenance using orthographic features,” IEEE Conf, 2009,pp. 1-10.

[17] M. Chandrasekaran, et al., ”Phishing email detection based on structural properties”, New York State Cyber Security Conference (NYS) , Albany, NY , 2006.

[18]Madhusudhanan Chandrasekaran, Ramkumar Chinchani, Shambhu Upadhyaya : ”PHONEY: Mimicking User Response to Detect Phishing Attacks”, Proceedings of the 2006 International Symposium on a World of Wireless, Mobile and Multimedia Networks IEEE, 2006.

[19] Ammar Almomani,Tat-Chee Wan, Ahmad Manasrah,Altyeb Altaher,Mahmoud Baklizi and Sureswaran Ramadass : ”An enhanced online phishing email detection framework based on Evolving connectionist system”, International Journal of Innovative Computing, Volume 9, Number 3, March 2013.

[20] Ammar Al-Momani,Wan T.C, K Al-Saedi, ”An Online Model on Evolving Phishing E-mail Detection and Classification Method,” Journal of applied science, vol. 11, pp. 3301-3307, 2011.

[21] Ammar Almomani, Tat-Chee Wan, Altyeb Altaher, ”Evolving Fuzzy Neural Network for Phishing Emails Detection,” Journal of Computer Science, vol.8,no.7, pp. 1099-1107, 2012.

[22] Machine Learning,http://en.wikipedia.org/wiki/Machinelearning, 4 November 2013.

[23] Content Analysis,http://en.wikipedia.org/wiki/Content analysis, 6 February 2014

[24]Summarization,http://people.dsv.su.se/hercules/textsammanfattningeng.html , 3 February 2014

[25] Natural language processing,http://en.wikipedia.org/wiki/Natural language processin 29 October 2013.

[26] Word sense disambiguation,http://en.wikipedia.org/wiki/Word sense disambiguation, 6 November 2013.

[27] NLTK,http://nltk.org/, 12 November 2013.

[28]Scikit-learn,http://scikit-learn.org/stable/supervised learning.html supervised learning, 11 January 2014.

[29] Python 2.7,www.python.org/, 13 November 2013.

[30] Tatsunori Mori, “Information gain Ratio as Term Weight”, http://aclweb.org/anthology/C02-1018, 22 january 2014.

[31]ROC,http://en.wikipedia.org/wiki/Receiveroperating characteristic, 19 March, 2014.

[32]Confusion Matrix, http://en.wikipedia.org/wiki/Confusion matrix, 26 March 2014.