International Journal of Recent Research and Review, Vol. XI, Issue 4, December 2018
ISSN 2277 – 8322

Detection of Phishing Websites Using Naïve Bayes Algorithms

Gyan Kamal[1], Monotosh Manna[2]
[1] *Scholar, Deptt. of Computer Science Engg., Yagyavalkya Inst. of Technology, Jaipur*
[2] *Asst. Professor, Deptt. of Computer Science Engg., Yagyavalkya Inst. of Technology, Jaipur*

***Abstract* - Phishing websites are form of mimicking the legitimate ones for the purpose of stealing user's confidential information such as usernames, passwords and credit card information. The prominence of phishing has risen over past years, with a number of unique attacks, reaching highest in year 2016. Attacks can be deployed with minimal cost and effort, enabling the attackers to launch large volumes of attacks in short spaces of time. This fast-paced nature of phishing makes automated detection processes critical for the safe-guarding of Internet users. Recently, the machine learning and data mining techniques have been a promising approach for detection of phishing websites by distinguishing between phishing and legitimate ones. The detection process in this approach is preceded by extracting various features from a website data set to train the classifier to correctly identify phishing sites. However, not all extracted features are effective in classification or even equivalent in their performance. In the present pursuit, we evaluate various machine learning algorithms with an optimization approach. Empirical results show that using the new proposed methodology an accuracy of 97.08% can be achieved by using Stacking, Bagging and Boosting along with Naïve Bayes, Decision Tree, and Random Forest Algorithms. This paper thoroughly investigates the use of machine learning for phishing detection, with features extracted from the URL only. It is one of the few techniques used to evaluate classification in a real-life scenario, using phishing and benign URL's retrieved from an environment where a large proportion of phishing attacks operate.**

Keywords - APWG, RMSE, RMSD, MAE, NAÏVE BAYES.

I. INTRODUCTION

Phishing can be better introduced as an attempt to crack the sensitive and confidential information such as usernames, passwords, and credit card details- indirectly money- often for malicious reasons, by disguising as a trust worthy entity in an electronic communication. Phishing is typically carried out by email spoofing or instant messaging and it often directs users to enter personal information at a fake website that looks and feel like a real most identical to the legitimate one. Communication supporting the social websites, auction sites, banks, online payment processor or IT administrators are often used to lure victims. Phishing emails may contain links to the websites that are infected with malware. One of the primary threats from phishing is identity theft. Consumers go to great lengths to protect their personal information, but a breach in security can expose a person to a multitude of threats, including credit card fraud, damaged credit, having an identity used for criminal activity, stolen bank information, unauthorized use of accounts (online and otherwise), or stolen money. It may also cause rare and intangible threats such as damage to credibility, loss of trust or embarrassment; having personal information stolen and can cost a great deal more than the lost cash. Authors in [1-4] have discussed various methods to detect phishing websites.

According to the Identity theft Resource Center, an average time spent repairing the damage caused by a stolen identity is approximately 600 hours and it can take years to completely recover [5-6]. Figure 1.1 shows the life cycle of phishing emails.

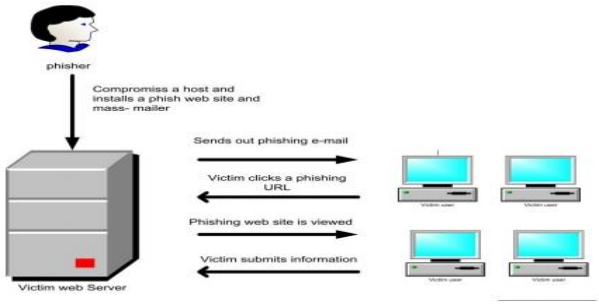


Fig. 1.1. Phishing E-mails based Life Cycle

Phishing is not just a "small-time" operation. Phishing is a business, and billions of dollars are being made by criminals while consumers and businesses are left to suffer the consequences.

Why does phishing activity increase?

Many reasons have contributed for the increase of Phishing activities. Points out that the necessity of technical resources to execute phishing attacks can be easily achieved through public and private sources. Equally, the automation of some Phishing technical resources have facilitated non-technical criminals to conduct phishing activities without any effort. One of the famous attack known as social engineering attack, increases rapidly because some Internet users are totally unaware of phishing and consequently cannot take any precaution when conducting online activities. Connected to systems such as bank, e-commerce systems, some Internet users lack knowledge concerning the policies of the system they are connected to, and ways for contacting system owners for issue related to privacy. This gives a door open to people conducting phishing to carry out their activities. Phishers are becoming more organized in their ways of thinking and operating. Their organization resulted to the creation of new ready-to-use phishing kit embedding items such as pre-generated HTML pages and emails for popular banks and e-commerce sites, scripts for processing user input, email and proxy server lists and even hosting services for phishing sites. With these kits creations, anyone connected to Internet can easily carry out phishing activities. According to APWG trends report of 2014, one additional reason of increasing of this activity is due to the cheapness and freeness of domain name registration. This is one major reason the number of people acquiring domain names for fake activities increases exponentially every day.

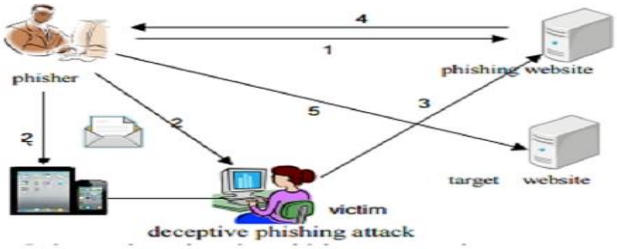


Fig. 1.2. Deceptive of Phishing Attacks Setup Stages

Deceptive attack steps:

1. The Phisher has to set a phishing website where all information entered to this site is posted to him. After user credentials are posted to the hacker, the hacker might effectively redirect the user to his/her original bank account while planning to connect to this account after the authorized user is disconnected.
2. After the phishing website is setup, the hacker has to broadcast phishing messages to potential victim's phones or PCs. This luring messages are meant to attract users to follow some links to bogus sites.
3. Non vigilant users follow links inserted in messages or e-mails they received, leading them to bogus site.
4. After the user connects and enters his Bank account number and other credentials, the phishing website will post them to the hacker.

II. METHODOLOGIES

We have used Naïve Bayes Algorithm for classification of phishing websites. For further optimization, we have used techniques like Bagging, Boosting and Stacking.

Naïve Bayes

Naïve Bayes is a simple technique for constructing classifiers models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. The

problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines. For some types of probability models, naïve Bayes classifiers can be trained very efficiently in a supervised learning setting. Web pages containing more external links than internal ones and password field input are classified as suspicious. Ram B Basnet *et al*. explained that a website content with more external links than internal links is an attempt to achieve some similarities and styles from external sources with the objective to steal user credential [7].

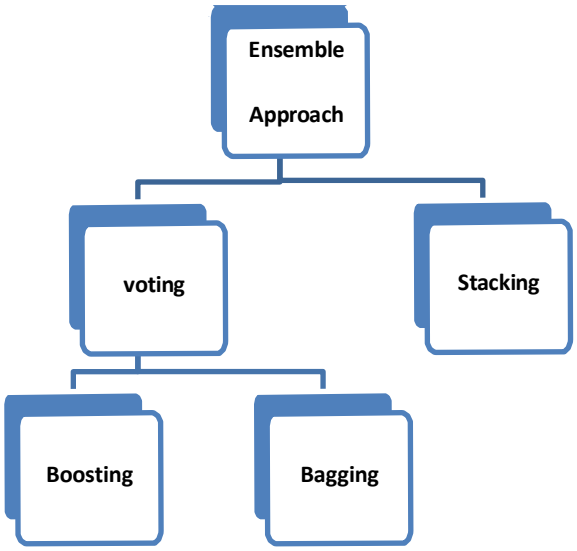


Fig. 2.1. Ensembling approach

Voting: In voting scheme, when classifiers are combined, the class assigned to a test instance will be the one suggested by most of the base level classifiers involved in the ensemble. Bagging and boosting are the variants of the voting schemes.

Bagging: Bagging is a voting scheme in which n models of same types are constructed. For an unknown instance, each model's predictions are recorded. That class is assigned which is having the maximum vote among the predictions from models.

Boosting: Boosting is very similar to bagging in which only the model construction phase differs. There will be n classifiers which themselves will have individual weights for their accuracies. Finally, that class is assigned which is having maximum weight. An example is Ada boost algorithm.

Stacking: In stacking, the predictions by each different model is given as input for a Meta level classifier whose output is the final class.

III. PERFORMANCE EVALUATION

Flowchart:

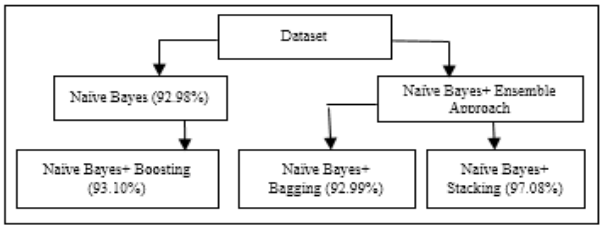


Fig. 3.1. Dataset of Naïve Bayes Evaluation

The legitimate e-mails are from both the 2002 and 2003 ham collections, easy and hard from [1].

Accuracy:

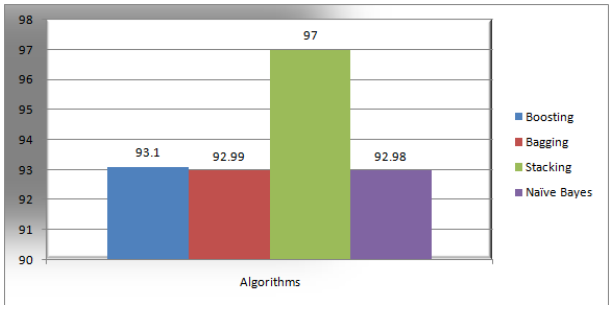
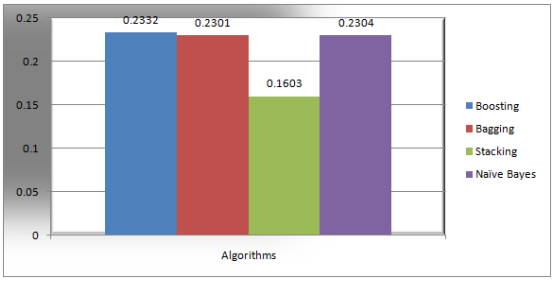


Fig. 3.2. Accuracy in Algorithm

Root Mean Square Error:

The root mean square deviation (RMSD) or root-mean-square error (RMSE) (or sometimes root-mean-square-error) is a frequently used measure of the differences

between values (sample or population values) predicted by a model or an estimator and the values observed.



Mean Absolute Error:

In statistics, mean absolute error (MAE) is a measure of difference between two continuous variables. Assume *X* and *Y* are variables of paired observations that express the same phenomenon. Examples of *Y* versus *X* include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement. Consider a scatter plot of *n* points, where point *i* has coordinates (x_i, y_i). Mean Absolute Error (MAE) is the average vertical distance between each point and the identity line. MAE is also the average horizontal distance between each point and the identity line.

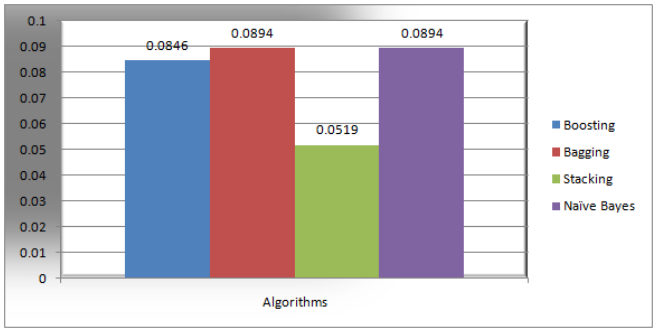


Fig.3.3. MAE X and Y Points

Related work:

Here, we have reviewed common intelligence phishing classification approaches with the help of Naïve Bayes in weka platform. In this we work, we have used Boosting, Bagging and stacking variants.

IV. PROPOSED SOLUTION

Filtering email content helps to identify Phishing scams spam and many other types of deceptive attacks. Researchers used collection of features extract from e-mails contents to detect scam mails. Some of these features are not efficient enough to accurately identify Phishing. Hence, we found our motivation from the used of ad e-mail by phishers as a means to achieve deceptive Phishing attacks. This idea has been taken into consideration in our work with a good set of words we use as feature in our set of features to efficiently detect and alert deceptive e-mails.

V. RESULTS

Naïve Bayes with Boosting

Correctly Classified Instances	1157	85.5137 %
Incorrectly Classified Instances	196	14.4863 %
Kappa statistic	0.7308	
Mean absolute error	0.1421	
Root mean squared error	0.2645	
Relative absolute error	37.9918 %	
Root relative squared error	61.1711 %	
Total Number of Instances	1353	

Fig.5.1. Output of Boosting

Naïve Bayes with Bagging

Correctly Classified Instances	1215	89.8004 %
Incorrectly Classified Instances	138	10.1996 %
Kappa statistic	0.8186	
Mean absolute error	0.1016	
Root mean squared error	0.2214	
Relative absolute error	27.1586 %	
Root relative squared error	51.19 %	
Total Number of Instances	1353	

Fig.5.2. Output of Bagging

Naïve Bayes with Stacking

Correctly Classified Instances	702	51.8847 %
Incorrectly Classified Instances	651	48.1153 %
Kappa statistic	0	
Mean absolute error	0.3741	
Root mean squared error	0.4324	
Relative absolute error	100 %	
Root relative squared error	100 %	
Total Number of Instances	1353	

Fig.5.3. Output in Stacking

VI. CONCLUSION

Phishing has caused many losses all over the world and continue to increase its number of victims tremendously. It appears in many forms or types with distinct modes of operation. The variety of phishing operational mode gives us a hint to pay more attention on some features that could help to efficiently detect phishing attacks. Therefore to address the problem of phishing through e-mail, we proposed a successful phishing detection framework that uses features that have prove to be good in the literature and yielded high accuracy using machine learning techniques. Hence, the detection accuracy at this level can greatly increase by populating the list of ad and porn words in the database. It is important to note that the alerting system only output Phishing mails, Ad email, and Phishing ad emails. From our experience, we noticed that by taking into consideration ad and porn e-mails, we have been able to alert most Phishing e-mails.

VII. REFERENCES

[1] WeiboChu, BinB.Zhu, FengXue, XiaohongGuan, and ZhongminCai, Protect sensitive sites from phishing attacks using features extractable from in accessible phishing urls ,Proceedings of IEEE International Conference on Communications, ICC 2013,Budapest, Hungary, June9-13,2013, pages1990–1994, 2013

[2] M.Dashand H. Liu. Feature selection for classification. Intelligent Data Analysis,1:131–156, 1997.

[3] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and IanH. Witten, The weka data mining software, An update SIGKDDExplor.Newsl.,11(1):10–18,November 2009.

[4] A.G.Janecek, W.N.Gansterer, M.A.Demel, and G.F.Ecker, On the Relationship Between Feature Selection and Classification Accuracy, JMLR: Workshop and Conference Proceedings 4, pages 90–105, 2008.

[5] Ian Fette, Norman Sadeh, and Anthony Tomasic, Learning to detect phishing emails, Proceedings of the16th International Conference on World Wide Web, WWW'07, pages649–656, 2007.

[6] Almomani, Ammar & Gupta, B B & Atawneh, Samer & Meulenberg, Andrew & Almomani, Eman. (2013). A Survey of Phishing Email Filtering Techniques. IEEE Communications Surveys &amp Tutorials. 15. 2070-2090. 10.1109/SURV.2013.030713.00020.

[7] Ram B. Basnet,et al., Feature selection for improved phishing detection,Proceedings of the 25thInternational Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems: Advanced Research in Applied Artificial Intelligence, IEA/AIE'12, pages252–261, 2012.