International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 9 • Issue 1 pp. 259-265 Sept 2017 - March 2018 www.csjournals.com

Phishing Email Filtering Techniques Using Machine Learning

Meenu, Sunila Godara
Guru Jambheshwar University of Science & Technology, Hisar, Haryana.

Abstract: Phishing technique is used to steal personal information for the purpose of theft and sending fake e-mail messages that come into view from lawful company. By doing this, person's confidential and private information can be stolen. In this paper we will study machine learning, filter selection methods and review the work done in phishing detection.

I. Introduction

Spam is a brand of canned cooked meat made by Hormel Foods Corporation. It was firstly commenced in 1937. It became more popular worldwide after its use in World War II. By 2003, Spam was widely used by six continents sold containing 41 countries and branded in over 100 countries (except in the Middle East and North Africa). In 2007, the seven billionth can of Spam was put on the market. Spam is electronic junk mail or unsolicited email. Conversely, if a longer known person finds your email address and sends a message to you, this is not treated as spam, although it is unwelcome. Real spam is normally email promotion for various products sent by a mailing list or newsgroup.

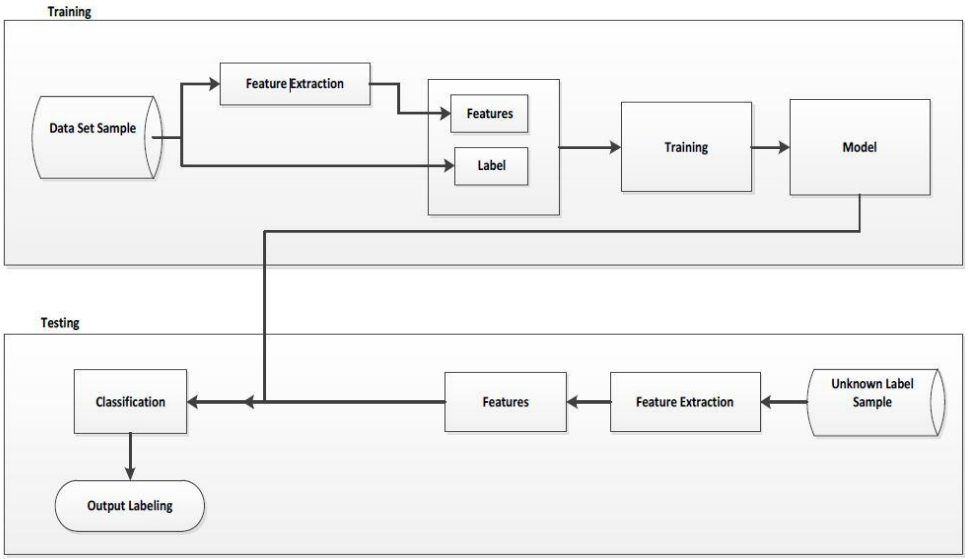

Figure (1) Automated Phishing Email Detection

The majority of the spam can be treated as unwanted e-mail but not all of the unwanted e-mails might not be spam. Wikipedia, the biggest encyclopedia on the Internet gives the following definitions: Spam: "E-mail spam (...) involves sending nearly identical messages to thousands (or millions) of recipients." Spamming: "Spamming is the abuse of any electronic communications medium to send unsolicited messages in bulk." Figure 1 shows the procedure of phishing which is started by sending emails to deal attacks of individual's[7][8].

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 9 • Issue 1 pp. 259-265 Sept 2017 - March 2018 www.csjournals.com

Machine Learning

Machine Learning has ability to be taught without explicitly programmed. Individual ability is incomplete and he/she is not able to find all the phishing. But the machine learning can make a machine intelligent and prevent from intrusion . Types of machine learning techniques are[1][2][3]:

1. Supervised learning
2. Unsupervised learning
3. Semi Supervised

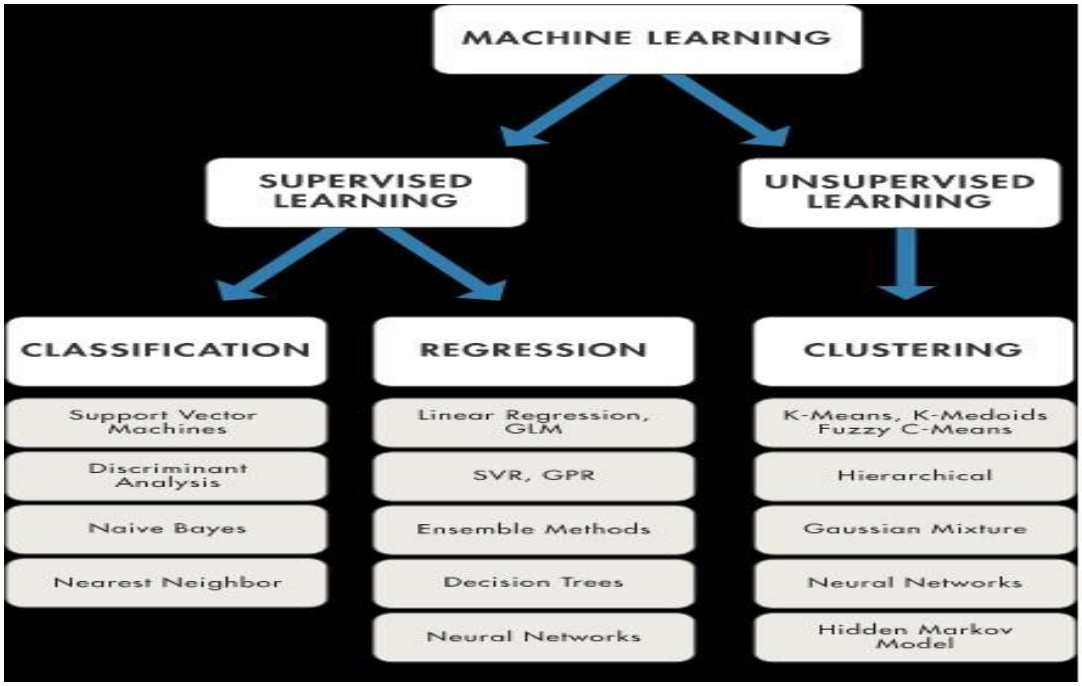

Figure (2)

At the moment machine learning is used in all segments of technology, that we don't still realize it while we are widely using it. The data is very massive and computation time is also increased, but due to Machine Learning people can process large data in minimum time with increased accuracy. It may be of following types:

1) Supervised learning: Overseen learning is a created and productive game plan in traditional topical gathering and has been grasped and investigated for evaluation revelation with alluring results . Critical directed course of action estimations are: Naïve Bayes, a generative classifier that assessments prior probabilities of P(X|Y) and P(Y) from the readiness data and produces the back probability of P(Y|X) in light of these previous probabilities; Support Vector Machine (SVM), a discriminative classifier that makes no previous suppositions reliant on the arrangement data and clearly measures P(Y|X) and the drowsy learning figuring K-Nearest Neighbors (KNN), which doesn't require prior advancement of a gathering model. In both topical and feeling gathering, Naïve Bayes and SVM are the most broadly perceived and ground-breaking coordinated learning computations. The best limitation related with controlled learning is that it is sensitive to the sum and nature of the arrangement data and may bomb when getting ready data are uneven or inadequate. Conclusion area at the sub-record level raises additional challenges for managed learning based approaches in light of the way that there is little information for the classifier[1][2][3].

2) Unproven learning: In content order, it is once in a while tough to make marked conceiving records, yet it is anything but challenging to crease the unlabeled archives. The hearsay learning techniques defeat these troubles. Customary point models, for example, LDA and PLSA are unsupervised techniques for removing inert themes in content archives. Subjects are including, and each component (or theme) is an

appropriation over (highlight) terms. The confinement of unsupervised methodologies is that they ordinarily need a huge volume of information to be prepared precisely. Completely unsupervised models frequently produce indiscernible themes on the grounds that the target elements of point models don't constantly correspond well with human decisions. In spite of this detriment, unsupervised adapting still offers us an approach to pick up learning about the information with no comment[1][2][3].

3) Semi-Supervised learning (SSL): SSL models drive from either directed or solo procedures. Curiously with coordinated acknowledging, which increases from checked data just, SSL gains from both named and unlabeled data.

The various feature selection methods which can be used are described as:

- Chi Square test

Chi Square Test is used in statistics to test the independence of two events. Given dataset about two events, we can get the observed count O and the expected count E. Chi Square Score tells how much the expected value E and observed value O derivate from each other. To estimate the χ^2 value of a spam table1 is used.

	Positive class	Negative class	Total
spam occurs	A	B	A+B = M
Ham occurs	C	D	C+D = N – M
Total	A+C = P	B+D = N – P	N

Table1. Confusion Matrix

Based on the null hypothesis that the two events that are independent, expected value E_A can be calculated using the following formula:

$$E_A = (A + C)\frac{A + B}{N}$$

- Pearson's Correlation: It is used to find linear dependence between two continuous variables X and Y. Its value varies from -1 to +1. Pearson's correlation can be calculated as:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

$$cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

The formula for Pearson's Correlation is given as:

$$\rho X,Y = \frac{E[(X-\mu_X)(Y-\mu_Y}{\mu_X \mu_Y}$$

- Kendall Correlation:

Kendall's rank connection is one of a few insights that measure the connection between rankings of various ordinal factors or distinctive rankings of a similar variable. At the end of the day, it gauges the similitude of orderings when positioned by the amounts. Both this coefficient and Spearman's connection coefficient are intended for use with non-parametric and non-regularly appropriated information.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 9 • Issue 1 pp. 259-265 Sept 2017 - March 2018 www.csjournals.com

- Fisher scores :

Fisher score is supervised feature selection method and selects each feature separately according to their fisher score, which provides a suboptimal subset of features, is given as:

Let, $p(i^{g,l})$ denotes the probability that $i^{g,l}$ is selected .then define $p(i^{g,l})$

$$p(i^{g,l})=\frac{f(i^{g,l})}{\sum_{l-1}^{N} f(i^{g,l})}$$

Big data and cloud computing are gaining significance for their working, machine learning is widely used as technology to analyze those big chunks of data, reduction of data scientists work by using an automated process .

II. Literature Review

This section reviewed various papers related to Phishing Discovery Using Machine Learning.

4. Toolan et.al proposed a new C5.0 algorithm to classify into Phishing and non-Phishing categories by taking 5 features and outperformed many existing methods. 8,000 emails were taken in which half were phishing and the other half were non-Phishing.
5. Abu-Nimeh et.al proposed a detection tool for mobile protection mobile against attacks. The client-server used Additive Regression Trees to improve their predictive accuracy and eliminated the overhead of variable selection.
6. Gansterer et.al proposed a filtering system which classified the mails into three classes such as: legitimate, spam, and phishing emails. This classification was performed using recently collected email's features. System achieved 97% accuracy among the three groups, by using two binary classifiers.

9 Dr. Ma et.al used an algorithm with a set of orthographic features to cluster phishing emails automatically and eliminating redundant features. This clustering and feature selection technique succeeded in providing highly efficient results. Ma applied the global k-mean model with a little modification and generated the values of the objective function over a range of tolerance values of selected features subsets. The objective function values assisted in recognizing the suitable clusters based on the distribution of these values.

10 Basnet et.al studies a detection approach that utilizes readily acquired features from the email's content without resorting to heuristic-based phishing features. This approach relied on Confidence-Weighted Linear Classifiers proposed by Basnet. images are generated by Phishers from the message's text that only graphical data passes the phishing filter.

11 Dr.Wu et.al focused on spoofing emails and Microsoft Outlook TM services by developing a sender authentication protocol (SAP). This authentication protocol verifies the authenticity of the sender by testing the claimed-sender1 with the archived emails. The enhanced OutlookTM has an add-in that tests feasibility while it remained the same user-friendly interface of the original version, and this the SAP add-in will be started automatically once the OutlookTM operates.

12 Khonji et.al (2011) used 47 features for the Email to classify the phishing emails in the study and they gave a brief description on each feature, the list covers all the structures of the Email.

13 Alguliev et.al developed new genetic algorithm for clustering spam messages and solving clustering problems. The proposed algorithm used the policy of maximizing the similarity between messages in clusters, and the objective function was defined by k-nearest neighbor algorithm. Performance of such algorithm was limited by the constant support of chromosomes. By which slow convergence is achieved. Further, penalty function was used speed up the convergence process and leaving infeasible chromosomes.

14 Alguliev et.al proposed new clustering method. Spam messages were collected, and Genetic algorithm with penalty function is used for solving clustering problem. In addition to, the classification of new spam messages coming to the bases of anti spam system. The proposed system is not only capable to detect purposeful information attacks but also to analyze origins of the spam messages from collection, it is possible to define and solve the organized social networks of spammers.

15 Al-Momani et.al proposed Phishing Evolving Neural Fuzzy Framework by using adaptive evolving fuzzy neural network . Root Mean Square Error (RMSE) and Non-Dimensional Error Index (NDEI) were used to measure performance.

16 Altaher et.al relied on Adoptive Evolving Fuzzy Neural Network (EFuNN) to create Phishing Evolving Neural Fuzzy Framework (PENFF) to detect of unknown "zero-day" phishing emails by handling all similar feature vectors to establish rules for prediction. Therefore, PENFF approach relies on the similarity of features included in the email's body and URL.

17 Zhang et.al used cross validation for phishing emails detection. Multilayer feed forward neural networks with dissimilar numbers of hidden units and activation functions was used and provided more accurate results. these results were obtained with few training samples.

18 Al Momani et.al(2013) proposed a new model called Phishing Dynamic Evolving Neural Fuzzy Framework (PDENF) with improved results in terms of recall, precision, F-measure and accuracy compared with other methods. The model has done prediction of emails in online mode .

19 Akinyelu et.al (2014)classified phishing emails using forest machine learning mechanism. 2000 phishing emails were used for testing and attained accuracy (99.7%) with low false negative (FN) and false positive (FP) rates.

20 Nizamani et.al (2014) applied several classification techniques such as: SVM, NB, J48 and CCM, by taking different features sets.

21 In 2015, Kathirvalava kumar et.al proposed a multilayer neural network for phishing detection using feed forward pruning algorithm that takes data as features from the email and applied a weight trimming strategy. With pruning approach minimizing the number of features are minimized so less computation time was required for classification of emails .

22 Roosevelt C. Mosley, et.al [23] examined the utilization of connection, bunching, and affiliation investigations to online networking. This is exhibited by dissecting protection Twitter posts. The consequences of these investigations help recognize watchwords and ideas in the online networking information, and can encourage the use of this data by safety net providers. As safety net providers investigate this data and apply the consequences of the examination in significant zones, they will probably proactively address potential market and client issues all the more successfully.

23 Huifeng Tang, et.al [24] talked about four issues, i.e., subjectivity order, word slant arrangement, archive supposition characterization dependent on AI procedures, and conclusion extraction issue. Despite the fact that we had the option to get genuinely great outcomes for the audit order task through the decision of proper highlights and measurements, yet we recognized various issues that make this issue troublesome.

24 Robert Malouf, et.al [25] suggested that relational association assessment is a critical instrument for performing ordinary language getting ready tasks with easygoing web compositions. A database of postings from a US political trade site was accumulated, close by self reported political heading data for the customers. A variety of presumption assessment, content gathering, and casual network examination systems were associated with the postings and evaluated against the customer's self descriptions.

25 Scott S. Piao, et.al [26] proposed a framework which depends on existing semantic lexical assets and NLP apparatuses, expecting to make a system of assessment extremity relations among archives and references. This is an electronic framework which enables clients to get to the references gathered from reports and recover those archives connected to every one of the references with various feeling extremity relations, to be specific endorsement, nonpartisan or objection relations. Different methodologies will be tried including recognizing semantic direction of emotional words with regards to references and AI utilizing physically clarified information

26 Matt Thomas, et, al [27] explored whether one can decide from the transcripts of U.S. Congressional floor discusses whether the discourses speak to help off or resistance to proposed enactment. To address this issue, we abuse the way that these talks happen as a major aspect of an exchange; this enables us to utilize wellsprings of data in regards to connections between talk portions, for example, regardless of whether a given articulation shows concurrence with the assessment communicated by another. We find that the consolidation of such data yields considerable upgrades over characterizing addresses in detachment.

27 Maite Taboada, et.al [28] given an exploration goal to extricate data on the notoriety of various writers, in view of works concerning the writers. The venture means to make a database of writings, and

computational instruments to concentrate content naturally. This paper depicts the underlying phases of an undertaking following the artistic notorieties of six creators somewhere in the range of 1900 and 1950, and the appropriateness of existing procedures for separating assessment from writings that examine and scrutinize these writers.

III. Conclusion

The Microsoft Consumer Safety Index survey concluded that the yearly worldwide collision of phishing email was US $5 billion. Supplementary, the cost of renovating their crash was US $6 billion .As enormous work is done in phishing email detection task, but there is no predefined set of features that can be used for phishing detection. Same nondeterministic set-up is applied by all classification algorithms. Finally, for enhancement in accuracy of the detection model best set of features, best classification algorithm and integration of multiple classification algorithms are highly required.

References

1. Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007, October) ieee. A comparison of machinelearning techniques for phishing detection. InProceedings of the anti-phishing working groups 2nd annual eCrime researchers summit (pp. 60-69). ACM IEEE.
2. Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2009, June). Distributed phishing detection by applying variable selection using Bayesian additive regression trees. In Communications, 2009. ICC'09. IEEE International Conference on (pp. 1-5). IEEE.
3. Adida, B., Chau, D., Hohenberger, S., & Rivest, R. L. (2006). Lightweight email signatures. In Security and Cryptography for Networks (pp. 288-302). Springer Berlin Heidelberg.Akinyelu, A. A., & Adewumi, A. O. (2014). Classification of phishing email using random forest machine learning technique. Journal of Applied Mathematics IEEE.
4. Alguliev, R. M., Aliguliyev, R. M., & Nazirova, S. A. (2011). Classification of textual e-mailspam using data mining techniques. Applied Computational Intelligence and Soft Computing,10 IEEE.
5. Al-Momani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Al-Momani, E. (2013). A survey of phishing email filtering techniques. Communications Surveys & Tutorials, IEEE, 15 (4), 2070-2090.
6. Al-Momani, A., Gupta, B. B., Wan, T. C., Altaher, A., & Manickam, S. (2013). Phishingdynamic evolving neural fuzzy framework for online detection zero-day phishing email. 56 IEEE.
7. Al-Momani, A., Wan, T. C., Altaher, A., Manasrah, A., Al-Momani, E., Anbar, M., Ramadass, S. (2012). Evolving fuzzy neural network for phishing emails detection, Journal of Computer Science, 8, 1099 IEEE.
8. Almomani, A., Wan, T. C., Manasrah, A., Altaher, A., Baklizi, M., & Ramadass, S. (2013). An enhanced online phishing e-mail detection framework based on evolving connectionist system. International Journal of Innovative Computing, Information and Control (IJICIC), 9(3), 169-175 IEEE.
9. Altaher, A , Al-Momani, A., Wan, T. C, Manasrah, A., Al -Momani, E., Anbar, M., ... & Ramadass, S. (2012). Evolving fuzzy neural network for phishing emails detection. Journal of Computer Science, (7), 1099 IEEE.
10. Apacheorg. (2016). Apacheorg. Retrieved 18 November, 2016, from http://spamassassin.apache.org/publiccorpus/ Azad, B. Identifying Phishing Attacks.
11. Basnet, R. B., & Sung, A. H. (2010). Classifying phishing emails using confidence-weighted linear classifiers. In International Conference on Information Security and Artificial Intelligence (ISAI) (pp. 108-112)IEEE.
12. Bergholz, A., Chang, J. H., Paass, G., Reichartz, F., & Strobel, S. (2008, August). ImprovedPhishing Detection using Model-Based Features. InCEAS.
13. Cao, Y., Han, W., & Le, Y. (2008). Anti-phishing based on automated individual white-list. In Proceedings of the 4th ACM workshop on Digital identity management (pp. 51-60) IEEE.
14. Chandrasekaran, M., Narayanan, K., & Upadhyaya, S. (2006, June). Phishing email detectionbased on structural properties. In NYS Cyber Security Conference (pp. 1-7). 57 IEEE.
15. Chhabra, S. (2005). Fighting spam, phishing and email fraud (Doctoral dissertation,University of California Riverside).\
16. Gansterer, W. N., & Pölz, D. (2009). E-mail classification for phishing defence. In Advances in Information Retrieval (pp. 449-460). Springer Berlin Heidelberg.
17. Jain, A., & Richariya, V. (2011). Implementing a web browser with phishing detection techniques. arXiv preprint arXiv:1110.0360.
18. Khonji, M., Iraqi, Y., & Jones, A. (2013). Enhancing phishing E-Mail classifiers: a lexical URL analysis approach. International Journal for Information Security Research (IJISR), 2(1/2).

19. Kumar, R. K., Poonkuzhali, G., & Sudhakar, P. (2012, March). Comparative study on email spam classifier using data mining techniques. In Proceedings of the International MultiConference of Engineers and Computer Scientist (Vol. 1, pp. 14-16).
20. Ma, L., Yearwood, J., & Watters, P. (2009, September). Establishing phishing provenance using orthographic features. In eCrime Researchers Summit, 2009. eCRIME'09. (pp. 1-10).IEEE.
21. Monkeyorg. (2016). Monkeyorg. Retrieved 18 November, 2016, from http://monkey.org Nizamani, S., Memon, N., Glasdam, M., & Nguyen, D. D. (2014). Detection of fraudulent emails by employing advanced feature abundance. Egyptian Informatics Journal, 15(3), 169-174.
22. Pandey, M., & Ravi, V. (2012, December). Detecting phishing e-mails using text and data mining. In Computational Intelligence & Computing Research (ICCIC), 2012 IEEE International Conference on (pp. 1-6). IEEE.
23. Mosley Jr, R. C. (2012). Social media analytics: Data mining applied to insurance Twitter posts. In Casualty Actuarial Society E-Forum (Vol. 2, p. 1).
24. Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. Expert Systems with Applications, 36(7), 10760-10773.
25. Malouf, R., & Mullen, T. (2008). Taking sides: User classification for informal online political discourse. Internet Research, 18(2), 177-190.
26. Piao, S., Ananiadou, S., Tsuruoka, Y., Sasaki, Y., & McNaught, J. (2007, January). Mining opinion polarity relations of citations. In International Workshop on Computational Semantics (IWCS) (pp. 366-371).
27. Thomas, M., Pang, B., & Lee, L. (2006, July). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In Proceedings of the 2006 conference on empirical methods in natural language processing (pp. 327-335). Association for Computational Linguistics.
28. Taboada, M., Gillies, M. A., & McFetridge, P. (2006, May). Sentiment classification techniques for tracking literary reputation. In LREC workshop: towards computational models of literary analysis (pp. 36-43).