International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 1 pp. 363-367 Jan 2018-June 2018 www.csjournals.com

Spam Detection using Neural Network (NN)

[1]Sanjeev Kumar, [2]SunilaGodara
[1,2]Department of CSE, Guru Jambheshwar University of Science & Technology, Hisar, India

Abstract: Human capacity is limited and he/she cannot prevent and detect all the phishing but the machine can be made intelligent. Neural Network (NN) is a machine learning technique and widely used for Spam Detection. Spamming is the method for mishandling an electronic informing framework by sending spontaneous mass messages. In this paper, Neural Network (NN) machine learning technique is utilized as spam detector. It can make a model that show the estimation of an objective variable dependent on different information factors. The NN is a supervised learning model that has learning algorithms and the ability to analyze data for classification. Given a set of training examples, NN can decide whether an email belongs to the "spam" or "good" email category. Microsoft Azure is used as a tool to compute various evaluation metrics (accuracy, F score, precession and recall). NN using Pearson Correlation as feature selection outperforms others.
Keywords: phishing, spam, feature selection, machine learning, NN.

I. Introduction

The misuse of electronic messaging systems to casually send unsolicited emails is called "spamming". It is very common for email user to find a high rate of spam emails from unknown senders in his mailbox. Spamming has also introduced cyber fraud on the internet, most of which starts from an email from an unreliable source containing a URL that, when opened, compromises one's personal information. Spamming is economically viable because spammers can manage their mailing lists at a low cost. Due to the minimal investment by the spamming business, the number of spammers and spam emails has been increased. This has resulted in a system in which every email has become a suspect, leading to substantial investment in counter measures, such as the development of spam filtering software, anti-spam software, the creation of domain name server black lists (DNSBL) and white lists, and analysis of spammer activities. Figure 1 shows Automated Spam Detection using NN Machine Learinng technique.

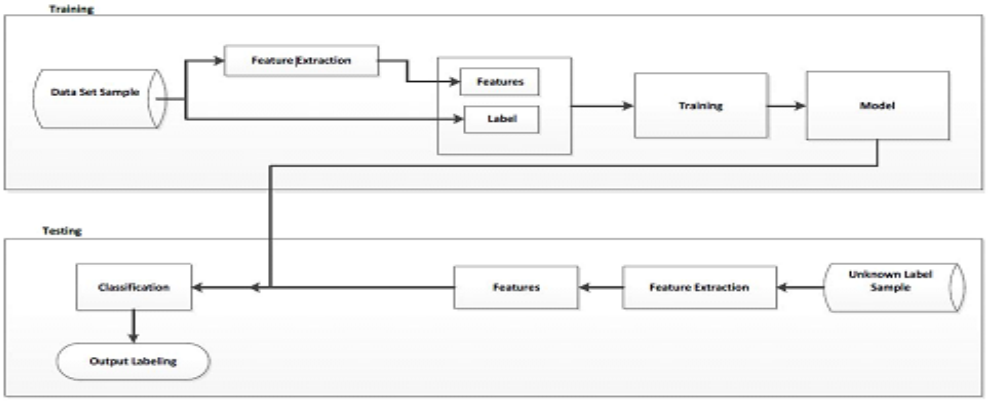

Figure 1: Automated Spam Detection

Microsoft Azure platform provides tools for machine learning. The NN is a supervised learning model that has learning algorithms and the ability to analyze data for classification. Given a set of training examples, NN can decide whether an email belongs to the "spam" or "good" email category. Pang et al. [1] , and that standard machine learning procedures completely beat human-created baselines.

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 1 pp. 363-367 Jan 2018-June 2018 www.csjournals.com

Notwithstanding, the three machine learning strategies we utilized (Naive Bayes, most extreme entropy arrangement, and bolster vector machines) don't execute too on assessment classification as on customary theme based classification. Witten et al. [2] break down the impact of different highlights in spam identification.Work watch that the audit spammer reliably composes spam. This gives another view to distinguish survey spam and can recognize if the creator of the audit is spammer.

McGregor et al. [3] show a strategy, in view of machine realizing, that can separate the follow into groups of traffic where each bunch has different traffic qualities. Jonathon et al. [4] Recognize a bit of content as indicated by its creator's general inclination toward their subject, be it positive or negative. Work shows that match regarding space and time is additionally vital, and presents primer examinations with preparing information marked with emails, which has the capability of being autonomous of area, subject and time. Kotsiantis et al. [5] portrays different administered machine learning characterization procedures. Abu-Nimeh et al. [7] examined about the precision of a few machine learning techniques like Logistic Regression, Classification and Regression Trees, Bayesian Additive Regression Trees, Support Vector Machines, Random Forests, and Neural Networks for anticipating phishing messages. An informational collection of 2889 phishing and genuine messages is used as a part of the relative examination. In this research work, Spam Detectionis performed using Neural Networks. The Azure machine learning studiois used in this research work.

Basnet et al. [11] Study an identification approach that uses promptly gained highlights from the email's substance without using heuristic-based phishing highlights. This methodology depended on Confidence-Weighted Linear Classifiers proposed by Basnet. Pictures are created by Phishers from the message's content and this graphical information passes the phishing channel.Ping et al. [12] proposed a b-bit hashing direct learning calculation for SVM to tackle vast scale attack . The Count-Min (CM) and Vowpal Wabbit (VW) calculations, which have basically identical changes from arbitrary projections . Results represent that hashing is more precise than VW for the paired information.

Almomani et al. [13] present a review of the cutting edge inquire on assaults. This is the first extensive overview to examine strategies for assurance against phishing email assaults in detail. A relative report and assessment of filtering techniques was done. Kumar et al. [14] utilized TANAGRA information mining tool on spam dataset to assess the productivity of the messages classifier. Calculations were connected on the informational index. Using Fisher spam results accomplished 99% precision in recognizing spam.

II.Neural Network(NN)

A neural system is an arrangement of interconnected layers. The information sources are the primary layer, and are associated with a yield layer by a non-cyclic chart included weighted edge. Most prescient errands can be refined effectively with just a single or a couple of shrouded layers.

A neural network as shown in Fig 2, is a set of interconnected nodes.The first layer is input layer which is connected to the hidden layer and this hidden layer is connected to the output layer.

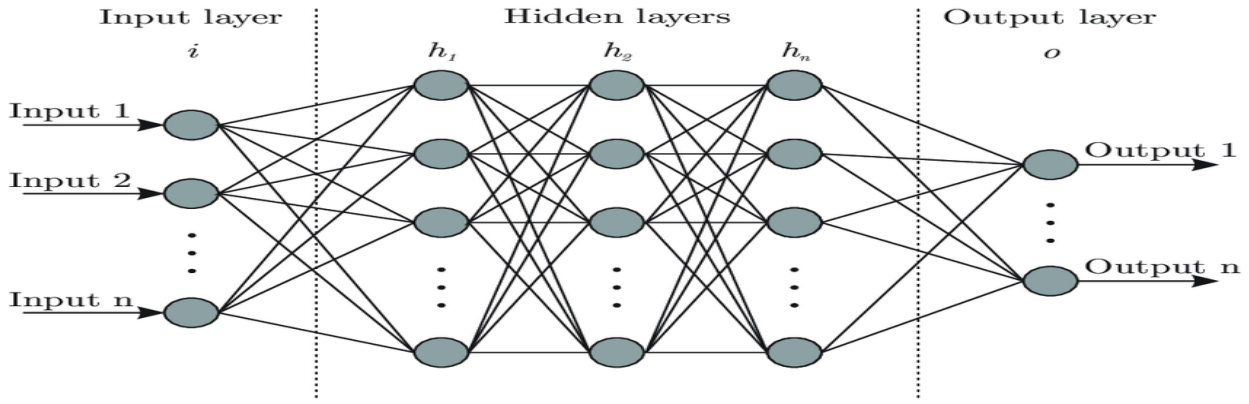

Fig2:NN working Model

Input layer : this layer is used for the input
Hidden layer: This layer represents the intermediate calculation and calculates threshold weighted sum of the input.
Output layer: represent the output.

III.Evaluation approach

This section describes about the data set and about evaluation metrics that are used in comparison.

A. Data Set descriptionand Evaluation metrics

The information used contains 2,000 marked messages for preparing and 100 named messages for testing. Each message is marked either spam or ham (not spam). While assessing spam following measurements are used:
Accuracy: this will gauge the level of the right consequence of an order to demonstrate.
Precision: this is a level of genuine forecast that is right.
Recall: this s a small amount of positive occurrence that was anticipated as positive and give all the right outcome returned by demonstrating.
F-Score: it is figured as the heaviness of accuracy and reviews normally.

IV.Experimental Results

In this section Experimental results shows the predictive accuracy,f1 score, precession and recall of NN using various feature selection methods as shown in table 1 like Pearson correlation,chi squared and Kendall correlation.Vowpal Wabbit is a fast machine learning framework used by Feature Hashing is used in this research work.Feature selection refers to the process of applying statistical tests to inputs, given a specified output, to determine which columns are more predictive to the output. Azure Machine Learning also supports feature value counts as an indicator of information value. Various feature selection methods are given as:

Chi square method for feature selection[7]: Chi squared is a statistical method that measure expected value and tells actual result are how closer to each other. This method assumes that variables are randomly drawn from independent variables . Based on the null hypothesis that the two events are independent, we can calculate the expected value $E_A$using the following formula:

$$\frac{E_A}{X+Z} = \frac{X+Y}{Q}$$

So,

$$E_A = (X+Z)\frac{X+Y}{Q}$$

Using the formula of Chi Square test:

$$\chi^2 = \frac{1}{d}\sum_{k=1}^{n}\frac{(O_k - E_k)^2}{E_k}$$

Pearson Correlation method for feature selection[11]: this model is otherwise called a measurable model and for any two factors, it restores the quality of relationship. Pearson relationship coefficient is registered by taking two factors and partition these factors by the result of their standard deviation. Any difference in scale in the two variable does not impact coefficient. It is used as a measure for quantifying linear dependence between two continuous variables X and Y. Its value varies from -1 to +1. Pearson's correlation is given as:

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

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.

Method	Requirements
Pearson Correlation	Label can be text or numeric. Features must be numeric.
Kendall Correlation	Label can be text or numeric but features must be numeric.
Chi Squared	Labels and features can be text or numeric. Use this method for computing feature importance for two categorical columns.

Table1: various feature selection methods and their requirements

Feature scoring method	Accuracy	F score	Precision	Recall
Pearson Correlation	0.9431	.9601	.9430	1
Chi Square	0.942	.959	.9407	0.999
Kendall	0.941	.9579	.9334	0.999

Table 2:Comparison of various feature selection methods on basis of accuracy ,F score, precession and recall

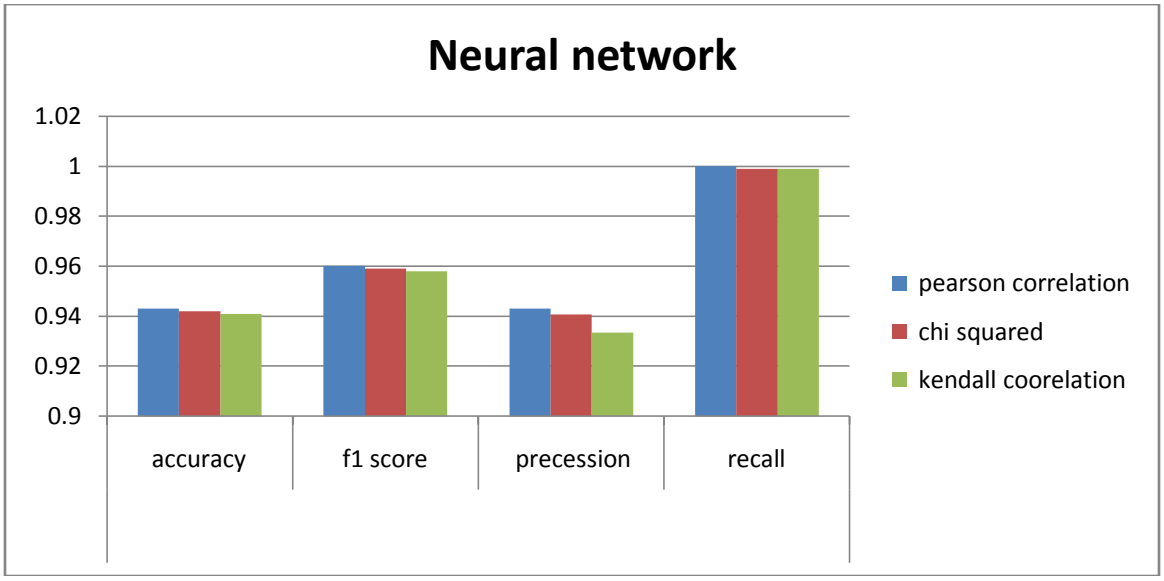

Fig3: Comparison of accuracy, f1 score,precision an recall of various feature selection methods on NN

Table 2 compares neural network using various feature selection methods like pearsoncorrelation,chi squared method and kendall correlation . As shown in table 2 accuracy of neural network by using pearson correlation method for feature selection is 0.9431, by using chi squared test for feature selection is 0.942 and by using kendall correlation for feature selection is 0.941 . F1 score of neural network by

using pearson correlation method for feature selection is 0.9601, by using chi squared test for feature selection is 0.959 and by using kendall correlation for feature selection is 0.9579 . Precision of neural network by using pearson correlation method for feature selection is 0.9430, by using chi squared test for feature selection is 0.9407 and by using kendall correlation for feature selection is 0.9334 . Recall of neural network by using pearson correlation method for feature selection is 1, by using chi squared test for feature selection is 0.999 and by using kendall correlation for feature selection is 0.999 .Fig 3 shows graphical representation of comparison of accuracy, f1 score,precession an recall of various feature selection methods on NN. Enhancement in results is obtained due to Vowpal Wabbit which is a fast machine learning framework used by Feature Hashing ,which is used to hashes feature word into n memory indexes ,by using hash functions.

V. Conclusion

This paper proposesa framework using Neural Network model (NN) machine learning systems to beat the spam issue. Feature selection methods,Pearson CorrelationChi Square and Kendall Correlationare applied toselect the features. Neural network using Pearson Correlation gives better accuracy, Precision and Recall as compared to others.

REFERENCES

1. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up: slant arrangement utilizing machine learning systems." In Proceedings of the ACL-02 meeting on Empirical techniques in normal dialect preparing Vol.10, pp. 79-86, 2002.
2. Witten, Ian H., Eibe Frank, Mark A. Lobby, and Christopher J. Buddy." Information Mining: Practical machine learning devices and systems." 2016.
3. McGregor, Anthony, Mark Hall, Perry Lorier, and James Brunskill. "Stream bunching utilizing machine learning strategies." In International Workshop on Passive and Active Network Measurement ,Springer, Berlin, Heidelberg, pp. 205-214, 2004.
4. Read, Jonathon. "Utilizing emojis to lessen reliance in machine learning methods for slant characterization." In Proceedings of the ACL understudy investigate workshop, relationship for Computational Linguistics, pp. 43-48, 2005.
5. Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Regulated machine taking in: An audit of arrangement strategies." Emerging man-made reasoning applications in PC building 160 pp 3-24.,2007.
6. Rathi, M., & Pareek, V."Spam Mail Detection through Data Mining-A Comparative Performance Analysis". International Journal of Modern Education and Computer Science,(12), 31, 2013.
7. Abu-Nimeh, Saeed, Dario Nappa, Xinlei Wang, and Suku Nair. "An examination of machine learning procedures for phishing identification." In Proceedings of the counter phishing working gatherings second yearly eCrime analysts summit, pp. 60-69 , 2007.
8. Sommer, Robin, and Vern Paxson. "Outside the shut world: On utilizing machine learning for arrange interruption location." IEEE , pp. 305-316 , 2010.
9. Kolari, Pranam, Akshay Java, Tim Finin, Tim Oates, and Anupam Joshi. "Distinguishing spam writes: A machine learning approach." In AAAI, vol. 6, pp. 1351-1356. 2006.
10. Crawford, Michael, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter, and Hamzah Al Najada. "Overview of audit spam location utilizing machine learning systems." Journal of Big Data 2, no. 1: 23,2015
11. Basnet, R. B., and Sung, A. H. (2010). "Ordering phishing messages utilizing confidence weighted direct classifiers".In International Conference on Information Security and Artificial Intelligence (ISAI) pp. 108-112,2010.
12. Li, Ping, Anshumali Shrivastava, Joshua L. Moore, and Arnd C. König. "Hashing algorithms for large-scale learning." In Advances in neural information processing systems, pp. 2672-2680. 2011.
13. Almomani, Ammar, B. B. Gupta, Samer Atawneh, A. Meulenberg, and Eman Almomani. "A survey of phishing email filtering techniques." IEEE communications surveys & tutorials 15, no. 4 pp.2070-2090.,2013.
14. Kumar, R. K., Poonkuzhali, G., and Sudhakar, P. Similar investigation on email spam classifier utilizing information mining procedures. In Proceedings of the International Multi Conference of Engineers and Computer Scientist Vol. 1, pp. 14-16,march-2012.