Evaluation of Phishing Email Classification Features: Reliability Ratio Measure

Melad Mohamed Al-Daeef , Nurlida Basir, and Madihah Mohd Saudi

Abstract— **Heuristic-based anti-phishing systems are widely implemented to detect phishing attacks. Selecting most reliable classification features however, is a challenging task. *Information Gain IG, Gain Ratio GR, Term Frequency-Inverse Document Frequency TF-IDF, Chi-Square* are examples of measures that have proven their excellence in text classification field. These measures have also been used to evaluate phishing classification features. Phishing emails however, are difficult to be detected based only on their subject and content texts since they are usually constructed to look like legitimate ones. Text classification measures may produce high error rates if they are naively employed to detect phishing instances. Some attempts therefore have been done to adapt them to evaluate phishing classification features. *Average Gain AG* for example, which is an *IG*-dependent measure, was used to adapt *IG* measure to be used in phishing classification field. In this study, *Reliability Ratio RR* measure is proposed to evaluate the reliability of phishing email classification features. Experimental results have proven the effectiveness of the proposed *RR* measure compared with other evaluated measures such as *IG* and *AG*.**

Index Terms— **Evaluation measure, Phishing email, Classification, Feature reliability, Suspicion level.**

I. INTRODUCTION

Heuristic-based tools are commonly used to combat phishing attacks. Heuristics are more reliable than other approaches such as black and white lists especially in detecting zero hour attacks. Results' accuracy of heuristic-based tools however, depends on the quality of employed classification features. In many cases, some features might be employed even they are not enough informative which leads to inaccurate discriminative decisions [1]. Real-world datasets, especially large ones are commonly contain noisy data that will also cause classification feature to produce inaccurate results. There are two general types of noise sources, attribute or feature noise and class noise. Attribute noise is the errors in attribute values. Possible sources of class noise include the contradictory instances, i.e., same instance with different class labels, or instances that labeled with wrong classes [2],[3]. Data cleaning process which is usually laborious and time consuming is error prone process and may also cause noise in datasets [4]. Not all classification features therefore can be at the same reliability level. If some of them have been extracted from wrongly labeled instances, they will definitely produce high *False Positive FP* or *False Negative FN* results [4],[5],[6]. Numerous of evaluation measures thus, have been used to evaluate the reliability of phishing classification features.

Information Gain IG, Gain Ratio GR, Chi-Square and *Term Frequency-Inverse Document Frequency TF-IDF* are examples of evaluation measures that basically implemented, and have proven their excellence, in text classification field [7],[8]. These measures however, have been widely used to evaluate email classification features. Such measures worked well for evaluating the reliability of spam email classification features since spam emails can be classified based on their content and subject texts. Phishing emails are usually constructed to look like legitimate ones, text classification measures may therefore produce high rates of false results when they are naively applied to evaluate phishing classification features [9]. This may highly occur if evaluated features are heterogeneous in their nature, or their values are not in the same range. For example, if the values of some features are in continues form (real or integer numbers) whereas the values of others are in binary or categorical form (fall into a set of finite values such as [0-1]) [10]. Such a problem becomes more severe when features are extracted from different email parts because they will have variant occurrences. *Keyword-based* features for example will definitely have more occurrences than *URL* or *Subject-based* features [11]. Combining heterogonous features together in a single clustering algorithm is another problem that requires further processing [10],[12]. In addition to that, text classification oriented measures may not well applicable to evaluate phishing classification features since many of legitimate emails and web sites may include sensitive keywords in their contents [13]. Studies have shown the limitations of *IG, GR, Chi-Square*, and *TF-IDF* measures that make them unsuitable to evaluate phishing classification features. These limitations are discussed in section II. Researchers in many studies have attempted to overcome the limitations of such measures. *Average Gain AG* measure for example, was introduced in [14] to overcome the limitations of *IG* and *GR* measures.

In this study, *Reliability Ratio RR* measure was proposed to evaluate the reliability level of phishing classification features. Based on its *RR* value, each feature is assigned to one of: *High, Medium* or *Low Suspicion Level SL* categories. The implementation of *RR* measure was inspired from the fact that, reliable feature will definitely produce higher *TP*

Manuscript received June 29, 2016; revised July 17, 2016.

Melad Mohamed Al Daeef is a PhD Student at the Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM), Bandar Baru Nilai 71800, Negeri Sembilan Darul Khusus, Malaysia Hand phone: 0060-18250-3435; e-mail: meladmohalda@gmail.com

Dr. Nurlida Basir is a lecturer at the Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM), Bandar Baru Nilai 71800, Negeri Sembilan Darul Khusus, Malaysia e-mail: nurlida@usim.edu.my

Assoc. prof. Dr. Madihah Mohd Saudi is a lecturer at the Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM), Bandar Baru Nilai 71800, Negeri Sembilan Darul Khusus, Malaysia e-mail: madihah@usim.edu.my

than *FP* results. *TP* stands for *True Positive*, i.e. phishing email is correctly identified as phishing, whereas, *FP* stands for *False Positive*, i.e. legitimate email is incorrectly identified as phishing. Section III shows the experiments of calculating *RR*, *IG* and *AG* values of all participated classification features. Experimental results show that, *RR* measure can overcome the limitations of *IG*, and thus can be relied upon to participate in building reliable anti-phishing systems. Conducted experiments also show the straightforward process of calculating *RR* compared with the process of *AG* calculation. This makes *RR* more favoured than *AG* measure although they have achieved same results. In addition to that, *RR* measure can also tolerate class noise of analyzed datasets by not omitting less informative features from classification process. Less informative features are instead, assigned to a *Low SL* category and allow anti-phishing systems have more decision choices.

The rest of this paper is organized as follows; section II gives an overview on some research studies in that, the importance of phishing classification features was evaluated using aforementioned evaluation measures. Results from three different experiments using *RR*, *IG* and *AG* measures to evaluate the same set of 12 phishing classification features are presented in Section III. Lastly, the conclusion and future work of this paper are presented in Section IV.

II. RELATED WORK

This section reviews some studies in which, different evaluation measures were used to evaluate the reliability of phishing classification features.

In [15], researchers have implemented *IG* measure to evaluated the importance of 40 features extracted from three different ham, spam, and phishing emails' datasets. *IG* values was then used to assign evaluated features to three different groups; best, medium and worst. Researchers at the end have deemed nine features that appeared in the top 10 over the three datasets as the best features. *IG* measure was also used in [16] to evaluate the effectiveness of a small set of 7 phishing classification features. As stated by the authors of [16], many of potential features have not been considered in their experiment. [17] is another study in which, *IG* measure was used to chose the best classification features amongst 47 features to classify phishing emails. In [18] also, researchers have employed *IG* measure to chose best 10 amongst 22 classification features to classify malicious short URLs. Although it has been used in many studies, *IG* measure however is criticized for its bias toward features with higher values even they are not enough informative [14],[19],[20],[21]. *GR* measure has been used as an alternative to overcome the limitations of *IG* measure. In [19] for example, researchers have used *GR* to evaluate 30 classification features to classify emails as: ham, spam or phishing. As an opposite of *IG*, *GR* bias toward features with small values [21].

Chi-Square measure was also used in many studies to evaluate phishing classification features. In [22] researchers have evaluated the importance of 17 features to predict phishing websites. It was also used in [8] as a feature selection technique to classify emails using text-based approach. *Chi-Square* however, behaves erratically toward features with very small counts. This is a common phenomenon in text classification field where some classification words have uncommon occurrences [23]. Such a behavior of *Chi-Square* measure, especially with heterogeneous features, will badly affect the results of evaluation process.

TF-IDF is a content-based measure that used to weight words in text classification processes. It is often used in information retrieval and text mining fields to evaluate the importance of a specific word to a document, or an email to a dataset [24],[25]. *TF-IDF* measure was used in [25] to evaluate selected keywords that extracted from subject and body parts of analyzed messages to filter spam emails. In CANTINA [26], which is a content-based approach, authors have employed some other features to avoid high *FP* results that caused by using *TF-IDF* measure.

In addition to limitations of implemented evaluation measures, researchers in many studies did not take into account the noisy data found in analyzed datasets. Such noisy data will definitely affect the accuracy of obtained results [5]. The herein work proposes *Reliability Ratio RR* measure as an attempt to correctly evaluate the reliability level of phishing email classification features. Proposed *RR* measure can also tolerate noisy data found in analyzed datasets.

III. FEATURE SUSPICION LEVEL

The *Suspicion Level SL* category of each feature is determined after its *RR* value is calculated. *RR* measure was motivated based on the fact that, the most reliable and informative feature is the one that produces higher *TP* than *FP* results. Based on its *RR* value, each feature then is assigned to one of; *High, Medium* or *Low SL* categories.

In order to prove the effectiveness of the proposed *RR* measure, same classification features were evaluated in three different experiments using *RR, IG* and *AG* measures. Obtained *RR* results are then compared with the results obtained from employing *IG* and *AG* measures. Same two datasets of more than 13000 legitimate and phishing emails were used in all three experiments. Phishing emails dataset comprised of 3240 emails from [27]. Legitimate emails dataset comprised of 10000 emails from three different sources, they are [28],[29],[30]. Evaluated features are presented in each of Table I, Table II and Table III. Section III.A presents the formula and the process of implementing *RR* measure. In sections III.B and CIII.C, *SL* categories of all features were determined using *IG* and *AG* measures respectively. Achieved results from employing the three measures were then compared to prove the effectiveness of the proposed *RR* measure.

A. Reliability Ratio RR Measure

RR value of a given feature *f* is the ratio between the percentages of *TP* and *FP* results that produced by this feature. Since analyzed datasets are different in their sizes, *TP* and *FP* results therefore have presented in the percentage form. P_{TP} and P_{FP} in Table I represents the percentage values of *TP* and *FP* results of participated classification features. P_{TP} and P_{FP} values are then used to calculate *RR* value of each feature to determine its *SL* category as shown in Table I. The *SL* of any phishing email is then determined based on the *Suspicion Level* category of classification

ISBN: 978-988-14047-4-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

feature(s) upon which this email was identified as phishing. Equation (1) is used to calculate *RR* values;

$$R_R(f) = \frac{P_{TP}(f)}{P_{FP}(f)} \quad (1)$$

where $R_R(f)$ is the *RR* value of a given feature *f*, $P_{TP}(f)$ is the percentage value of *TP* result that produced by the feature *f*, and $P_{FP}(f)$ is the percentage value of *FP* result that caused by the same feature.

1) Results

When aforementioned datasets were analyzed, 2892 out of 3240 phishing emails have been correctly identified as phishing, this number represents the total of *TP* results that achieved by all 12 participated features. On the other side, there was a number of 685 out of 10000 legitimate emails were incorrectly identified as phishing, this number represents all *FP* results that caused by all 12 features. $P_{TP}(f)$ and $P_{FP}(f)$ values of a given feature *f* are calculated using *TP* and *FP* results produced by the feature *f*, and the total of *TP* and *FP* results that produced by the all 12 participated features. *TP* and *FP* results of each feature are presented in Table II. Based on that, P_{TP} and P_{FP} values of *Imghttps* classification feature for example, are calculated as follows; $P_{TP}(Imghttps) = \frac{228}{2892} = 0.079$, whereas $P_{FP}(Imghttps) = \frac{5}{685} = 0.007$. Same procedure was applied to calculate P_{TP} and P_{FP} values of all other features.

Results in Table I show that, some features have achieved $R_R(f) > 1$, whereas some others have their $R_R(f) < 1$. It seems that, features with $P_{TP} > P_{FP}$ values have achieved *RR* > 1, whereas features with $P_{TP} < P_{FP}$ have achieved low *RR* values. *RR* value of *Imghttps* feature for example was 11.285 which is the highest *RR* value compared with other features. Equation (1) was applied on 0.079 P_{TP} and 0.007 P_{FP} results of *Imghttps* feature as follows;

$$R_R(Imghttps) = \frac{0.079}{0.007} = 11.285$$

RR value 11.285 means that, 0.079 is 11.285 times of 0.007, i.e. *Imghttps* feature has contributed 11.285 times in producing *TP* than in causing *FP* results. The 0.031 *RR* value of *FormTag* feature on the other side means that, *TP* result that produced by *FormTag* feature is only 0.031 times of the *FP* result that caused by this feature, i.e. 0.002 is only 0.031 times of 0.064. 0.031 is a very small *RR* value compared with *RR* value of *Imghttps* feature which is 11.285. Obtained results have put *Imghttps* at the top of the list as the most informative feature, and *FormTag* feature at the tail of this list as less informative feature.

2) Thresholds

This section shows how to determine the thresholds that are used to draw the borders between *SL* categories. Three *SL* categories are defined in this work, they are; *High, Medium* and *Low*. Since *RR* values of some features are below 1, whereas other *RR* values of some features are higher than 1 as shown in Table I, thus number 1 is used as the first threshold point between *Low* and *Medium SL* categories. Based on that, 3 out of 12 classification features have assigned to the *Low SL* category, they are, *MoreThanOneDomainURL, @Character* and *FormTag*.

The arithmetic mean was applied on $R_R(f)$ values of the remaining 9 features to determine the threshold point between *High* and *Medium SL* categories. Equation (2) was applied on $R_R(f)$ values of the features from 1 to 9 for that purposes.

Table I The Distribution of Classification Features Amongst *SL* Categories

No	Feature Name	P_{TP}	P_{FP}	$R_R(f)$	SL
1	Imghttps	0.079	0.007	11.286	High
2	OnMouseOver	0.033	0.007	4.714	
3	URLHEXcode	0.026	0.006	4.333	
4	DMNSemantics	0.056	0.016	3.500	Medium
5	URL_IP	0.410	0.126	3.254	
6	FldrNameLength	0.085	0.039	2.179	
7	URLKeyWord	0.643	0.434	1.482	
8	DMNDashes&Dots	0.313	0.264	1.186	
9	PortNumber	0.065	0.064	1.016	
10	MoreThanOne-DomainURL	0.049	0.054	0.907	Low
11	@Character	0.002	0.053	0.038	
12	FormTag	0.002	0.064	0.031	

$$T_{sh}(SL_n, SL_{n+1}) = \sum_{f=1}^{n} R_R(f)\frac{1}{n} \quad (2)$$

whereas $T_{sh}(SL_n, SL_{n+1})$ is the threshold point between two *SL* categories, $R_R(f)$ is the *RR* value of a given feature *f*, and *n* is the number of features that to be distributed amongst the *Medium* and *High SL* categories. Second threshold point is calculated as follows;

$$\begin{aligned} T_{sh}(SL_1, SL_2) &= (11.286 + 4.714 + 4.333 + 3.500 \\ &\quad + 3.254 + 2.179 + 1.482 + 1.186 \\ &\quad + 1.016)/9 \\ &= 32.950/9 = 3.661 \end{aligned}$$

Based on $T_{sh}(SL_1, SL_2)$ result, features with $R_R(f) > 3.661$ are assigned to the *High SL* category, whereas the features with $3.661 > R_R(f) > 1$ are assigned to the *Medium SL* category. Features with $R_R(f) < 1$ have already been assigned to the *Low SL* category. Based on that, 3 out of 9 features are assigned to the *High SL* category, they are; *Imghttps, OnMouseOver,* and *URLHEXcode*, whereas the other 6 features are assigned to the *Medium SL* category, they are; *DMNSemantics, URL_IP, FldrNameLength, URLKeyWord, DMNDashes&Dots,* and *PortNumber*.

B. Information Gain

IG is the expected reduction in entropy that caused by splitting the dataset according to a given feature to evaluate the reliability of that feature in classification process. Entropy [31] is a very common measure used in information theory that characterizes the impurity of a collection of datasets. Equation (3) is used to calculate the Entropy $\mathrm{E}(s)$ of a given collection of datasets *S*. Entropy is calculated herein as a prior requirement to calculate *IG* values.

$$E(s) = \sum_{i=1}^{n} -p_i \log_2 pi \quad (3)$$

where *n* is the number of classes in the entire collection of datasets, in this study there are two classes; legitimate and phishing emails. And p_i is the probability that a particular feature belongs to class i. Equation (4) is used then to calculate the *IG* value of each classification feature.

ISBN: 978-988-14047-4-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$IG(s,f) = E(s) - \sum_{v \in values(f)} \frac{s_v}{s} E(s_v) \qquad (4)$$

where $IG(s,f)$ is the IG of a given feature f over the dataset s, $E(s)$ is the entropy of the entire dataset as calculated using (3), s_v is the number of features in s where f has the value v, and $E(s_v)$ is the entropy of this subset of the dataset.

Table II presents the *TP, FP, TN* and *FN* results that obtained from analyzing the two mentioned datasets. These results are used to calculate *IG* values of all features. *IG* results as in Table II show that, *URLKeyWord* feature has obtained the highest *IG* value compared with other features.

Table II Features Arranged in
Descending Order Based on Their *IG* Values

Feature Name	TP	FP	TN	FN	IG
URLKeyWord	1858	297	9703	1382	0.2543
URL_IP	1184	86	9914	2056	0.1702
DMNDashes&Dots	906	181	9819	2334	0.1016
Imghttps	228	5	9995	3012	0.0331
FldrNameLength	247	27	9973	2993	0.0299
DMNSemantics	163	11	9989	3077	0.0212
PortNumber	187	44	9956	3053	0.0181
MoreThanOneDomain-URL	142	37	9963	3098	0.0132
OnMouseOver	95	5	9995	3145	0.0127
URLHEXcode	76	4	9996	3164	0.0102
FormTag	7	44	9956	3233	0.0002
@Character	5	36	9964	3235	0.0002

C. Average Gain

To ovoid the limitations associated with *IG* measure, researchers in [14] for example have introduced an improved feature selection measure which called *Average Gain AG.* It was motivated after the idea of penalizing features with high values by dividing their *IG* values by the number of their occurrences. Equation (5) [14] was used to calculate *AG* value of each feature as follows;

$$AG(s,f) = \frac{IG(s,f)}{|f|} \qquad (5)$$

where $AG(s,f)$ is the *Average Gain* of a given feature f over the dataset s, $IG(s,f)$ is the IG value of the feature f as calculated in section B, and $|f|$ is the occurrences number of a given feature f, here it stands for *(TP+FP)* results of each evaluated feature. Table III shows that, evaluated features are arranged in descending order based on their *AG* values.

D. Discussion

When the results of *RR* are compared with the results of *AG* measures, it can be seen that there is a slight difference in the sort or the arrangement of the features as shown in Table I and Table III. Although of this arrangement difference, all features however, still have the same distribution amongst the defined *SL* categories in both of the two tables. *IG* results as in Table II on the other side show *URLKeyWord* at the top of the list supposing it as the most informative feature. This however not necessarily be always the reality since it has been claimed that, many legitimate emails and web sites may contain sensitive key words in their contents [11],[13]. *IG* measure furthermore, has been criticized for its bias toward features with high occurrences [14],[19],[20],[21]. *IG* values in Table II show that, classification features have been arranged almost based only on their *TP* occurrences. Although many of the features have high *FP* frequencies, they sit however at the top of the list. These limitations of *IG* were eliminated using *Average Gain* as has been introduced in [14].

Table III Features Arranged in
Descending Order Based on Their *AG* Values

Feature Name	IG	TP+FP	AG	SL
Imghttps	0.0331	233	0.0001421	High
URLHEXcode	0.0102	80	0.0001275	
OnMouseOver	0.0127	100	0.0001270	
URL_IP	0.1702	1270	0.0001255	Medium
DMNSemantics	0.0212	174	0.0001218	
URLKeyWord	0.2543	2055	0.0001180	
FldrNameLength	0.0299	274	0.0001091	
DMNDashes&Dots	0.1016	1087	0.0000935	
PortNumber	0.0181	231	0.0000784	
MoreThanOneDomain-URL	0.0132	179	0.0000611	Low
@Character	0.0002	41	0.0000049	
FormTag	0.0002	51	0.0000039	

IV. CONCLUSION AND FUTURE WORK

Selecting the most reliable and informative classification feature(s) still a challenge that limits the functionality of anti-phishing systems. *Information Gain IG, Gain Ratio GR, Term Frequency-Inverse Document Frequency TF-IDF* and *Chi-Square* are examples of evaluation measures that were used for this purpose. Although their excellence in text classification field, the nature of these measures however, limit their wellness when they were implemented to evaluate phishing classification features. Researchers have introduced alternative measures to overcome the limitations of some existing measures. *AG* measure for example, was implemented to overcome *IG's* measure associated limitations.

Reliability Ratio RR measure was proposed in this work to perfectly evaluate the efficiency of phishing classification features, it was motivated from the fact that, a given feature *f* is reliable if produces higher *TP* than *FP* results. The feature becomes more reliable as its *RR* value goes higher. Based on the *RR* value and a determined threshold point, each feature has been assigned to one of *High*, *Medium*, or *Low SL* categories as presented in Table I. The reliability of 12 phishing *URL*-based classification features has been evaluated in this work using *IG,AG* and *RR* measures in three different experiments conducted on same phishing and legitimate emails datasets. Results of *IG* experiment show that, evaluated features have almost arranged based on their *TP* occurrences without considering their *FP* results. RR experiment results on the other side show that, the same features have arranged based on their participation in produced *TP* and caused *FP* results. In order to validate its effectiveness, results obtained from implementing the proposed *RR* measure were compared with the results obtained from implementing the *AG* measure. Their results were almost identical in terms of features' distribution amongst *Suspicion Level* categories as shown in Table I and Table III. The benefit of the proposed *RR* measure on *AG* measure however is its straightforward calculation process which can be conducted only using *TP* and *FP* results that produce by each evaluated feature. Going through a complex calculation process on the other side, is a mandatory requirement to calculate *AG* values as has been shown in section III. In addition to that, the proposed *RR* measure can also tolerate class noise found in analyzed datasets. *RR* measure do not omit less informative features from classification process, such features are instead, assigned to a *Low SL* category to allow anti-phishing systems to have more decision choices.

As a future work, *RR* results are going to be compared with the results obtained from applying other measures such as *GR, TF-IDF* and *CHI-Square* on the same feature set and datasets.

REFERENCES

[1] Xiang, G., et al., *Cantina+: A feature-rich machine learning framework for detecting phishing web sites.* ACM Transactions on Information and System Security (TISSEC), 2011. **14**(2): p. 21.
[2] Yang, Y., X. Wu, and X. Zhu, *Dealing with predictive-but-unpredictable attributes in noisy data sources*, in *Knowledge Discovery in Databases: PKDD 2004*. 2004, Springer. p. 471-483.
[3] Zhu, X., X. Wu, and Q. Chen. *Eliminating class noise in large datasets*. in *ICML*. 2003.
[4] Maletic, J.I. and A. Marcus. *Data Cleansing: Beyond Integrity Analysis*. in *IQ*. 2000: Citeseer.
[5] Frénay, B. and M. Verleysen, *Classification in the presence of label noise: a survey.* Neural Networks and Learning Systems, IEEE Transactions on, 2014. **25**(5): p. 845-869.
[6] Gomez, J.C., E. Boiy, and M.-F. Moens, *Highly discriminative statistical features for email classification.* Knowledge and information systems, 2012. **31**(1): p. 23-53.
[7] Lee, C. and G.G. Lee, *Information gain and divergence-based feature selection for machine learning-based text categorization.* Information processing & management, 2006. **42**(1): p. 155-165.
[8] Zareapoor, M., *Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection.* 2015.
[9] Fette, I., N. Sadeh, and A. Tomasic. *Learning to detect phishing emails*. in *Proceedings of the 16th international conference on World Wide Web*. 2007: ACM.
[10] Hamid, I.R.A. and J.H. Abawajy, *An approach for profiling phishing activities.* Computers & Security, 2014. **45**: p. 27-41.
[11] Al-Daeef, M.M., N. Basir, and M.M. Saudi. *A Method to Measure the Efficiency of Phishing Emails Detection Features*. in *Information Science and Applications (ICISA), 2014 International Conference on*. 2014: IEEE.
[12] Abur-rous, M.R.M., *Phishing website detection using intelligent data mining techniques. Design and development of an intelligent association classification mining fuzzy based scheme for phishing website detection with an emphasis on E-banking*. 2011, University of Bradford.
[13] Tian, X.-P., G.-G. Geng, and H.-T. Li. *A framework for multi-features based web harmful information identification*. in *Computer Application and System Modeling (ICCASM), 2010 International Conference on*. 2010: IEEE.
[14] Wang, D. and L. Jiang. *An improved attribute selection measure for decision tree induction*. in *Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on*. 2007: IEEE.
[15] Toolan, F. and J. Carthy. *Feature selection for Spam and Phishing detection*. in *eCrime Researchers Summit (eCrime), 2010*. 2010: IEEE.
[16] Liping Ma, et al., *Detecting Phishing Emails Using Hybrid Features.* 2009.
[17] Vaishnaw, N. and S. Tandan, *Development of Anti-Phishing Model for Classification of Phishing E-mail.* Development, 2015. **4**(6).
[18] Nepali, R.K. and Y. Wang. *You Look Suspicious!!: Leveraging Visible Attributes to Classify Malicious Short URLs on Twitter*. in *2016 49th Hawaii International Conference on System Sciences (HICSS)*. 2016: IEEE.
[19] Gansterer, W.N. and D. Pölz, *E-mail classification for phishing defense*, in *Advances in Information Retrieval*. 2009, Springer. p. 449-460.
[20] Fahad, A., et al., *Toward an efficient and scalable feature selection approach for internet traffic classification*. Computer Networks, 2013. **57**(9): p. 2040-2057.
[21] Novaković, J., P. Štrbac, and D. Bulatović, *Toward optimal feature selection using ranking methods and classification algorithms.* Yugoslav Journal of Operations Research ISSN: 0354-0243 EISSN: 2334-6043, 2011. **21**(1).
[22] Mohammad, R.M., F. Thabtah, and L. McCluskey, *Intelligent rule-based phishing websites classification.* Information Security, IET, 2014. **8**(3): p. 153-160.
[23] Ladha, L. and T. Deepa, *Feature selection methods and algorithms.* International journal on computer science and engineering, 2011. **1**(3): p. 1787-1797.
[24] Ayodele, T., S. Zhou, and R. Khusainov, *Email classification using back propagation technique.* International Journal of Intelligent Computing Research (IJICR), 2010. **1**(1/2): p. 3-9.
[25] Sharma, A.K. and R. Yadav. *Spam Mails Filtering Using Different Classifiers with Feature Selection and Reduction Technique*. in *Communication Systems and Network Technologies (CSNT), 2015 Fifth International Conference on*. 2015: IEEE.
[26] Zhang, Y., J.I. Hong, and L.F. Cranor. *Cantina: a content-based approach to detecting phishing web sites*. in *Proceedings of the 16th international conference on World Wide Web*. 2007: ACM.
[27] *PhishingCorpus.http://monkey.org/~jose/wiki/doku.php*.
[28] Cormack, G.V. and T.R. Lynam. *TREC 2005 Spam Track Overview*. in *TREC*. 2005.
[29] *Spamassassin public corpus, http://spamassassin.apache.org/publiccorpus*
[30] Shetty, J. and J. Adibi, *The Enron email dataset database schema and brief statistical report.* Information sciences institute technical report, University of Southern California, 2004. **4**.
[31] Shannon, C.E., *A mathematical theory of communication.* ACM SIGMOBILE Mobile Computing and Communications Review, 2001. **5**(1): p. 3-55.