IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-16, May 2019 E-ISSN: 2347-2693

A Novel optimal Email Feature Selection Protocol (OEFS) for Detecting Spam Emails

P.Mano Paul[1*], I. Diana Jeba Jingle[2]

[1] Department of Computer Science and Engineering, Presidency University, Bangalore, India
[2] Department of Computer Science and Engineering, Christ University, Bangalore, India

[*]*Corresponding Author: pmanopaul@gmail.com, Tel.: +91-94444-12470*

DOI: https://doi.org/10.26438/ijcse/v7si16.3439 | Available online at: www.ijcseonline.org

***Abstract*—** In this paper, we propose a hybrid rule-based approach, named as Optimal Email Feature Selection (OEFS) Protocol for selecting optimal features to reduce the searching time in detecting spam emails. The OEFS protocol performs email spam detection in four stages: Feature Selection, Normalization of selected features, Rank Assignment and Optimal Feature Selection. The OEFS protocol has been executed and designed for large data amount of data by achieving accurate feature generation. The performance of OEFS analyzed using different protocols in existing systems. The protocol defines here an optimality for email spam detection and correction which provides an optimal solution and outperforms all email filtering protocols like PEP and CRVSM.

***Keywords*—**Optimal Feature, Normalization, Score Assignment, Spam Email

I. INTRODUCTION

Signal Processing to Analyze Malware (SPAM) can be named as a suspicious, informal and fraudulent message that extremely affects the email as it is considered as the best mode of communication which costs nothing to users. Email Spamming [12] devours the cyber resources such as system memory and system size in large volume. It was stated that "Over 85% of the overall emails that are sent are considered to be spam" in a recent report. Networks like Yahoo Mail, Googlemail, Hotmail, and iCloud Mail, are considered as large scale networks and they are extremely affected by email spamming and they are unable to provide efficient service successfully.

Content-based spam filtering methods are used to isolate the spam contents of an email, and they are categorized into different types (as shown in Fig. 1) namely: 1) Adaptive, 2) Rule-based, Language-based and Learning-based filters. Rule-based filters afford accurate detection of spam in an inexpensive way and are suitable for large datasets. Moreover, they are found to be more appropriate for the effective filtering of spams in email contents.

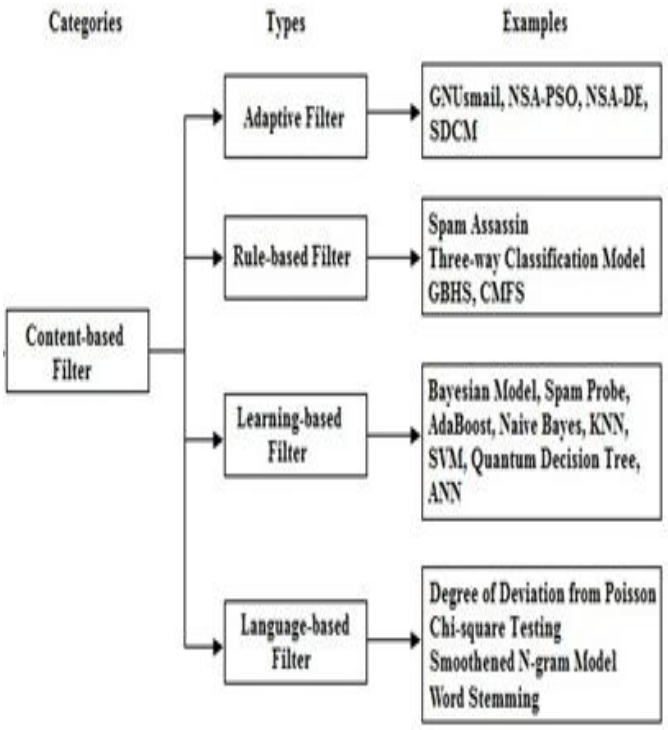

Figure 1 Classification of Rule-based Filters

We propose an Optimal Email Feature Selection (OEFS) Protocol which is a rule-based hybrid method to filter out the spam email messages. The OEFS protocol exploits the

properties of two existing different feature selection approaches that are well-known in filtering spam emails. The proposed OEFS protocol performs feature selection in four different stages namely, Feature Selection stage, Normalization process stage, Score Assignment stage and Optimal Feature Selection stage. We designed the OEFS protocol as a rule-based filtering approach since it is suitable for large volumes of data and attains optimality in accurate feature generation with reasonable complexity. The OEFS performance has been examined using suitable tests. The proposed protocol affords a best solution for spam email detection and beats the existing detection protocols, i.e., PEP and CRVSM.

The paper has been structured as follows: Section I presents the Introduction on SPAM and various content-based filtering methods. Section II presents the survey related to rule-based filters for email spam detection. Section III presents the proposed OEFS protocol, Section IV provides an analysis on the results obtained and performance of OEFS over other protocols. Finally Section V concludes the work.

II. RELATED WORK

In this section, In rule-based filtering method[7], [8], [9] related rules are mined from the data and graded using a ranking approach for predicting and isolating the email. Rule-based filtering methods improves the algorithm's efficiency because, no training period is required for generating rules and such filters can be easily installed. But the difficulty with rule-based filters is, the rules require persistent updating, and therefore it is a troublesome job to keep these rules. Email spammers can simply modify the rules if overtly open. These rules are not flexible to new patterns. Three-way classification [11], Spam Assassin [10], CMFS [2] and GBHS [13] methods are certain rule-based spam filtering methods that were introduced in the past.

Spam Assassin [10] is a rule-based filter that generates different rules for validating every received email and verifies whether the rules match the email messages and computes overall score for that email. The overall score is then matched with the expected score; based on the identified match, the email is finally classified as spam. The problem with this method is, it creates more amount of false alarms.

The three-way classification method stated in [11] uses three different evolutionary indicator-based and decomposition-based multi-objective rule-based algorithms and it optimizes the filtering performance. Accurate classification is accomplished in this method in an inexpensive way. The problem with this method is, it performs high computations with increased complexity.

Global best harmony search (GBHS) [13] method selects two different predetermined thresholds to achieve optimality and accuracy. At the initial level, the most discriminative features are selected using these thresholds, plus an optimal document frequency (ODFFS) method and an optimal term frequency (OTFFS) method. At the final level, the left over unimportant features are selected using only OTFFS and ODFFS methods. GBHS method attains better accuracy and optimality. The problem with this method is, it takes more time for the feature selection process and due to this spam detection process is delayed.

The comprehensive measurement (CMFS) approach [2] measures the importance of a term found among different emails and in an email message. The CMFS approach attains dimensionality lessening of features represented in the feature space and is suitable for large volumes of data. The problem with this approach is, it performs high computations with increased complexity.

III. PROPOSED METHODOLOGY

The proposed Optimal Email Feature Selection (OEFS) protocol has been designed to accurately classify the email messages as spam or ham ones with reduced amount of false alarms. The OEFS protocol also detects spam emails in a short time because feature selection process is rapidly performed with reduced searching time. The OEFS is a rule-based hybrid protocol that exploits the properties of two different feature selection approaches such as: Feature Selection using Comprehensive Measure (CMFS) [2] and Chi-square testing method [1]. Comprehensive Measure-based approach attains dimensionality reduction and can be used for large amount of data; whereas, Chi-square testing achieves automatic filtering of email contents efficiently; and can examine how well the features and associated classes are interdependent. The problem with these approaches are, they perform high computations and their complexity are at acceptable levels. The proposed OEFS protocol performs feature selection in four different stages namely: Feature Selection stage, Normalization stage, Score Assignment stage and Optimal Feature Selection stage. The proposed OEFS protocol for detecting spam emails is depicted in Figure 2.

The Feature Selection Stage: All incoming emails are subjected to feature extraction as soon as it arrives. We use CMFS approach and Chi-square testing approach to extract the essential features. The CMFS method calculates the term frequency for term t_i [2] as,

$$\mathrm{CMFS}(t_i) = \frac{(\mathrm{tf}(t_i, c_z) + 1)^2}{(\mathrm{tf}(t_i) + |C|)(\mathrm{tf}(t, c_z) + |V|)}$$

where $tf(t_i, c_z)$ states the frequency of term t_i in class c_k, $tf(t, c_z)$ denotes the amount of frequency of all the terms in c_z, |V| represents the number of terms in feature vector space, $tf(t_i)$ represents the frequency of t_i in the training set, and |C| represents the number of classes. Chi-square testing can be stated as follows:

$$X^2(t_i) = \frac{N(a_{ki}d_{ki} - b_{ki}c_{ki})^2}{(a_{ki} + b_{ki}) + (a_{ki} + c_{ki}) + (b_{ki} + d_{ki}) + (c_{ki} + d_{ki})}$$

where N is the amount of emails, a_{ki} is the frequency of occurrence of t_i in c_k; b_{ki} is the frequency of occurrence of t_i not in class c_k; c_{ki} is the frequency of occurrence of c_k that does not have t_i; d_{ki} is the number of times neither t_i nor c_k occurs.

Normalization Stage: The terms obtained by CMFS feature selection method are now normalized using the following formula,

$$CMFS(t_i) = \frac{CMFS(t_i)}{\max_{t \in F_d} CMFS(t_i)}$$

Likewise, the terms obtained by Chi-square testing feature selection method are normalized using the following formula,

$$X^2(t_i) = \frac{X^2(t_i)}{\max_{t \in F_d} X^2(t_i)}$$

Score Assignment Stage: The terms that are normalized are graded in decreasing order based on the values obtained from $CMFS(t_i)$ and $X^2(t_i)$ respectively.

Optimal Feature Selection Stage: Our resolution is to choose the finest number of features and so we generated two pre-defined thresholds: TH1>0 and TH2<1 for choosing the optimal number of features. The ranked terms that are higher than TH1 and less than TH2 were chosen as the optimal feature; The remaining left out terms t_x that are unimportant are also chosen as, $CMFSX^2(t_x) = CMFS(t_x) \times X^2(t_x)$ and they are graded in decreasing order to support the optimal selection process. This is to provide a guarantee that these distinguishing terms can also be chosen to achieve optimality. These thresholds balances the amount of features chosen at the initial stage. These dynamic thresholds varies for different datasets.

Optimal Email Feature Selection (OEFS) Protocol

Input: N, number of features extracted from incoming mail
F_I = set of features in spam and ham database
T_d = number of terms in F_I
F_{CMFS} = number of features selected by $CMFS(t_i)$
F_{X^2} = number of features selected by $X^2(t_i)$
F_{temp} = temporary set of features
$TH1, TH2$: pre − defined threshold values; $TH1 > 0, TH2 < 1$
Output: F_O = Optimal set of features
Step 1: Initialize parameters: $F_{CMFS} = 0, F_{X^2} = 0, F_O = null, T_d = null$
Step 2: For each term t_i in F_I; $i = 0$ to $T_d - 1$
  calculate $CMFS(t_i)$ and $X^2(t_i)$ //feature selection
  Normalize $CMFS(t_i)$ and $X^2(t_i)$
 End for
Step 3: Score t_i for $CMFS(t_i)$ and $X^2(t_i)$ in descending order
 Denote the score for $CMFS(t_i)$ as $\{S_{di}\}$ and $X^2(t_i)$ as $\{S_{bi}\}$ respectively
Step 4: For each t_i in $\{S_{di}\}$
 If $\{S_{di}\} > TH1$ && $F_{CMFS} < N$
  $F_O = F_O + t_i$
  $F_{CMFS} = F_{CMFS} + 1$
 Else
  $F_{temp} = F_{temp} + t_i$
 End if
 End for
Step 5: For each t_i' in $\{S_{bi}\}$
 If $\{S_{bi}\} > TH2$ && $F_{CMFS} + F_{X^2} < N$
  $F_O = F_O + t_i'$
  $F_{X^2} = F_{X^2} + 1$
 Else
  $F_{temp} = F_{temp} + t_i'$
 End if
 End for
Step 6: If $(N - F_{CMFS} - F_{X^2}) > 0$
 For each term t_x in F_{temp}
  Calculate $CMFSX^2(t_x) = CMFS(t_x) \times X^2(t_x)$
 End for
  Score t_x for $CMFSX^2$ in descending order
  $F_O = F_O + (N - F_{CMFS} - F_{X^2})$
 End if

Figure 2 Proposed Optimal Feature Selection Algorithm

The thresholds TH1 and TH2 were chosen based on Optimal Harmony Search (OHS) Algorithm presented in [13] and it performs threshold selection in four steps. During the first step, the parameters required for threshold selection are specified is done; during the second step, the harmony space is initially set; during the third step, the modified harmony space is updated based on some values and during the fourth step, the final harmony space is created with thresholds as

TH1>0 and TH2<1. The two thresholds acquired through OHS algorithm are then used by OEFS to choose the optimal set of features which is then provided as input to any classifier to efficiently detect spam emails. We found through previous research that Fuzzy Support Vector Machine (FSVM) [13] is best suited to classify a email message as spam and ham.

IV. PERFORMANCE RESULTS AND ITS ANALYSIS

The performance evaluation of OEFS protocol was verified with Intel core – i3 processor at 3.8GHz with 2GB RAM. To validate OEFS protocol we use LingSpam dataset which contains an total of 2983 email out of which 2412 are legitimate emails and other 481 are emails which containing spam. A Java Programming used to build the selected features on vector space model using Net beans IDE 8.1 platform. Initially with the corpus datasets to perform stop word removal for removing character that are not alpha-numeric. Here the email addressed is presented in multidimensional vector space, which initially extracted the messages and its feature selection was performed using Chi-square testing and CMFS methods. Here again we define two thresholds: $TH1 > 0$ and $TH2 < 1$ which is used to accurately detect an email which contains Spam content. These methodologies of thresholds were used using Optimized Harmony Search Algorithm for its feature selection which was in [13], in which the threshold levels with low value have the high chance of less false negatives and the threshold which holds with high value always has the low chance of false positives. By this the algorithm consumes less false positives and negatives.

The performance analysis has been measured through various results of OEFS protocol and proceeds through different metrics such as Precision (P), Recall (R), Detection Time (DT) and Detection Accuracy (DA). Here we used Fuzzy Support Vector Machine (FSVM) which categories a Spam and Non-Spam by this classifier which was executed with 5-fold measures which was validated using Chi-square testing methods and CMFS for the performance of OEFS protocol. The threshold used here was balance the feature selection process executed by its $\mathrm{CMFS}(t_i)$ also with its $X^2(t_i)$. The obtained threshold values through LingSpam corpus dataset with TH1 and TH2 we obtained as 0.915 and 0.935 respectively. P is calculated as the precision:

$$P = \frac{FP}{FP + TN}$$

where FP [3] is defined as the amount of calculated ham emails that are mistakenly classify as spam (also called as false positive) and TN is defined as the total amount of ham emails that are suitably classify (which is also called as true negative). R is calculated as the recall:

$$R = \frac{FN}{FN + TP}$$

where FN [4] is defined as the amount of calculated spam emails which are misclassified as ham (also called as false negatives) and TP is the amount of spam mails that were accurately classified (also called as true positive).

The Detection Accuracy [5] depends on the P and R values and is defined as follows:

$$DA = 1 - \left(\frac{P * R}{1 + (P * R)}\right)$$

where P is the precision and R is the recall. The higher the P and R values, the lower is the DA and vice versa.

The Detection Time (DT) [6] is the calculated amount of time taken which was arrives at the server to detect spam email content. Detection Time is comprised of two factors based on the accuracy of OEFS protocol which reduces the searching time by selecting optimal set of feature and the efficiency of the FVSM classifier.

Here we analyze the P and R Parametric values which vary the optimal set of features as number of selected email as 100, 200, 300, 400 and 500 for OEFS, Chi-Square Testing and CMFS. The comparison of P and R for OEFS, Chi-square and CMFS were shown in Fig. 3 and Figure 5 respectively. Here, obtained results shows that when less number of feature the Precision and recall value increases and as the number of feature increases the P and R decreases. We observed that OEFS outperforms Chi-square testing and CMFS by producing less number of FPs and FNs when compared to chi-square and CMFS methodologies.

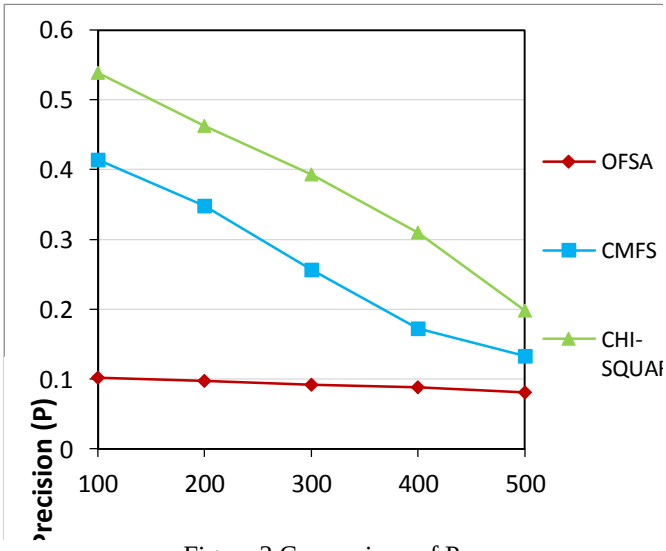

Figure 3 Comparison of P

The performance of OEFS protocol is also analyzed for ongoing emails; whenever a new email arrives, they are

stored in a new dataset. Now, by varying the number of emails (i.e., 10000, 20000, 30000, 40000 and 50000 emails) with 500 senders, we carried out the experiment to analyze the optimal number of features needed by each protocol to effectively carry out the detection. Fig. 4 shows the comparison of number of selected features for varying number of incoming emails. We observed that when the number of incoming emails increases, the number of features also increases, and hence optimal feature selection cannot be achieved efficiently. However, OEFS outperforms the other protocols by selecting optimal set of features. The P and R values are affected by both static emails collected from LingSpam Corpus and incoming emails in real-time stored in our new dataset.

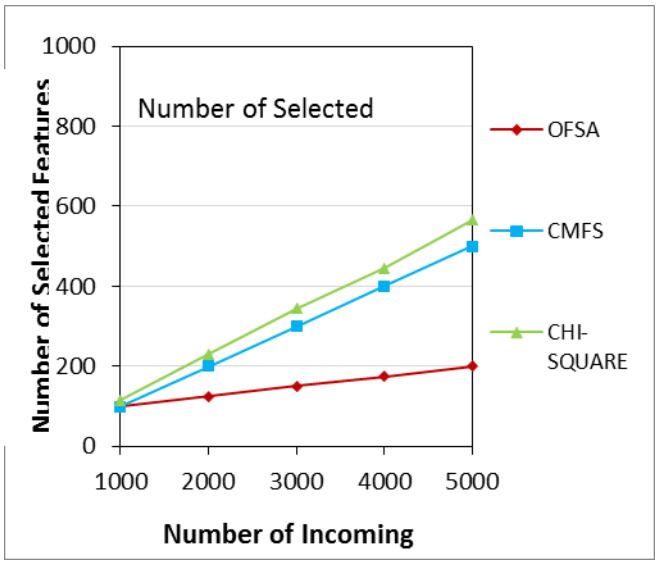

Figure 4 Number of Selected Features for different Protocols

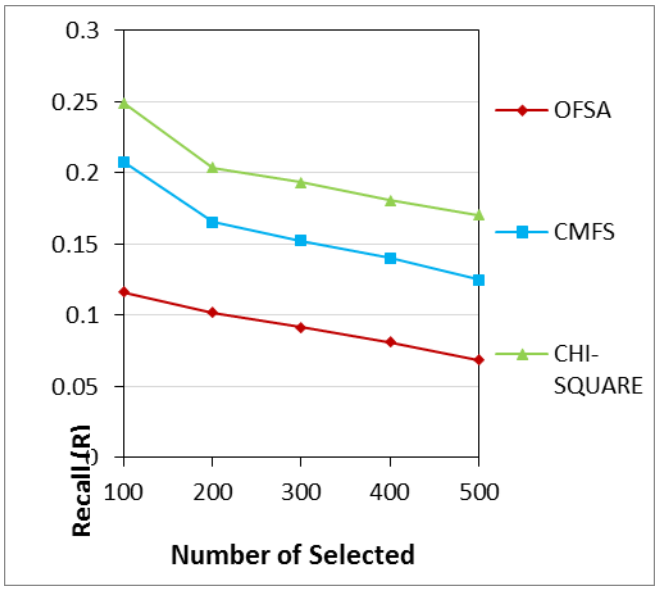

Figure 5 Comparison of Recall

We analyzed the Detection Accuracy of OEFS, CMFS and Chi-Square Testing approaches by varying the number of selected set of features for different number of emails. It is observed that the DA remains more or less the same for OFSA for different number of features. This is because of the two thresholds used for selecting optimal number of features which produced accurate P and R values to achieve good detection accuracy. Fig. 6 shows the comparison of DA for OFSA, CMFS and Chi-Square Testing. TABLE I shows the performance analysis of the three protocols in terms of Precision, Recall and Detection Accuracy.

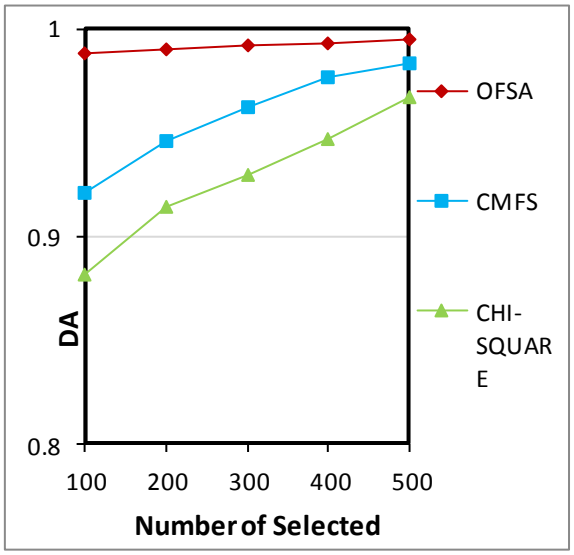

Figure 6 Comparison of Detection Accuracy

We examined the DT for different number of optimal features (i.e., 100 to 500) for the three protocols. We observed that when the optimal features is few say, 100, the DT has been is less but it raises as the number of optimal features grows (say, from 200, 300, 400 and 500). OEFS outperforms the other protocols by detecting spam with short delay. Fig. 7 shows the comparison of DT for OFSA, CMFS and Chi-Square Testing. The results show that OFSP takes 0.008 seconds to detect a spam email whereas, CMFS takes 0.018 seconds and Chi-Square Testing takes 0.02 seconds and our proposed protocol outperforms the other protocols by performing detection within a short delay. This is due to the efficiency of the proposed algorithm which selects optimal set of features.

TABLE I PERFORMANCE ANALYSIS OF DIFFERENT PROTOCOLS

Protocols	FP	TN	FN	TP	P	R	DA
OFSA	194	2218	33	448	0.080	0.069	0.99
CMFS	320	2092	60	421	0.133	0.125	0.98
CHI-SQUARE	477	1935	82	399	0.198	0.170	0.97

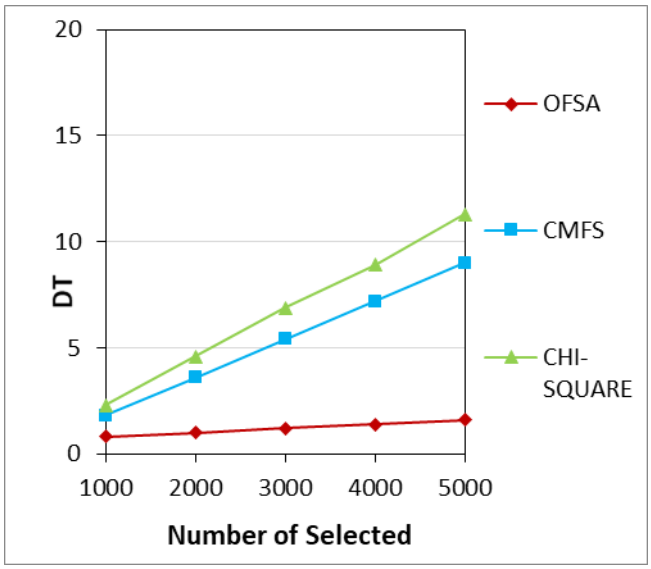

Figure 7 Comparison of Detection Time

V. CONCLUSION

Thus the proposed Optimal Email Feature Selection (OEFS) Protocol reduces the searching time of email spam detection by selecting optimal number of features. The OEFS protocol has been designed and implemented for large volumes of data and achieves high level of accuracy by generating optimal set of features. The performance analysis of OEFS protocol is done using different terms: Precision, Recall, Detection Time and Detection Accuracy. The performance results show that OEFS protocol outperforms existing detection protocols like, CRVSM and PEP by achieving better results.

REFERENCES

[1] Hiroshi O., Hiromi A., Masato K., "Feature selection with a measure of deviations from Poisson in text categorization", Expert Syst. with Appln., Pages 6826–6832., Vol. 36, Issue: 3, Apr 1, 2009

[2] Jieming Y., Yuanning L., Xiaodong Z., Zhen L., Xiaoxu Z., "A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization", Inf. Processing & Mgmt., Vol. 48, Issue 4, Pages 741-754, 2012

[3] Jingle, D.J. and Rajsingh, E.B.,"ColShield: an effective and collaborative protection shield for the detection and prevention of collaborative flooding of DDoS attacks in wireless mesh networks," Human-centric Comp. and Inf. Sci., Springer Publications, Vol. 4, Issue 8, 2014

[4] Jingle, D.J., Rajsingh,E.B. and Paul, M., "Distributed Detection of DoS Using Clock Values in Wireless Broadband Networks," Int. J. of Engg. and Advanced Tech. (IJEAT), Vol. 1, Issue 5, June 2012.

[5] Jingle, D.J and Rajsingh, E.B., "DDDOST: Distributed detection of DOS attack using timers in wireless broadband networks," Int. Conf. on Advanced Comp. (ICoAC), IEEE, ISSN : 2377-6927, DOI : 10.1109/ICoAC.2012.6416795, 2012.

[6] Jingle, D.J. and Rajsingh,E.B., "Defending IP Spoofing Attack and TCP SYN Flooding Attack in Next Generation Multi-hop Wireless Networks," Inter. J. of Inf. and Net. Sec. (IJINS), Vol. 2, Issue 2, Dec 2012.

[7] Paul, M. and Ravi R., "A Collaborative Reputation-based Vector Space Model for Email Spam Filtering", J. of Comput. and Theoret. Nanosci., Vol. 15, No.2, Pages 474-479, February 2018, doi.org/10.1166/jctn.2018.7128, American Scientific Publishers, 2018

[8] P. Mano Paul, Dr. R. Ravi, "A novel Email Spam Detection protocol for next generation networks" Taga J. of Graphic Tech., Tech. Ass. of the Graphic Arts, Vol.14, Swansea Printing Technology Ltd, Pages 124-133, 2018

[9] P. Mano Paul and R. Ravi, "Cooperative Vector Based Reactive System For Protecting Email Against Spammers In Wireless Networks", J. of Elec. Engg., Vol.18, Edition:4, ISSN 1582-4594, Dec. 2018.

[10] Tu Ouyang, Soumya Ray, Mark Allman, Michael Rabinovich, "A large-scale empirical analysis of email spam detection through network characteristics in a stand-alone enterprise", Comp. Net., Vol. 59, 11 Feb. 2014, Pages 101-121, Elsev. B.V, http://dx.doi.org/10.1016/j.comnet.2013.08.031, 2013

[11] Vitor Basto-Fernandes, Iryna Yevseyeva, José R. Méndez, Jiaqi Zhaod, Florentino Fdez-Riverola, Michael T.M. Emmerich, "A spam filtering multi-objective optimization study covering parsimony maximization and three-way classification", Appl. Soft Comp., Vol. 48, Pages 111-123, 2016

[12] Wazir Zada Khan, Muhammad Khurram Khan, Fahad Bin Muhaya, Muhammad Y Aalsalem and Han-Chieh Chao, "A Comprehensive Study of Email Spam Botnet Detection", IEEE Comm. Surv. & Tut..

[13] Youwei Wang, Yuanning Liu, Lizhou Feng, Xiaodong Zhu, "Novel feature selection method based on harmony search for email Classification", Knowl. - Based Syst., Vol. 73, Pages 311 - 323, http://dx.doi.org/10.1016/j.knosys.2014.10.013, Elsevier B.V., January 2015

Authors Profile

Dr. P. Mano Paul is the Assistant Professor of Presidency University, Bangalore, India. He received his Bachelor of Engineering degree in Information Technology from Noorul Islam College of Engineering, Manonmaniam Sundaranar university, India in 2002. He received his Master of Engineering degree from Department of Computer Science and Engineering from Manonmanian Sundaranar University, India in 2005. He completed his Ph.D in the department of Computer Science and Engineering from Anna University, Guindy campus, chennai, India. His research interests is in the area of Cyber Security, Mobile Ad-hoc Networks and Network Security.

Dr. I. Diana Jeba Jingle is the Assistant Professor of Christ University, Bangalore, India. He received her Bachelor of Technology in Information Technology from Sun college of Engineering and Technology, Anna University, India in 2006 and she received her Master of Engineering degree in Computer Science Engineering from Francis Xavier College of Engineering, Anna University, India in 2008. She obtained her Ph.D in Computer Science and Engineering from karunya University, India. Her area of research interests is in wireless networks, Cyber security, Mobile Ad-hoc Networks, Internet of Things, Internet of Everythings etc.

