Effect of Header-based Features on Accuracy of Classifiers for Spam Email Classification

Dr. Priti Kulkarni[1], Prof & Dr. Jatinderkumar R. Saini[2]*
Symbiosis Institute of Computer Studies and Research
Symbiosis International (Deemed) University, Pune, India

Prof & Dr. Haridas Acharya[3]
Allana Institute of Management Sciences
Pune, India

***Abstract*—Emails are an integral part of communication in today's world. But Spam emails are a hindrance, leading to reduction in efficiency, security threats and wastage of bandwidth. Hence, they need to be filtered at the first filtering station, so that employees are spared the drudgery of handling them. Most of the earlier approaches are mainly focused on building content-based filters using body of an email message. Use of selected header features to filter spam, is a better strategy, which was initiated by few researchers. In this context, our research intends to find out minimum number of features required to classify spam and ham emails. A set of experiments was conducted with three datasets and five Feature Selection techniques namely Chi-square, Correlation, Relief Feature Selection, Information Gain, and Wrapper. Five-classification algorithms-Naïve Bayes, Decision Tree, NBTree, Random Forest and Support Vector Machine were used. In most of the approaches, a trade-off exists between improper filtering and number of features. Hence arriving at an optimum set of features is a challenge. Our results show that in order to achieve the objective of satisfactory filtering, minimum 5 and maximum 14 features are required.**

***Keywords*—Email classification; Chi-Square; correlation; relief feature selection; wrapper; information gain; Naive Bayes; J48; spam; support vector machine; random forest; NBTree**

I. Introduction

Email communication has become an essential part of all spheres of personal life as well as professional life. But all the emails are not relevant for every user. Day by day the email traffic is increasing, making it imperative to filter spam emails. According to a survey conducted by Radicati 2017, total emails sent and received per day would reach to 319.6 billion by the end of year 2021 [1]. As per Infocomm survey 2016 for internet usage, 'Sending and receiving emails' (94%) and 'Information Search' (92%) are two main activities on internet [2].

Spam finds the first mention as early as in 1975in RFC 706 by John Postel. According to RFC 2505, mass unsolicited emails, sent in large volumes to target the consumers, are called spam emails. Text Retrieval Conference (TREC) defines spam as "unsolicited, unwanted email sent indiscriminately, directly or indirectly, by a sender having no current relationship with the user"(spam track) [10],[11]. According to survey conducted by GFI software in 2014, spam emails consume bandwidth and detract the user from the work. The purpose of sending spam differs from a person-to-person and organization-to-organization. It is used to send phishing, advertising emails or to spread viruses and worm.

An email contains headers and body. Email header field format is defined in RFC 822, RFC 2822. One may classify an email inspecting the body content and headers. Email header contains useful information (Metadata). Contents of the body can be a text, pictorial data, or even sound. This is a purely unstructured part of the email. Our work intends to find out:

a) The minimum number of features in header, necessary to identify spam email.

b) The effect of identified features on the accuracy of classification of email.

c) The best combination of features selection technique and classification algorithm.

This paper consists of three sections, in the first section different approaches for spam classification, as found in literature, are discussed. The next section presents the details about data collection and experiments carried out in for this research. The discussion on results of the experiments follows, along with the conclusion.

II. Literature Review

There are four commonly used techniques for spam classification namely,

a) Use of blacklist [14]

b) Protocol-based approach

c) Use of keywords or content filtering

d) Header based [20],[28],[21],[5],[36],[13]

In the first case, a list of email the network administrator maintains addresses or domain name databases. The classifier matches new record with blacklisted database and simply rejects some mails and puts them onto the spam folder. However, this blacklist requires continuous updation of list. The blacklist approach may fail if the sender's address is fake [37]. The Second approach is protocols based where traffic coming from specific IP address can be blocked. But IP addresses can be easily forged [17], [6]. In the third method of keyword or content filtering [16], spammers bypass the filter by embedding text into images. Such models provides better filtering, however it come with two disadvantages,

a) It is time consuming.

b) The process is language dependent [9].

*Corresponding Author.

That is why this paper is focused on the fourth approach of header based filtering. Spam classification helps us to filter the unwanted emails from the email Inbox. There have been various attempts to classify the spam email based on using email header [20],[21],[5],[36],[37],[38],[13],[4], using email body [3],[41],[35],[29],[27],[30],[7],[31],[32],[33],[34] and also using both body and header [18],[23],[21],[15],[42] and statistical features [19],[25]. The email header classification is performed using techniques such as Naïve Bayes (NB), Decision Tree (DT) [40][43], and Support Vector Machine (SVM) [23],[24],[20],[13],[26] Random Forest (RF) [4],[13]. When these techniques were adopted by the researchers using various features and datasets, Random Forest showed better performance than the other techniques. Selecting appropriate set of features is important because that influence accuracy of classifier [8]. Author in [5] used total 26 features derived from behaviour from headers and syslog of emails with back-propagation neural networks (BPNN) and achieved accuracy of 99.6%. But one of the drawbacks of using BPNN is its unstable time to Converge. The number of features and training data affects the performance of BPNN. So the results can fluctuate. In [13] authors have used IP address and subject with other four features which resulted into accuracy of 96.7%. But IP address may get forged. So we have not considered IP address in this research. Our attempt is to suggest optimized features without use of any text data from subject and Body of email. Therefore, we have use combination of different features from literature and by study of personal spam data.

III. Research Methodology

The experiments are conducted in two phases; first Feature Selection techniques are applied on datasets which generate subset of features. In second phase, the resultant feature subsets are used for classification to find the effect on accuracy of classifier. The minimum number of features with classifier is selected as result.

The steps are as follows,

1) Input: Email datasets.
2) Extract Email header features.
3) Apply feature selection techniques.
4) Select subset of features generated by feature selection techniques.
5) Apply classification on Email datasets with selected feature subsets.
6) Classify email into spam and non-spam.
7) Note down the accuracy of the classifiers.

IV. Data Collection and Pre-Processing

We collected emails as reference database to carry the necessary experiments. These emails were collected from personal email account during a period of last 7 years. The two Benchmark corpora available publicly, namely Spam Assassin Corpus and CSDMC2010 corpus are also used in this experiments. These datasets contains spam and ham files. Description of data collected for experimental purpose is given in Table I.

TABLE. I. Description of Email Databases used in Experimentation

Sr. No	Data Set	Description	Spam	Ham	Total
1	S1	Personal Emails	1845	4687	6532
2	B1- Spam Assassin	Benchmark Databases	500	250	750
3	B2-CSDMC 2010		1378	2949	4327

A. Use of Features in Spam Classification

RFC 822 and RFC 2822 are the standard formats, which define email structure and various email header fields. Therefore, the email header field as features are adopted from the above two. The list of features was obtained by study of personal database and from literature. Some of the earlier researchers have not addressed the following six header fields:

- Content-Transfer-Encoding,
- Authentication-Results,
- Presence of ?,!symbols in 'from',
- Presence of ? symbols' in Reply-To' and
- Presence of ? and = symbols in 'Subject'
- presence of $symbol in message-id

So in this experiment an attempt is to rectify the situation, by considering above features.

The features are grouped into two categories,

1) Base features: The features, which are used directly from definitions given in RFCs; as specified in the following list.

Let, $S(U\_B_f)$= { To, Bcc, CC, Received, Return-path, From,Subject,Received-SPF,Authentication-Results,Message-ID, Reply-To, X-Mailer, Content-Transfer-Encoding }.

In further discussions the term $S(U\_B_f)$ (set of universal base features) is used to refer to above ten features, for the ease of explanation.

2) Derived features: The features, which are constructed from, base features

$S(D_f)$ = {BCC_notempty_To_empty, Message-ID_domainname, Received-Count, Reply-To_domain, Return_Path_Domain, Span_time, Total_Recp,}

In further discussions the term $S(D_f)$ (set of derive features) is used to refer to above seven features ,for the ease of explanation,

Set of features used for experiment by combining base features and derived features is:

S(f)=S(U_Bf) US(Df)

{Authentication-Results, BCC_notempty_To_empty, Content-Transfer-Encoding, From, Message-ID, Message-ID_domainname, Received-Count, Received-SPF, Reply-To_domain, Reply-To, Return_Path_Domain, Return_Path, Span_time, Subject_symbol, To, Total_Recp, X-Mailer}

Table II shows the list of features along with its descriptions used in this study.

TABLE. II. THE LIST OF NUMBER OF FEATURES, ALONG WITH THEIR DESCRIPTION, SELECTED FROM LITERATURE AND BASED ON THE STUDY OF OUR DATASET

Base features involved	**Values extracted**	**Derived feature label**	**Description**	**Reference**
To	To is Empty		Check value of "To" header field exists or if it contains "Undisclosed Recipients" or "<>" symbol	[37], [44]
	To is Undisclosed			[36],[9]
	To contains <>			Proposed feature
BCC,TO	To is empty and BCC is not Empty	BCC_not empty_ To_empty	Check if "BCC" contains email address and "To" do not have any email address.	[36]
To	To_number of address	Total_Recp (To+CC+ BCC)	Total number of email addresses in "To" field	[13] [44]
CC	CC_number of address		Total number of email addresses in "CC" field	[45]
BCC	BCC_number of address		Total number of email address in "To" field	[4]
Received	Count number of received fields	Received_ count	Contains total number of "Received" fields	[4]
Received	Time difference between first received field and last received field, extracted time converted into UTC	Span time	Total travelling time of email from source machine to destination machine.	[4]
From	From contains ?		Check for presence of ?,!,<> symbols in from header field.	proposed feature
From	From contains!			
From	From contains <>			
Subject	Subject contains ?	Subject_ symbol	Check subject field contains symbol "?,="	proposed feature
Subject	Subject contains =			
Base features	**Values extracted**	**Derived feature label**	**Description**	**Reference**
Received-SPF	Received-SPF="bad"		Check Received-SPF field for values as bad, softfail, fail, bad,	[12]
Received-SPF	Received-SPF="softfail"			
Received-SPF	Received-SPF="fail"			
Authentication- result	dkim="bad"		Check Authentication field, dkim value which allows to check email came from authentic domain	proposed feature
Authentication- result	dkim="softfail"			
Authentication- result	dkim="fail"			
Message-id, From	domain name	Message-ID_From_domainname	Check domain in "From" and "Message-id" are not same	[4], [44]
Message-id	Dollar symbol present		Check if message id contains any $ symbol	proposed feature
Reply-To	Reply-To is Empty/exists		Check "Reply -To" is exists or contains"?"	[44]
Reply-To	Reply-To is "?"			proposed feature
Reply-To	Reply-To _domain	Reply-To _domain	Check domain in "From" and "Reply-To" are not same	[44]
X-Mailer	X-Mailer_exist		Check whether X-Mailer exists & checks for valid value of X-Mailer	[4], [13]
Content- Transfer-Encoding	Content-Transfer-Encoding is exists		Check if content transfer Encoding exists/contains no value.	proposed feature
Return-Path	return-path=" " or return-path="bounce"		Check values in "Return-Path" if exists, check if it contains "bounce" word	[44]
Return-Path, From	Return path is NOT matching with From address	Return path_From Domain	Check domain in "return_path" and "From" are same	[45]

V. EXPERIMENT

As mentioned earlier, experiments were conducted on three dataset emails as described in Table I. A code is developed in python to extract email header data according to Table II. Our proposed model evaluates email using these 17 features. Each feature is assigned score of 1 (one) if condition is satisfied otherwise it is marked as 0 (zero). The sum of scores was calculated in the end. In this experiment, chi-squared [19], correlation based Feature Selection[39], Information Gain, and relief [22] and Wrapper Feature Selection techniques are applied to find significant features of an email. Classifiers namely Naïve Bayes, Decision Tree, Random Forest, NBTree, Support Vector Machine were used in the experiment.

The data mining tool Weka has been used for applying the machine learning techniques. All the Feature Selection methods and classifiers were adopted in Weka as a selectable runtime parameter. Collected data were arranged in a CSV file in the following format: feature 1, feature 2, feature 3, feature n, class label (Class label indicating two classes, Spam and Ham.) 10 fold cross validation technique is used for data validation. This method uses 90% of the data for training and 10% for testing.

The average weight of each feature generated by all Feature Selection techniques is calculated and listed in Fig. 1.

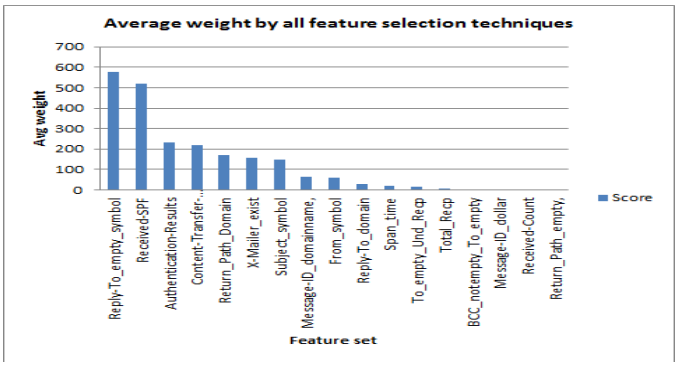

Fig. 1. Average Weight by all Feature Selection Techniques.

It can be clearly observed that our proposed features namely content-transfer-encoding, and Authentication-result, belong to the first five features by weight and have significant contribution to spam classification. The next two features Subject_symbol and From_symbol are among the top ten features. However, our proposed features namely BCC_notempty_To_empty and Message-ID_dollar do not have any significant contribution in spam classification.

VI. RESULTS AND DISCUSSION

In this experiment, we have not considered any text feature value from either body or subject. Following are the conventions used in Table III, Table IV and Table V.

FSM1-Chi Squared Feature Selection; FSM2-Correlation based Feature Selection, FSM3- Information Gain, FSM4-Relief Feature Selection, FSM5- Wrapper Feature Selection.

Classifiers:

NB-Naïve Bayes, DT-Decision Tree, RF -Random Forest, NBTree-Naïve Bayes Tree, SVM-Support Vector Machine

As Table III indicates, for dataset S1, the results showed accuracy of 93.53**%** with 17 header features. The maximum features are generated by Relief technique (RT), i.e. 14 features. It maintains best balance between false positive rate and true positive rate. Accuracy of RF is improved by 0.03% with 14 features. With accuracy of 93.56%, Random Forest (RF) outperformed the other four classifiers--Naïve Bayes (NB), Decision Tree (DT), NBtree and Support Vector Machine (SVM).Further, Next to RF, DT classifier also performs well. Naïve Bayes shows stable performance when features are increased from 11 to 14. As number of features reduced, performance of DT and RF decreased. When number of features varied between 11 and 14, Support Vector Machine performed well. However when features are reduced from 11 features to five features, performance of Support Vector Machine decreased by 0.9%.

TABLE. III. PERFORMANCE OF FEATURE SELECTION TECHNIQUES ON THE ACCURACY OF CLASSIFIERS ON DATASET S1

FSM	No of features selected	NB	DT	RF	NBTree	SVM
**	17	89.72	93.24	**93.53**	91.68	90.53
FSM1	11	89.72	92.39	92.71	91.49	90.54
FSM2	5	89.86	90.45	**90.65**	90.65	90.45
FSM3	11	89.72	92.39	92.71	91.49	90.54
FSM4	14	89.72	93.25	**93.56**	91.47	90.54
FSM5	13	89.72	93.25	93.53	91.78	90.54

TABLE. IV. PERFORMANCE OF FEATURE SELECTION TECHNIQUES ON ACCURACY OF CLASSIFIER ON DATABASE B1

FSM	No of features selected	NB	DT	RF	NBTree	SVM
	17	79.33	85.2	85.73	80.8	79.46
FSM1	6	79.46	81.86	81.33	80.93	76.13
FSM2	5	80.66	81.06	**81.33**	80.66	76.13
FSM3	6	79.46	81.86	81.33	80.93	76.13
FSM4	12	80.66	83.6	**83.73**	80.93	77.2
FSM5	7	80.26	81.06	80.66	80.66	77.06

TABLE. V. PERFORMANCE OF FEATURE SELECTION TECHNIQUES ON ACCURACY OF CLASSIFIER ON DATABASE B2

FSM	No of features selected	NB	DT	RF	NBTree	SVM
**	17	72.48	93.28	94.71	93.42	85.73
FSM1	14	79.57	90.67	91.06	91.01	85.41
FSM2	5	70.58	85.66	**86.17**	86.02	84.69
FSM3	14	79.57	90.67	91.06	91.01	85.41
FSM4	13	72.48	93.28	**94.78**	93.81	85.73
FSM5	8	71.53	93.44	71.53	93.51	84.71

Moreover, Correlation based Feature Selection technique generated five features, which are minimum number of features. When features are reduced from 17 to 5, accuracy of Random Forest (RF) is reduced by 2.36%. In short, RF and NBTree classifiers give high accuracy of 90.65% as compared to the other three. With five features, accuracy of NB is the lowest among all. However, even otherwise, NB did not perform so well as other four classifiers even with more number of features.

On benchmark dataset B1 of the size of 750 data records, Random Forest performs best (83.73% accuracy) with maximum of 12 features. Correlation based Feature Selection method generated 5 features, the minimum number in this dataset resulting into accuracy of 81.33% with RF classifier. On benchmark dataset B2, RF shows better performance, giving accuracy of 94.78% as compared to other two datasets. In this dataset also, relief Feature Selection technique with RF classifier outperformed others even with 13 features. In the same way, Random Forest performed better with accuracy of 91.6% when Chi Square and IG generated 14 maximum features. Correlation based FS generates 5 features. With minimum number of 5 features accuracy reduce by 4.89%. One of the common observations is that Random Forest method with Relief as Feature Selection technique performs better on all the datasets.

Following are the set of minimum and maximum number of features:

S_min_5={Total_Recp, Subject, Received-SPF, Authentication-Results, Reply-To}

S_max_14={Authentication Results, BCC_notempty_To_empty, Content-Transfer-Encoding, From_symbol, Message-ID_domainname, Received SPF, Reply-To_domain, Reply-To_empty_symbol, Return path_FromDomain, Span_time, Subject_symbol, X-Mailer_empty, To_empty_Und_Recp, Total_Recp}

VII. Conclusion

In this paper, we evaluated performance of five Feature Selection techniques and five classifiers on email headers. Our header based approach for Feature Selection showed that minimum five features generated by correlation based Feature Selection technique performed well on all three datasets with varying accuracy 70.58% to 90.65%. Relief Feature Selection technique generated the maximum fourteen features with varying accuracy of 91.06% to 94.78%.This implies that the features we proposed namely, Authentication-result and content-transfer encoding play significant role in identifying spam emails. The result of our experiment result shows that Random Forest performs better than all other classifiers in terms of accuracy as well as number of features.

References

[1] The Radicati Group, Inc. https://www.radicati.com/wp/wp-content/uploads/2017/01/ Email-Statistics-Report-2017-2021-Executive-Summary.pdf

[2] Annual Survey On Infocomm Media Manpower For 2016(https://www.imda.gov.sg/-/media/imda/files/industry-development/fact-and-figures/infocomm-survey-reports/infocomm-media-manpower-survey-2016_public-report.pdf?la=en)

[3] Ayodele T., Zhou S., Khusainov R.: Email Classification Using Back Propagation Technique, International Journal of Intelligent Computing Research, 2010.

[4] Al-Jarrah, O., Khater, I., & Al-Duwairi, B. (2012). Identifying potentially useful email header features for email spam filtering. In The Sixth International Conference on Digital Society (ICDS) (Vol. 30, p. 140).

[5] C.-H. Wu, "Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks," Expert Systems with Applications, vol. 36, pp. 4321-4330, April, 2009

[6] Chirita, P. A., Diederich, J., & Nejdl, W. (2005, October). MailRank: using ranking for spam detection. In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 373-380). ACM.

[7] Daeef, A. Y., Ahmad, R. B., Yacob, Y., Yaakob, N., & Azir, K. N. F. K. (2016). Multi Stage Phishing Email Classification. Journal of Theoretical & Applied Information Technology, 83(2).

[8] De Stefano C, Fontanella F, Marrocco C, et al. A GA-based Feature Selection approach with an application to handwritten character recognition. Pattern Recognition Letters 2013

[9] F. Salcedo-Campos, J. Dнaz-Verdejo, P. GarcнaTeodoro, Segmental parameterization and statistical modelling of e-mail headers for spam detection. Information Sciences, no. 195(2012), 2012, pp. 45-61.

[10] Gordon Cormack and Thomas Lynam.(2005). Spam corpus creation for TREC. In Proceedings of Second Conference on Email and Anti-Spam, CEAS'2005,

[11] Gordon Cormack, TREC 2006 Spam Track Overview (http://trec.nist.gov/pubs/trec15/papers/SPAM06.OVERVIEW.pdf)

[12] Gorling, S. (2007). An overview of the Sender Policy Framework (SPF) as an anti-phishing mechanism. Internet Research, 17(2), 169-179.

[13] Hu, Y., Guo, C., Ngai, E. W. T., Liu, M., & Chen, S. (2010). A scalable intelligent non-content-based spam-filtering framework. Expert systems with applications, 37(12), 8557-8565.

[14] Jung J, Sit E. An empirical study of spam traffic and the use of DNS black lists. In: Proceedings of fourth ACM SIGCOMM conference on internet measurement, Taormina, Sicily, Italy; October 2004.

[15] Jason Chan, Irena Koprinska, Josiah Poon, Co-training with a Single Natural Feature Set Applied to Email Classification", In:Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence 2004

[16] Kelly Jackson Higgins, Dark Reading. Botnets Battle Over Turf. http://www.darkreading.com/document.asp? doc id=122116, Apr. 2007

[17] Ramachandran, A., Feamster, N., & Vempala, S. (2007, October). Filtering spam with behavioral blacklisting. In Proceedings of the 14th ACM conference on Computer and communications security (pp. 342-351). ACM.

[18] Kiritchenko S., Matwin S., Abu-Hakima S. (2004) Email Classification with Temporal Features. In: Klopotek M.A., Wierzchon S.T., Trojanowski K. (eds) Intelligent Information Processing and Web Mining. Advances in Soft Computing, vol 25. Springer, Berlin, Heidelberg

[19] Gomez, J. C., Boiy, E., & Moens, M. F. (2012). Highly discriminative statistical features for email classification. Knowledge and information systems, 31(1), 23-53.

[20] M. Ye, et al., "Spam Discrimination Based on Mail Header Feature and SVM," In Proc. Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference on Dalian Oct. 2008

[21] Jyh-Jian Sheu, Ko-Tsung Chu, Nien-Feng Li, Cheng-Chi Lee, 2017, An efficient incremental learning mechanism for tracking concept drift in spam filtering, PLoS one

[22] Kira, Kenji and Rendell, Larry (1992). The Feature Selection Problem: Traditional Methods and a New Algorithm. AAAI-92 Proceedings.

[23] Lai, C. (2007). An empirical study of three machine learning methods for spam filtering. Knowledge-Based Systems, 20(3), 249–254.

[24] Wang, M. F., Jheng, S. L., Tsai, M. F., & Tang, C. H. (2011, July). Enterprise email classification based on social network features. In 2011

International Conference on Advances in Social Networks Analysis and Mining (pp. 532-536). IEEE.

[25] Martin, S., Nelson, B., Sewani, A., Chen, K., & Joseph, A. D. (2005, July). Analyzing Behavioral Features for Email Classification. In CEAS

[26] Patidar, V., Singh, D., & Singh, A. (2013). A Novel Technique of Email Classification for Spam Detection. International Journal of Applied Information Systems, 5(10).

[27] Ruan, G. and Tan, Y. (2010). A Three-Layer BackPropagation Neural Network for Spam Detection using Artificial Immune Concentration. Soft Computing, 14(2), pp. 139-150.

[28] Sheu J.J, "An Efficient Two-phase Spam Filtering Method Based on E-mails Categorization " International Journal of Network Security, vol.9, pp. 34-43, July 2009.

[29] Senthamarai Kannan Subramanian and N. Ramaraj, 2007. Automated Classification of Customer Emails via Association Rule Mining. Information Technology Journal, 6: 567-572.

[30] Schmida M. R., Iqbalb F., Fungc B. (2015) E-mail authorship attribution using customized associative classification. The Proceedings of the Fifteenth Annual DFRWS Conference, Volume 14, Supplement 1, August 2015, Pages S116–S126.

[31] Sahın, E., Aydos, M., & Orhan, F. (2018, May). Spam/ham e-mail classification using machine learning methods based on bag of words technique. In 2018 26th Signal Processing and Communications Applications Conference (SIU). IEEE.

[32] Saini, J. R., & Desai, A. A. (2010). "Analysis of Classifications of Unsolicited Bulk Emails". International Journal of Computer and Information Engineering, 4(1), 115-119.

[33] Saini J. R., Desai A. A.(2010), "A Survey of Classifications of Unsolicited Bulk Emails", in National Journal of Computer Science and Technology (NJCST), 2010. ISSN 0975-2463

[34] Saini J. R.,Desai A.A,(2010), "A Supervised Machine Learning Approach with Re-training for Un-structured Document Classification in UBE", published in INFOCOMP Journal of Computer Science; ISSN: 1807-4545.

[35] Trevino, A. 2007. Spam Filtering Through Header Relay Detection.

[36] Wang C-C. Sender and receiver addresses as cues for anti-spam filtering. Journal of Research and Practice in Information Technology 2004;36(1):3–7.

[37] Wang, C. C., & Chen, S. Y. (2007). Using header session messages to anti-spamming. Computers & Security, 26(5), 381-390.

[38] W. Li, W. Meng, Z. Tan, and Y. Xiang,(2014) "Towards designing an email classification system using multi-view based semi-supervised learning," in 13th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2014, 2015, pp. 174-181.

[39] Yang, Q., & Gras, R, 2010, December. How dependencies affect the capability of several Feature Selection approaches to extract the key features. In Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on(pp. 127-134). IEEE.

[40] Youn, S., & McLeod, D. (2007). A comparative study for email classification. In Advances and innovations in systems, computing sciences and software engineering (pp. 387-391). Springer, Dordrecht.

[41] Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998, July). A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization: Papers from the 1998 workshop (Vol. 62, pp. 98-105).

[42] Zhang, L., Zhu, J., & Yao, T. (2004). An evaluation of statistical spam filtering techniques. ACM Transactions on Asian Language Information Processing (TALIP), 3(4), 243-269.

[43] Ying, K. C., Lin, S. W., Lee, Z. J., & Lin, Y. T. (2010). An ensemble approach applied to classify spam e-mails. Expert Systems with Applications, 37(3), 2197-2201.

[44] Qaroush, A., Khater, I. M., & Washaha, M. (2012). Identifying spam e-mail based-on statistical header features and sender behavior. In Proceedings of the CUBE International Information Technology Conference (pp. 771-778). ACM.

[45] Zhang, D. Y., & Yang, L. (2014). Implementation of Mail Classification Using Neural Networks of the Second Type Spline Weight Functions. In Applied Mechanics and Materials (Vol. 513, pp. 687-690). Trans Tech Publications Ltd.