The Fusion of Features to Improve Spam E-mail Classification

Pramod Prakash Ghogare[1], Manoj P. Patil[2]
[1]*Dept. of Computer Application, KCES's Institute of Management and Research, Jalgaon, India*
[2]*School of Computer Sciences, KBC North Maharashtra University, Jalgaon, India*
Email: [1]pramod.ghogare@yahoo.com, [2]mpp145@gmail.com

***Abstract*—All over the world, the Internet and E-mail are dominant communication tools. The internet provides a different way of communication but, on the other hand, increases the exploitation of e-mail. The growth of unwanted e-mails has encouraged the development of numerous spam e-mail filtering techniques. The spammers are devising new methods each time, anti-spamming techniques fail to filter out spam e-mails. Spam e-mail is difficult for the sustainability of the internet and worldwide business. This paper describes an experimental analysis of spam e-mail classification using multiple feature synthesis. The Subject, E-mail address, and IP address of e-mail were selected to classify an e-mail. The experimental result signifies the performance of the algorithm on the standard SpamAssassin dataset.**

Keywords: E-mail, Spam, Spam Classification, Naïve Bayes.

I. Introduction

E-mail is a long-standing tool for communication over the internet. The requirement of e-mail addresses is increasing over different mobile apps and websites; that makes e-mail addresses a resource to reach users. The spammer sends advertisements, product detail, discount offer to these clients through e-mail. Different spamming methods are used to bypass e-mail filters. So it is essential to build a proper filtration technique for filtering these unnecessary e-mails.

Spam e-mails are very frustrating things, which divert the user. The valuable resources used by spam e-mails for the fallow task are the primary concern to filter e-mail. A spam e-mail is one that the user not intended to receive. "A spam is anonymous, unsolicited bulk e-mail." "Unsolicited, unwanted e-mail that was sent indiscriminately, directly or indirectly, by a sender having no current relationship with the user" [1]. A spam e-mail is sent in bulk without the user's permission. These spam e-mails can be advertisements, product images, text in images, offers, discounts offer, asking for registration or donation, asking for debit card or bank details. "Real spam is that e-mail for advertising of product sent to list groups" [2]. Spam irritates a user degrades work efficiency by troubling users constantly [3]. Sometimes real e-mail is removed due to frustration by spam e-mails. An e-mail server takes much time for processing and removing spam e-mails [4]. In countries like Russia, it is legal to use spamming for the increasing turnover of business using spam e-mails. It is expected that 58 billion junk e-mails will be sent every day in the upcoming years: the e-mail service provider, internet service provider, network administrator affected by spam e-mails. The amount of money, time, and resources engaged by spam e-mails are the main reason to stop spamming.

In the early days, spam was directly sent to users and was called direct spam. Modem Pool is the type of spam in which dial-up connections were used to generate spam e-mails. Trojan horses are used for downloading malware and crippling viruses spread onto several machines through spam, which allows them to control the system from a remote location [5]. Word obfuscation is the technique of spamming which changes the position of letters in a word that makes it different than it is. For example, the word 'Available' is easily traceable but 'av@il@ble" which is somewhat hard to track for spam filters whereas it is readable for a human [6]. The spams are various types, such as health spam. These contain the benefits of purchasing and using health products [7]. Financial spams request for bank account details [8]. An e-mail requesting to buy shares, Market is down purchase share, these types of e-mail are in the stock category [7]. An advertisement of a political party asking for a vote or donation is a type of political spam [9]. Adult spam contains adult or pornographic content [7]. Phishing spam appearances look like official e-mails and may be used for committing fraud transactions [9].

II. Spam Classification

It is a method to classify spam e-mail based on e-mail properties. The spam classifying methods are broadly classified into origin-based and content-based.

A. Origin-Based Spam Classification

Origin-based filters use network information to classify spam e-mail. The e-mail source details are

ISBN: 978-93-91355-11-1

compared with the history of the spammers; if the incoming mail is matched, it is categorized as spam [10]. An e-mail header has information like Sender e-mail address, Sender IP address, Name of Sender, Reply-To, Subject, Content-Type, Message-ID, Delivered-To, Date, From, To, Received [11, 12], this information can be used to fetch sender and metadata of e-mail to spam e-mail classification [13]. Hassan et al. applied the Naïve-Bayes method using specific header field-based attributes, the overall improvement in spam filtering was achieved [14]. Sao et al. (2015) found that naïve Bayesian classifiers have accurate than a support vector machine. The error rate was very low for Naïve Bayesian Classifier [15].

The subject of an e-mail can play a vital role in classification; the user can see the subject in the mailbox before opening the e-mail. Spammer keeps the e-mail subject somewhat attractive that can lead the user to open the mail.

B. Naïve Bayes Classification

The Naïve Bayes (NB) classifier was proposed in 1998, the probability of an upcoming event can be recognized from the previous events [16]. The possibility of spam is more if some words frequently occur in spam e-mails but not in legitimate e-mails. Each word has a definite chance of occurring in the spam or legitimate e-mail. The filter will mark the e-mail as spam if the probability of words exceeds a specific limit. Statistic-based spam filters use Bayesian probability calculation to combine individual token statistics to decide [17].

$$S[T] = \frac{C_{spam(T)}}{C_{spam(T)} + C_{Legitimate(T)}} \quad (1)$$

Where C Spam(T) and C Legitimate(T), are the numbers of spam or Legitimate messages containing token T, respectively. To calculate the possibility of a message M with tokens {T1, T2, T3, ..., TN}, combining each token's spam probability to evaluate the overall e-mail's spam probability. The product of each token is calculated for spam probability and compare with the product of the token's Legitimate probability. The e-mail is marked as spam if the overall spam probability S[M] is greater than the Legitimate probability H[M] [18].

C. Literature Review

Origin-based filters are based on network information to classify an e-mail. The sender address is compared with the history of the spam sender; if an incoming e-mail is from one of the spammer addresses, it is categorized as spam e-mail [10]. The header of an e-mail has various information related to the sender, source of an e-mail, and this can be used for classification.

Androutsopoulos et al. found that the Naive Bayes (NB) classifier has high spam recall and precision. Automatic anti-spam filtering has become an essential member of junk-filtering tools for the internet [19]. The Naïve Bayes classifier has high accuracy and speed with simplicity. Naïve Bayes classifier uses the capabilities of tokens and related probabilities, allowing classification decision and experimental performance [20, 21]. Rusland et al. found that NB can give the optimum precision in spam e-mail classification [22]. Chih-Chin concluded, the Naïve Bayes and Support Vector Machine yield better performance than k-NN [23]. Youn et al. found that Naïve Bayesian classifier gave better results than Neural Network and Support Vector Machine [24]. The Naive Bayes is more accurate on the client-side [25]. Awad et al. studied machine learning methods in terms of accuracy Naïve Bayes, and rough sets methods have a very satisfying performance among the other methods. Among the four machine learning methods, the KNN algorithm has the worst precision percentage. The disadvantage of text filtering is that they are time-consuming [26, 27]. Sao et al. created an e-mail spam classification system on the Lingspam dataset. The Naive Bayes classifier and Support Vector Machine classifier used for testing the dataset. The Naive Bayes classifier produces a better result than the Support Vector Machine when the number of the dataset is increased, and it was seen that the Naive Bayes classifier has a low error rate than the Support Vector Machine [28]. Chopra et al. found that the Bayesian algorithm is used to add an e-mail to the spam list, and only the legitimate e-mails are shown to the user [29]. Sharaff et al. found that the SVM and Naïve Bayes give overall best classification results without employing any feature selection techniques. The Naïve Bayes and J48 classification algorithms are consistent [30]. Hassan et al. applied the Naïve-Bayes classification technique using a machine learning-based dataset including specific header field-based attributes, the overall improvement in spam filtering were achieved [31]. The Naïve Bayes performed satisfactorily compared to the Support Vector Machine [32, 33, 34]. Naïve Bayes gave the most accurate results among Support Vector Machine and Random Forest [35]. Pandey et al. found that the SVM and Naïve Bayes methods and Logistic Regression methods have a very satisfying performance amongst the other technique. More research could be done to increase the result of the Naïve Bayes either through a hybrid system. Hybrid systems appear to be the most efficient approach to generate a successful anti-spam filter [36].

III. Proposed Methodology

This paper aims to describe and develop a classifier that can classify spam e-mail using the Naive Bayes algorithm. An e-mail contains evidence about the origin of an e-mail, Sender IP, Precedence, Errors-To, Sender, In-Reply-To, X-Spam Status, X-Spam Level, X-Mailer, X-Priority, X-Mime OLE, Content-Type, Message-ID,

ISBN: 978-93-91355-11-1

Delivered-To, Date, From, To, Received. Out of these features, only the subject, the e-mail address, and the sender's IP address are selected as classification features.

The framework is divided into two phases; Phase I is building the spam and Legitimate keywords database after processing e-mails from the training dataset. In Phase II, e-mail's subject, sender e-mail address, and sender IP addresses from each e-mail were extracted by a similar process used in Phase I and compared with the list of subject keywords, e-mail addresses, IP addresses in a separate list for spam and Legitimate keywords which were prepared in Phase I. If the extracted word from the subject from the testing e-mail matched with spam keywords, the count increases by 1.

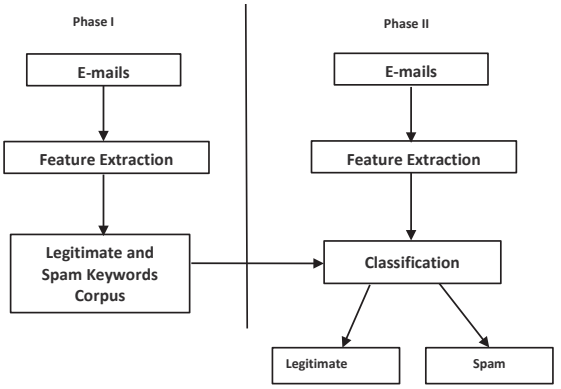

Fig. 1: Framework for SPAM E-mail Classification

A similar procedure was done for matching with a list of legitimate keywords, and if it was matched with Legitimate keywords, then the count of legitimate was increased by 1. After processing all words in the e-mail subject, the spam probability is calculated as per equation (1). This process applied to the e-mail address and IP addresses to calculate the probability of spam.

Algorithm 1: Classification Algorithm

Input: E-mail files from the datasets.

Output: Classified e-mails.

1. Phase – I
   a. Features extracted from each e-mail of the dataset.
   b. Extracted feature values are inserted in the list of keywords depending on the type of training e-mail that is spam or Legitimate.
   c. Duplicate values are removed from the keywords list.
2. Phase – II
   a. Features extracted from each e-mail present in the testing dataset.
   b. Compare the extracted value of the selected feature with each value in the keywords list.
      i. The value respective field is compared with Spam keywords if matched spam counter is increased by 1.
      ii. The value respective field is compared with Legitimate keywords if matched the Legitimate counter is increased by 1.
      iii. The probability of spam is calculated by Eq. (1) for the different fea-tures. If the probability of spam is greater than or equal to threshold 0.5, then testing e-mail is marked as SPAM; else, it is marked as LEGITIMATE.

IV. Dataset

The SpamAssassin dataset used, having a collection of spam and legitimate e-mails. Dataset consists of a total of 9349 e-mails, and 2398 are spam e-mails. To calculate an accurate result, the K-fold cross-validation method was applied. In the experiment, the dataset was divided into ten parts for performing 10-fold cross-validation. Nine parts were used for training, and the remaining part was used for testing. The probability of the testing e-mail was used to classify spam.

Table 1. SpamAssassin Dataset Details

Dataset	Period	Type of E-mail	Number of E-mails
SpamAssassin1	29-06-2004 to 11-03-2005	Spam	2398
		Legitimate	6951

Table 1 shows the details of the SpamAssassin dataset in the form of duration and count of the e-mails.

```
Received: from linux.midrange.com (dial-62-64-223-40.access.uk.tiscali.com [62.64.223.40])
          by linux.midrange.com (8.11.6/8.11.6) with SMTP id g6IDvRt21715
          for <gibbs@midrange.com>; Thu, 18 Jul 2002 08:57:28 -0500
Message-Id: <200207181357.g6IDvRt21715@linux.midrange.com>
From: "your long lost friend" <justokandgroovy@kunmail.com>
Date: Thu, 18 Jul 2002 14:58:29
To: gibbs@midrange.com
Subject: A rare and wonderful email really!
MIME-Version: 1.0
Content-Type: text/plain;charset="iso-8859-1"
Content-Transfer-Encoding: 7bit
X-Status:
X-Keywords:

Hi we are luke's secret following we love luke fictitious!
We are also your long lost friend! Hi
This email has nothing to do with lukefictitious.com
We wil be putting up our very own fan site soon
and wanted to let you know in advance!
Have a beautifull day!
```

Fig. 2: Spam E-mail Sample from SpamAssassin Dataset

Fig. 2 shows a sample spam e-mail from the SpamAssassin dataset. The e-mail contains two essential parts header and content or body. The header of the e-mail has elements as received, Message-is, From, Date, To, MIME, Content-Type, Content-Transfer-Encoding, X-status, X-Keywords, etc. For this particular experiment, only sender e-mail, send IP address and selected subject line.

[1]https://spamassassin.apache.org/old/publiccorpus/

V. Performance Measurement

The performance is measured in terms of accuracy as given in (5). The proposed algorithm calculates the probability of spam for each field.

Table 2: Performance Parameters

Classifier	Spam	Legitimate
Spam	Spam e-mails classified as spam (True-Positive)	Spam e-mails classified as Legitimate (False-Negative)
Legitimate	Legitimate e-mails classified as spam (False-Positive)	Legitimate e-mails classified as Legitimate (True-Negative)

$$Recall = \frac{TP}{TP + FN} \quad (1)$$

$$F - Score = 2 \text{ x } \left(\frac{Precision \ x \ Recall}{Precision \ x \ Recall}\right) \quad (2)$$

$$FalsePositiveRate = \frac{FP}{FP + TN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \text{ x } 100 \quad (5)$$

The performance was measured in the recall, f-score, false-positive rate, precision, and accuracy. The recall is the rate of true positives found from the total of a genuinely positive and false negative and calculated as per Eq. (2). The f-score provides an accurate measure of the performance of the test by using both precision and recall, calculated as per Eq. (3). The false-positive rate is the number of Legitimate e-mails classified as spam e-mails; concerning all Legitimate e-mails, the false-positive rate was calculated using Eq. (4). The precision is the true positives out of the total predicted positive, calculated as per Eq. (5). Accuracy is the ratio of true positive and true negative to all samples, calculated using Eq. (6).

VI. Result Analysis

The experiment results for classification are shown in the following tables and figures. Tables displays classification results in the form of recall, f-score, FPR, precision, and accuracy when an e-mail address is used as a feature. The 10-fold cross-validation method was used in the experiment, and it is referred to as a set of folds in the given tables.

Table 3: Classification Results for E-Mail Address

Set	TP	FN	FP	TN	Accuracy (%)	Recall	Precision	FP Rate	F-score
1	32	208	0	695	77.75	0.13	1.00	0.00	0.24
2	49	191	0	695	79.57	0.20	1.00	0.00	0.34
3	54	186	0	695	80.11	0.23	1.00	0.00	0.37
4	52	187	0	695	79.98	0.22	1.00	0.00	0.36
5	43	197	0	695	78.93	0.18	1.00	0.00	0.30
6	57	182	0	695	80.51	0.24	1.00	0.00	0.39
7	35	205	0	695	78.07	0.15	1.00	0.00	0.25
8	37	203	0	695	78.29	0.15	1.00	0.00	0.27
9	37	203	0	695	78.29	0.15	1.00	0.00	0.27
10	46	194	1	695	79.17	0.19	0.98	0.02	0.32
Mean	**44**	**196**	**0.1**	**695**	**79.07**	**0.18**	**1.00**	**0.00**	**0.31**

Table 4: Classification Results for IP Address

Set	TP	FN	FP	TN	Accuracy (%)	Recall	Precision	FP rate	F-score
1	164	76	41	654	87.49	0.68	0.80	0.06	0.74
2	170	70	103	592	81.50	0.71	0.62	0.15	0.66
3	204	36	97	598	85.78	0.85	0.68	0.14	0.75
4	158	81	32	663	87.90	0.66	0.83	0.05	0.74
5	205	35	90	605	86.63	0.85	0.69	0.13	0.77
6	187	52	94	601	84.37	0.78	0.67	0.14	0.72
7	202	38	52	643	90.37	0.84	0.80	0.07	0.82
8	200	40	3	692	95.40	0.83	0.99	0.00	0.90
9	209	31	0	695	96.68	0.87	1.00	0.00	0.93
10	188	52	96	600	84.19	0.78	0.66	0.14	0.72
Mean	189	51	61	634	88.03	0.79	0.77	0.09	0.77

ISBN: 978-93-91355-11-1

Table 5: Classification Results for E-mail Subject

Set	TP	FN	FP	TN	Accuracy (%)	Recall	Precision	FP rate	F-score
1	229	11	201	494	77.33	0.95	0.53	0.29	0.68
2	207	33	151	544	80.32	0.86	0.58	0.22	0.69
3	215	25	217	478	74.12	0.90	0.50	0.31	0.64
4	213	26	159	536	80.19	0.89	0.57	0.23	0.70
5	207	33	109	586	84.81	0.86	0.66	0.16	0.74
6	209	30	68	627	89.51	0.87	0.75	0.10	0.81
7	213	27	118	577	84.49	0.89	0.64	0.17	0.75
8	214	26	266	429	68.77	0.89	0.45	0.38	0.59
9	216	24	145	550	81.93	0.90	0.60	0.21	0.72
10	195	45	71	625	87.61	0.81	0.73	0.10	0.77
Mean	**212**	**28**	**151**	**545**	**80.91**	**0.88**	**0.60**	**0.22**	**0.71**

Table 3 represents the results obtained by the feature sender's e-mail address; the mean precision obtained is 1.0 that signifies the spam e-mails in the testing set are completely classified as spam. The false-positive obtained by the e-mails are almost negligible that denotes the feature is giving no misclassification when it comes to receiving Legitimate mail in the spam mailbox.

The false-positive obtained by the sender's IP address is increased as compared to Table 3; this signifies that fewer legitimate e-mails are classified as spam. If the comparison has been made between an e-mail address and IP address, it is notable that accuracy, f-score increased for IP address and with minimal false-positive rate. The IP address has the upper hand over the e-mail address in spam classification.

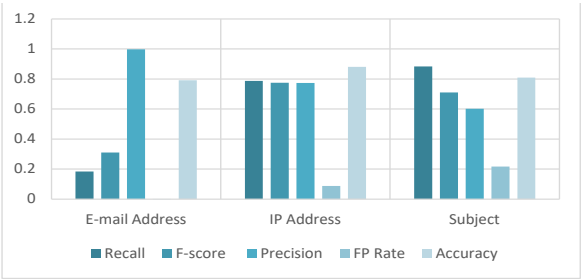

Fig. 3: Comparison of Recall, F-Score, Precision, FP Rate, and Accuracy

Table 5 signifies the decrease in accuracy and increase in the false-positive rate, the variation of words that exist in the subject line makes it less preferable than the IP address. Fig. 3 signifies that the subject as a feature for classification has more misclassification than the other two features due to the variation found in words in the subject line.

A. Feature Fusion

An improved method to get a better result by considering the results from all three features. This method keeps a counter set to form a decision; the counter increases whenever each feature classifies the e-mail as spam. For example, a testing e-mail classified as spam by subject and IP Address and e-mail address classifies it as legitimate, then the value of the counter set to 2. If only one out of three features classified it as spam, then the counter was set to 1. If the testing e-mail was classified as spam by all features, then the counter was set to 3. The final decision is made based on the counter's value; if the value of the counter is greater than or equal to 2, then the e-mail was classified as spam else marked as Legitimate.

Table 6 shows the results in the recall, F-score, FP rate, precision, and accuracy for classification considering all features to form a decision.

Table 6: Classification Results for E-mail Subject

Set	TP	FN	FP	TN	Accuracy (%)	Recall	Precision	FP rate	F-score
1	161	79	16	679	89.84	0.67	0.91	0.02	0.77
2	160	80	36	659	87.59	0.67	0.82	0.05	0.73
3	195	45	32	663	91.76	0.81	0.86	0.05	0.84
4	157	82	12	683	89.94	0.66	0.93	0.02	0.77
5	188	52	20	675	92.30	0.78	0.90	0.03	0.84
6	169	70	14	681	91.01	0.71	0.92	0.02	0.80
7	183	57	15	680	92.30	0.76	0.92	0.02	0.84
8	180	60	3	692	93.26	0.75	0.98	0.00	0.85
9	193	47	0	695	94.97	0.80	1.00	0.00	0.89
10	160	80	12	684	90.17	0.67	0.93	0.02	0.78
Mean	**175**	**65**	**16**	**679**	**91.31**	**0.73**	**0.92**	**0.02**	**0.81**

ISBN: 978-93-91355-11-1

The classification decision after considering three features gave a better f-score, false-positive rate, and accuracy. The mean accuracy obtained is greater than the highest accuracy found for the IP address. The improvement in the accuracy signifies better classification after considering the fusion of features. The fusion of features to form a decision increases the accuracy with lower misclassification.

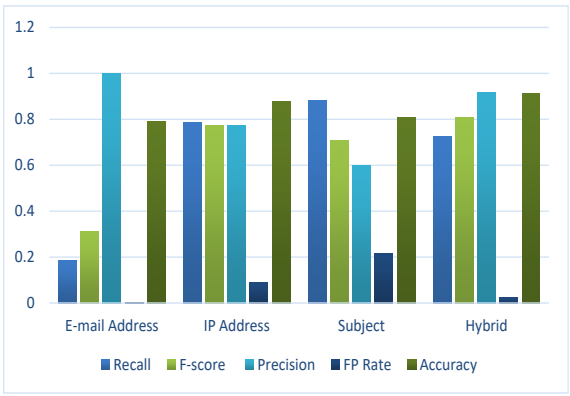

Fig. 4: Comparison of Individual and Combined Results for Different Features.

Fig. 4 reflects the improved results using hybrids decisions taken from all three features in the form of recall, precision, f-score, false-positive rate, and accuracy.

VII. Conclusion

This paper describes the spam e-mail classification using the fusion of features. The subject of the e-mail, the sender's e-mail address, and the IP address of the sender, these features were used for classification. When executed on the standard dataset of SpamAssassin with 10-fold cross-validation, the experimental gave good accuracy. It was found that an e-mail subject line is not an ideal feature to achieve better accuracy. The feature's fusion to form a classification decision provided rise inaccuracy with a low false-positive rate comparing to individual classification results.

References

[01] G. V. Cormack and T. Lynam, "Spam Corpus Creation for TREC," in *Second Conference on E-mail and Anti-Spam*, California, USA, 2005.

[02] N. Advilkar, P. Mane and D. Walunj, "SPAM MAIL FILTERING," *International Journal of Advanced Research in Computer Engineering & Technology,* vol. 5, no. 1, pp. 99-104, 01 2016.

[03] M. Siponen and C. Stucke, "Effective anti-spam strategies in companies: An international study," in *International Conference on System Sciences*, Kauia, HI, USA, 2006.

[04] Namrata and Suman, "Review Paper on Spam Detection Antiphishing Techniques," *International Journal of Computer Sciences and Engineering,* vol. 6, no. 5, pp. 1156-1161, 2018.

[05] D. Wang, D. Irani and C. Pu, "A Study on Evolution of Email Spam Over Fifteen Years," in *9th International Conference on Collborative COmputing: Networking, Application and Worksharing*, Austin, TX, USA, 2013.

[06] A. Bhowmick and S. M. Hazarika, "Machine Learning for E-mail Spam Filtering: Review,Techniques and Trends," in *Advances in Electronics, Communication and Computing*, 2016, pp. 583-590.

[07] E. P. Sanz, J. C. Cortizo Pérez and J. M. GOMEZ HIDALGO, "Email Spam Filtering," in *Advances in Computers*, vol. 74, 2008, pp. 45-114.

[08] E. Blanzieri and A. Bryl, "A survey of learning-based techniques of e-mail spam filtering," in *Artificial Intelligence Review*, 2009, pp. 63-92.

[09] P. G. Juneja and R. K. Pateriya, "A Survey on Email Spam Types and Spam Filtering Techniques," *International Journal of Engineering Research & Technology (IJERT),* vol. 3, no. 3, pp. 2309-2314, March 2014.

[10] N. Agrawal and S. Singh, "Origin (Dynamic Blacklisting) Based Spammer Detection and Spam Mail Filtering Approch," *International Conference on Digital Information Processing, Data Mining, and Wireless Communications,* pp. 99-104, 6-8 july 2016.

[11] H. Guo, B. Jin and W. Qian , "Analysis of Email Header for Forensics Purpose," in *International Conference on Communication Systems and Network Technologies*, Gwalior, India, 2013.

[12] Rekha and S. Negi, "A Review on Different Spam Detection Approaches," *International Journal of Engineering Trends and Technology (IJETT),* vol. 11, no. 6, pp. 315-318, May 2014.

[13] P. Kulkarni and H. Acharya, "Comparative analysis of classifiers for header based e-mails classification using supervised learning," *International Research Journal of Engineering and Technology,* vol. 3, no. 3, March 2016.

[14] M. M. Hassan and M. W. Hussain, "Header Based Spam Filtering Using Machine Learning Approach," *International Journal of Emerging Technologies in Engineering Research (IJETER),* vol. 5, no. 10, OCT 2017.

[15] P. Sao and K. Prashanthi, "E-mail Spam Classification Using Naïve Bayesian Classifier," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),* vol. 4, no. 6, pp. 2792-2796, 2015.

[16] T. Almeida, J. Almeida and A. Yamakami, "Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers," *Journal of Internet Services and Applications,* pp. 183-200, February 2011.

[17] M. N. Marsono, M. Watheq El-Kharashi and F. Gebali, "Binary LNS-based naı¨ve Bayes inference engine for spam control: noise analysis and FPGA implementation," *IET Computers & Digital Techniques,* vol. 2, no. 1, pp. 56-62, 2008.

[18] K. Li and Z. Zhong, "Fast Statistical Spam Filter by Approximate Classifications," in *Joint international Conference on Measurement and Modeling of Computer Systems*, Saint Malo, France, 2006.

[19] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras and C. D. Spyropoulos, "An Evaluation of Naive Bayesian Anti-Spam Filtering," in *11th European Conference on Machine Learning*, Barcelona, Spain, 2000.

ISBN: 978-93-91355-11-1

[20] E. Blanzieri and A. Bryl, "A Survey of Learning-Based Techniques of E-mail Spam Filtering," 2008.

[21] E. Yitagesu and M. Tijare, "E-mail Classification using Classification Method," *International Journal of Engineering Trends and Technology (IJETT),* vol. 32, no. 3, pp. 142-145, 2016.

[22] N. F. Rusland, N. Wahid, S. Kasim and H. Hafit, "Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets," in *International Research and Innovation Summit (IRIS2017)*, 2017.

[23] L. Chih-Chin, "An Empirical Study of Three Machine Learning Methods for Spam Filtering," *Knowledge-Based Systems,* vol. 20, 2006.

[24] S. Youn and D. McLeod, "A Comparative Study for Email Classification," in *Advances and Innovations in Systems, Computing Sciences and Software Engineering*, 2007.

[25] K. P. Clark, "A Survey of Content-based Spam Classiflers," 2008.

[26] W. A. Awad and S. M. ELseuofi, "Machine Learning Methods for Spam E-Mail Classification," *International Journal of Computer Science & Information Technology,* vol. 3, no. 1, FEB 2011.

[27] T. Kumaresan, S. SanjuShree, K. Suhasini and C. Palanisamy, "Machine Learning Algorithms Used In Spam Filtering Machine Learning Algorithms Used In Spam Filtering – A Study," *International Journal Of Advanced Information And Communication Technology,* vol. 1, no. 7, pp. 548-551, Nov 2014.

[28] P. Sao and K. Prashanthi, "E-mail Spam Classification Using Naïve Bayesian Classifier," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),* vol. 4, no. 6, pp. 2792-2796, 2015.

[29] N. Chopra and K. Gaikwad, "Image and Text Spam Mail Filtering," *International Journal of Computer Technology and Electronics Engineering (IJCTEE),* vol. 5, no. 3, pp. 15-18, June 2015.

[30] A. Sharaff, N. K. Nagwani and K. Swami, "Impact of Feature Selection Technique on Email Classification," *International Journal of Knowledge Engineering,* vol. 1, no. 1, pp. 59-63, 2015.

[31] M. M. Hassan and M. W. Hussain, "Header Based Spam Filtering Using Machine Learning Approach," *International Journal of Emerging Technologies in Engineering Research (IJETER),* vol. 5, no. 10, OCT 2017.

[32] V. Goswami, V. Malviya and P. Sharma, "Detecting Spam E-mails/SMS Using Naive Bayes, Support Vector Machine and Random Forest," *Innovative Data Communication Technologies and Application,* pp. 608-615, 2019.

[33] N. Mirza, B. Patil, T. Mirza and R. Auti, "Evaluating Efficiency of Classifier for E-mail Spam Detector Using Hybrid Feature Selection Approaches," in *International Conference on Intelligent Computing and Control Systems*.

[34] Pooja and K. K. Bhatia, "Spam Detection using Naive Bayes Classifier," *International Journal of Computer Sciences and Engineering,* vol. 6, no. 7, pp. 712-716, 2018.

[35] S. Mani, S. Kumari, A. Jain and P. Kumar, "Spam Review Detection Using Ensemble Machine Learning," in *Machine Learning and Data Mining in Pattern Recognition*, Springer, Cham, 2018, pp. 198-209.

[36] P. Pandey, C. Agrawal and T. N. Ansari, "A Hybrid Algorithm for Malicious Spam Detection in E-mail through Machine Learning," *International Journal of Applied Engineering Research,* vol. 13, no. 14, pp. 16971-16979, 2018.

ISBN: 978-93-91355-11-1