Shripriya Dongre et al. 2019, 7:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

A Survey: E-Mail Spam Classification using Machine Learning Techniques

M.E. Scholar (CSE) Shripriya Dongre
Jawaharlal Institute of Tech. Borawan ,
Shridongre11@gmail.com

HOD (CSE) Prof. Kamlesh Patidar
Jawaharlal Institute of Tech. Borawan
Kt125412@gmail.com

Abstract

E-mail is one of the most secure medium for online communication and transferring data or messages through the web. An overgrowing increase in popularity, the number of unsolicited data has also increased rapidly. To filtering data, different approaches exist which automatically detect and remove these untenable messages. There are several numbers of email spam filtering technique such as Knowledge-based technique, Clustering techniques, Learning-based technique, Heuristic processes and so on. This paper illustrates a survey of different existing email spam filtering system regarding Machine Learning Technique (MLT) such as Naive Bayes, SVM, K-Nearest Neighbor, Bayes Additive Regression, KNN Tree, and rules. However, here we present the classification, evaluation and comparison of different email spam filtering system and summarize the overall scenario regarding accuracy rate of different existing approaches.

Keywords: E-mail spam; unsolicited bulk email; spam filtering methods; machine learning; algorithm.

I. INTRODUCTION

In recent years, internet has been created several platforms for making human life become more secure. Among these; e-mail is a substantial platform for user communication. Email is nothing; simply it's called an electronic messaging framework which transmits the message from one user to another [1].

Nowadays, e-mail has turned into a typical medium [2] because of its several branches like Yahoo mail [3], Gmail [4], Outlook [5] etc, which are completely free for all web user by following some administration [6, 7]. At present, Email called a secure worldwide communication medium for its several functions. But sometimes email becomes more hazardous for some "Spam Email". Generally, Spam email called as junk email or unsolicited message which sent by spammer through Email. The process is, collected the address on the web and sends the message through domain's username. Actually, it has been produced for financial profits using the assortment of procedures [8] and instruments that incorporate spoofing, bonnets, open intermediaries, mail transfers, bulk mail instruments called mailers, and so forth. Spam filtering is a challenging undertaking for an assortment of reasons. For spam email, users are facing several problems like abuse of traffic, limit the storage space, computational power, become a barrier for finding the additional email, waste users time and also threat for user security [9, 10].

So, becoming email more secure and effective, appropriate Email filtering is essential. Several types of researches have been performed on email filtering, some acquired good accuracy and some are still going on. According to researcher's overview, Email filtering is a process to sort email according to some criteria. As there are various methods exist for email filtering, among them, inbound and outbound filtering is well known. Inbound filtering is the process to read a message from internet address and outbound filtering is to read the message from the local user. Moreover, the most effective and useful email filtering is

Spam filtering which performs through antispam technique. As spammers are proactive natures and using dynamic spam structures which have been changing continuously for preventing the anti-spam procedures and thus making spam filtering is a challenging task [9, 10]. Spam filtering is a process to detect unsolicited massage and prevent from entering into user's inbox. Now days, various systems have been existed to generate anti-spam technique for preventing unsolicited bulk email.

Most of the anti-spam methods have some inconsistency between false negatives (missed spam) and false positives (rejecting good emails) which act as a barrier for most of the system to make successful antispam system. Therefore, an intelligent and effective spam-filtering system is the prime demand for web users. Among various approach, Fiaidhi et al. [11] and Arora et al. [12] proposed method evaluate that, 70% today's business email's are spam [13]. Spam filtering has two major section; "Knowledge engineering" and "Machine learning".

Knowledge engineering is an arrangement of guidelines to determine the spam emails. In contrast, Machine learning is more efficient than knowledge engineering. It does not require any predefined rules. Naive Bayes, Support Vector Machines, Neural Networks, K-nearest neighbor, Rough sets, and artificial immune system are some prominent technique of Machine learning for spam filtering those are works by matching the regular expression, keywords from message text and so on.

II. SEVERAL EMAIL SPAM FILTERING METHODS

At present, number of spam email has increased for several criteria such as an advertisement, multi-level marketing, chain letter, political email, stock market advice and so forth. For restricting spam email, several methods or spam filtering system has been constructed by using various concept and algorithms. This section concluded by describing few of spam filtering methods to understand the process of spam filtering and its effectiveness.

1. Standard Spam Filtering Method

Email Spam filtering process works through a set of protocols to determine either the message is spam or not. At present, a large number of spam filtering process have existed. Among them, Standard spam filtering process follows some rules and acts as a classifier with sets of protocols. Figure.1 shows that, a standard spam filtering process performed the analysis by following some steps [14].

First one is content filters which determine the spam message by applying several Machines learning techniques [8, 10, 15-18]. Second, header filters act by extracting information from email header. Then, backlist filters determine the spam message and stop all emails which come from backlist file. Afterward, "Rules-based filters" recognize sender through subject line by using user defined criteria [19]. Next, "Permission filters" send the message by getting recipients pre-improvement. Finally, "Challenge response filter" performed by applying an algorithm for getting the permission from the sender to send the mail.

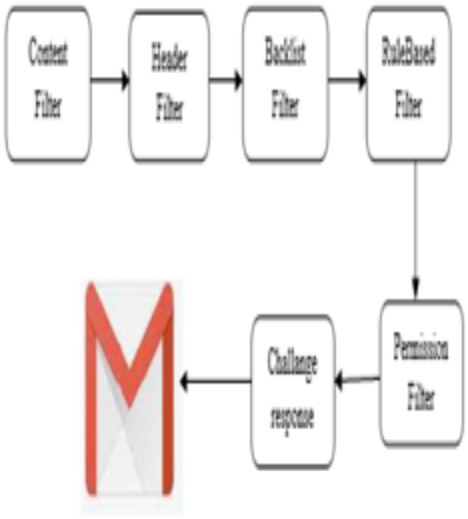

Figure 1: A Standard Process of Email Spam Filtering System.

2.Client Side and Enterprise Level Spam Filtering Methods.

A client can send or receive an email by just one clicking through an ISP. Client level spam filtering provides some frameworks for the individual client to secure mail transmission. A client can easily filter spam through these several existing frameworks by installing on PC. This framework can interact with MUA (Mail user agent) and filtering the client

inbox by composing, accepting and managing the messages [2].Enterprise level spam filtering is a process where provided frameworks are installing on mail server which interacts with the MTA for classifying the received messages or mail in order to categorize the spam message on the network. By this system, a user on that network can filter the spam by installing appropriate system [21, 22] more efficiently.

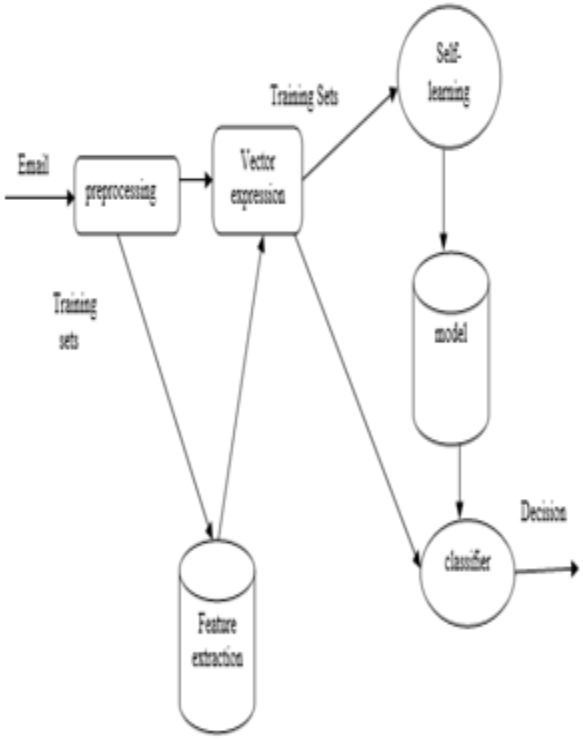

Figure 2: Spam Filtering Architecture by Applying Machine Learning Techniques.

Here, describes a sample of case base spam filtering architecture by applying Machine learning techniques [Fig. 3] in detail. The full process perform through several steps which followed by the figure 3. At the first step, extracted all email (spam email and legitimate email) from individual users email through collection model.

Then, the initial transformation starts with the pre-processing steps through client interface, highlight extraction and choice, email data classification, analyzing the process and by using vector expression classifies the data into two sets. Finally, machine learning technique is applied on training sets and testing sets to determine email whether it is spam or legitimate. The final decision makes through two steps; through self observation and classifier's result to make decision whether the email is spam or legitimate.

III. OVERVIEW OF SEVERAL EXISTING EMAIL SPAM FILTERING SYSTEMS FOR MACHINE

Mohammed et al. [11] [2013] proposed an approach for Classifying Unsolicited Bulk Email (UBE) using Python Machine Learning Techniques with the help of spam filtering which performs the work by creating a spam-ham dictionary from the given training data and applying data mining algorithm to filter the training and testing data. After applying various classifier on1431 dataset, the approach predicts that, Naïve Bays and SVM classifiers are the prominent classifier for spam filtering or classification.

Subramaniam et al. [23] [2012] implemented Naïve Bayesian Anti-spam Filtering Technique on Malay Language to investigate the utilization of Naïve Bayesian procedure to combat spam issue. An experiment conducted through Naïve Bayesian method for filtering Malay language spam and the result depicts that, propose approach has gained 69% accuracy. They realized that by reducing false positive and expanding training corpus the result would much better for classifying Malay language spam.

Sharma et al. [24] [2013] described Adaptive Approach for Spam Detection. This article consider SPAMBASE dataset and various machine learning technique such as Bays Net, Logic Boost, Random tree, JRip, J48, Multilayer Perception, Kstar, Random Forest, Random Committee are applied for classifying the spam. It measures the accuracy by grouping the spam/non-spam e-mails from labeled emails of a single account. The paper estimates that, total accuracy was 95.32% which depicts the quality of the proposed approach.

Banday et al. [25] [2008] discuss the procedures of statistical spam filters design by incorporating Naïve Bayes, KNN, SVM, and Bayes Additive Regression Tree. Here evaluates these procedures in terms of accuracy, recall, precision, etc. Though all machine learning classifiers are effective but according to this approach, CBART and NB classifiers has better capability to spam filtering. This approach estimates that during spam filtering calculations of false positive are more costly than false negative.

Awad et al. [1] [2011] proposed an ML- based approach on for Spam E-mail Classification. In this article present the most prominent machine learning strategies and its effectiveness regarding spam email classification. Here introduced Portrayals algorithms and the performance of Spam Assassin corpus. The result shows that, Naïve bays and rough sets methods are the promising algorithms for email classification. They perform their future research to improve the Nave Bays and Artificial immune system by hybrid system or by resolution the feature reliance issue.

Chhabra et al. [26] [2010] developed Spam Filtering using Support Vector Machine by considering Nonlinear SVM classifier with different kernel functions over Enron Dataset. Here considered six datasets and perform the analysis of datasets having diverse spam: ham ratio and makes satisfactory Recall and Precision Value.

Tretyakov et al. [27] [2004] discussed Machine Learning Techniques through Spam Filtering. In this article compared the precision between before eliminating false positive and after eliminating false positive. They represent the result that the result becomes more reliable considering both precision results (before eliminating and after eliminating false positive) either taking one.

Shahi et al. [28] [2013] developed Mobile SMS Spam Filtering for Nepali Text Using Naïve Bayesian and Support Vector Machine. The fundamental concern of this study was to look at the effectiveness of Naïve Bayesian and SVM Spam filters. The correlation of productivity between these Spam filters was done based on the precision and recall. Approach showed that Nave Bays produce better accuracy than SVM.

Kaul et al. [29] [2004] implemented Filtering Spam E-mail with Support Vector Machines. Here in this paper they consider a virtual machine called Spam Stop. Spam Stop performs on the large dataset to produce more accurate result. It has a drawback such as Spam Stop does not yet incorporate an assortment of standard pre-filtering mechanisms.

Suganya et al. [30] [2014] worked on short message and misspelling of data on online Social Networks (OSNs) user post. They used machine learning technique with content- based features for short message and Filtered Wall (FW) [31] to evaluate system for filtering spam massage. They categorized the classification process into two levels; first-level classifier performs on Neutral and Non-neutral through hard binary categorization and second level classifier performs through RBFN model [32].

Rathi et al. [33] [2013] proposed an approach using Data mining technique for finding the best classifier for email classification. They analyzed various data mining technique for measuring the performance of several classifiers through "with feature selection algorithm" and "without feature selection algorithm". After selecting the Best feature selection algorithm, they considered the selected algorithm for their feature selection purpose. They experiment their data by using several algorithms such as Naïve Bayes, Bayes Net, Support vector machine, and Function tree, J48,Random Forest and Random Tree. The whole dataset consists of 58 attributes and 4601 instances. Considering Random Tree algorithm highest accuracywas 99.72% and the lowest accuracy was 78.94% forNaïve Bayes algorithm.

Mohammed et al. [11] [2013] presents anapproach for filtering spam email using machine learning algorithms. At first, they filter Spam and Ham word from the training datasets by applying tokenization method based on these token create the testing and training table using various data mining algorithm. Then find the frequency of spam and ham tokens for measuring the probability which is suggested by Paul Graham [34]. For ham token, the probability value was0 and for spam token probability value was 1. They used Nielson Email-1431 [35] dataset and emphasized that the Naïve Bayes and Support Vector Machine are the most effective classifier.

Singh et al. [36] [2018] discussed the solution and classification process of spam filtering and presented a combining classification technique to get better spam filtering result. With the help of Data mining, they collected all the information of previous failures, success and current problems of spam filtering. In this method, researchers used binary value where 1 for spam email and 0 for not spam emails. But its success rate was very poor. So they apply NB, KNN, SVM, Artificial Neural Network classification method and find their accuracy. Based on these two techniques (machine learning and knowledge engineering) effectiveness, they adopt a classification technique for spam filtering. Moreover, here first collect data from user training set, compared and find the spam email and then use a global training set to

optimize the classification technique. Using this technique increases the precision rate at least 2%.

Abdul amid et al. [37] [2018] introduced performance analysis based approach by using some classification techniques such as Bayesian Logistic Regression, Hidden Naïve Bayes, Logit Boost, Rotation Forest, NNge, Logistic Model Tree, REP Tree, Naive ayes, Radial Basis Function (RBF) Network, Voted Perception, Lazy Bayesian Rule, Multilayer Perception, Random Tree and J48. The competence of these techniques classified through Accuracy, Precision, Recall, F-Measure, Root Mean Squared Error, Receiver Operator Characteristics Area and Root Relative Squared Error using Spam base dataset and WEKAdata mining tool. For conducting the performance and comparison, datasets are considered from UCI Machine Learning Repository. Considering Rotation Forest algorithm acquired the highest accuracy was 0.942 and the REP Tree algorithm showed the lowest accuracy was 0.891. They applied the F-measure method for finding precision and recall. The highest F-measure considered from Rotation forest algorithm and lowest Measure considered from Naïve Bayes algorithm. For finding the probability use ROC curves on randomly selected positive and negative instance and for Rotation forest algorithm the ROC curves carried the highest score was 0.98. In contrast, Random Tree having the lowest score which was 0.905. For finding the statistics result, they use kappa Statistics and the result was much better for Rotation Forest algorithm which approximately 0.879. This paper showed that, Rotation Forest classifier gained the best result with 0.942accuracies, then J48 with 0.923, Naïve Bayes with 0.885and Multilayer Perception with 0.932.

Sah et al. [38] [2017] proposed a method for detecting of malicious spam through feature selection and improve the training time and accuracy of malicious spam detection system. They also showed the comparison of difference classifier as Naïve Bayes (NB)and Support Vector Machine (SVM) based on accuracy and computation time. The proposed approach completed by four steps such as preparing the text data, creating word dictionary, Feature extraction process and training the classifier. For preparing text data researchers split the dataset into the training set(702 mails) and a test set (260 mails) and divided into spam and ham mails. Performed feature selection process by generating feature vector matrix. According to the approach, Naïve Bayes selected as good classifiers among others.

Verma et al. [39] [2017] proposed a method for spam detection using Support Vector Machine algorithm and feature extraction. This methodology works through several steps such as Email collections, pre-processing, feature extraction, SVM training, test classifier, top word predictors, test email and result. First they take a dataset from Apache Public corpus. In pre-processing section, they remove all special symbol, URL and HTML tags and also unnecessary alphabet. Then they mapped all word from the dictionary using Vocab file. SVM classifier applied on the training dataset. The Accuracy of the system was 98%.

Rusland et al. [40] [2017] perform the analysis using Naïve Bayes algorithm for email spam filtering on two datasets which are evaluated based on the accuracy, recall, precision and F-measure. Naïve Bayes algorithm is a probability-based classifier and the probability is counting the frequency and combination of values in a dataset. This research performed through three phases such as pre-processing, Feature Selection, and implementation through Naïve Bayes Classifier. First they remove all conjunction words, articles from the email body in pre-processing section. Made two datasets through WEKA tool; one is a Spam Data and another is the Spam Base dataset. The average accuracy was 8.59% by considering two datasets where Spam data get 91.13% and the Spam Base data get 82.54% accuracy. The average precision for Spam Base was 88% and for Spam data was 83%. They proposed that, Naïve Bayes classifier performs better on Spam Base data compared with Spam Data.

Yuksel et al. [41] [2017] use Support Vector Machine and Decision tree for spam filtering. The Decision tree used in data mining and the support vector machines as a supervised learning model which can analyze the data for spam classification. First data was divided into two sections; one is training and other is test data, then the algorithm was trained and evaluated through Microsoft Azure platform which provides tools for machine learning and compared results with decision tree and support vector machine algorithm. The result of SVM method was 97.6% and for Decision tree the result was 82.6%. The result estimate that, SVM classifier performed better than DT.

Choudhary et al. [42] [2017] presented a novel approach using machine learning classification algorithm for finding and classifying SMS spam by using Short Message Service (SMS). The first step in this approach is feature selection and for that, they work on presence of mathematical symbols: UGLs, Dots, special symbols, emotions, Lowercased words and Uppercased words, mobile number, keyword specific and the message length in the SMS. After that they created a system design and collected a dataset which contained 2608 emails out of 2408 collected SNS Spam Corpus. The SMS Spam Corpus v.0.1 consists two sets of messages as SMS Spam Corpus v.0.1 Small and SMS Spam Corpus v.0.1 Big. Using "WEKA tools" for five machine learning approaches; such as Naive Bayes, Logistic Regression, J48, Decision Tree and Random Forest. Evaluating result uses with True Positive Rate (TP) and True Negative Rate (TN). False Positive Rate (FP), False Negative Rate (FN), Precision, Recall, Measure and Receiver Operating Characteristics (ROC) area achieved 96.5% true positive rate and 1.02% false positive rate with Random Forest machine learning algorithm and it performs better algorithm with high rate accuracy.

DeBarr et al. [43] [2009] use Random Forest algorithms for classification of spam email then refining the classification model using active learning. They take data from RFC 822(Internet) email message and divided each email into two sections and converted each message to term frequency and inverse document frequency (TF/IDF) features. Here select an initial set of email message using clustering technique to label as training examples and for clustering used Partitioning Around Melodies (PAM) algorithm. After considering the cluster prototype messages for training they experiment with some algorithm Random Forest, Naive Bayes, SVM and kNN. Here Random Forest algorithm performs the best classifier with 95.2% accuracy.

IV. SUMMARY OF EXISTING E-MAIL SPAM CLASSIFICATION APPROACHES

Since last few decades, researchers are trying to make email as a secure medium. Spam filtering is one of the core features to secure email platform. Regarding this several types of research have been progressed reportedly but still there are some untapped potentials. Over time, still now e-mail spam classifications one of the major areas of research to bridge the gaps. Therefore, a large number of researches already have been performed on email spam classification using several techniques to make email more efficient to the users. That's why, this paper tried to arrange the summarized version of various existing Machine Learning approaches. In addition, in order to evaluates the most of the approaches like Random Forest, Naive ayes [11, 23, 43], SVM [8, 10, 18], kNN [27, 36], and Random Forest [15, 16] used reliable and well known dataset for benchmarking performance such as Spam Data [16], The Spam Assassin [44], The Spam base, Ecml-pkdd 2006 challenge dataset [45], PU corpora dataset [15], Enron dataset [46],Trec 2005dataset [47]. Some of these dataset are in a prepared structure e.g. ECML and data accessible in Spam baseUCI archive [20]. Among them, some of the classifiers also used novel methods applied in the feature selection for improving classification such as [1, 11].

Table -1: Summary of different existing email spam classification approaches regarding Machine Learning Techniques

Sr. No.	Author	Algorithms	Corpus or Datasets	Accuracy/Performance
1	Mohammed et al.	Naïve Bayes,SVM,KNN,Decision Tree,Rules	Email-1431	85.96% Accuracy Achieved
2	Subrmaniam et al.	Naive Bayesian	Collection ofspame mails from Google's Gmail Account	96.00% Accuracy Achieved
3	Sharma et al.	Various Machine LearningAlgorithms Adaptions	SPAMBASE	94.28% Accuracy Achieved
4	Banday et al.	Naive Bayes, K-NearestNeighbor, SVM, classification Bayes	Real life data set	96.69Accuracy Achieved

		Additive Regression Tree		
5	Awad et al.	Naive Bayes, SVM, k-Nearest Neighbor, Artificial NeuralNetworks,Rough Sets	Spam Assassin	99.46% Accuracy Achieved
6	Chhabra et al.	Nonlinear SVM classifier.	Enron dataset	For Dataset 3, spam: real,therationis1:3,for satisfactory RecallandPrecision Values
7	Tretyakov	Bayesian classification, k-NN, ANNs, SVMs	PU1 corpus	94.4% AccuracyAchieved
8	Shahi et al.	Naïve Bayes, SVM	Nepali SMS	92.74% Accuracy Achieved
9	Kaul et al	SVM	Sample emails	90% ~ 95%AccuracyAchieved
10	Sugany a et al.	Rule Baseed Method	Online Social Networks (OSNs) user post	Excellence Accuracyfor Given Datasets
11	Rathi et al.	Naive Bayes, BayesNet ,SVM, and RandomForest	Custom Collection	99.72% Accuracy Rate
12	Mohammed et al.	WordFilterization by Tokenization,Appling	Nielson Email-1431	Reported Satisfactory Accuracyfor Proposed Method
13	Singh et al.	Naive Bayes, k-Nearest Neighbor, SVM, Artificial Neural Network.	Custom Collection	Reported Improvement of precision rate at least2%
14	Abdulhamid et al.	Various Machine Learning Algorithms	UCI Machine Learning Repository	94.2% Accuracy Achieved
15	Sah et al.	Naïve Bayes, SVM	&Custom Collection	Reported good Accuracy overall
16	Verma et al.	Customised SVM	ApachePublic Corpus	98% Accuracy Rate Reported
	Ruslan d et al.	Modified Naive Bayeswithselective features	Spam Base, SpamData	Spam Baseget88%Precision Rateand Spam Dataget83%
18	ksel et al.	Microsoft Azure platform defined decision tree and SVM	Custom Collection	SVM Accuracy 97.6% Decision Tree Accuracy 82.6%
19	Choudhary et al.	Feature Engineered NaiveBayes	The SMS Spam Corpusv. 0.1	96.5% True Positive Rate Accuracy
20	DeBarr et al.	Random Forest algorithm	Custom Collection	95.2% Accuracy

V. DISCUSSION

From the observation, it seems that, the majority of email spam filtering process performed through Machine learning technique using Naïve Bayes and SVM algorithm. Most of the approaches adopt different dataset such as "ECML" data and Spam base UCI archive [20]. Among several papers, Mohammad et al. introduce a classifier for feature

selection which regarded as the most novel classifier for feature selection [1, 11]. Rathi et al proposed an approach considering "Naïve Bayes", "Bayes Net", "SVM" and "Random forest" algorithm and obtain the higher accuracy than others which approximately crossed 99.72% accuracy [32]. Another one is, Awad et al. which proposed an approach considering "Naïve Bayes", "SVM", "K-Nearest Neighbor", "Artificial neural Networks", "Rough sets" algorithm and obtain 99.46% accuracy which seems good on their effectiveness [1]. After the analysis it should predict that, "Naïve Bayes" and "SVM" algorithm is the most effective algorithm in machine learning technique and have the ability to better classification of email spam.

VI. CONCLUSION

This survey paper elaborates different Existing Spam Filtering system through Machine learning techniques by exploring several methods, concluding the overview of several Spam Filtering techniques and summarizing the accuracy of different proposed approach regarding several parameters. Moreover, all the existing methods are effective for email spam filtering. Some have effective outcome and some are trying to implement another process for increasing their accuracy rate. Though all are effective but still now spam filtering system have some lacking which are the major concern for researchers and they are trying to generate next generation spam filtering process which have the ability to consider large number of multimedia data and filter the spam email more prominently.

REFERENCES

[1]. Awad, W. A., &ELseuofi, S. M. (2011). Machine Learning methods for E-mail Classification. International Journal of Computer Applications,16(1).

[2]. Saad, O., Darwish, A., &Faraj, R. (2012). A surveyof machine learning techniques for Spamfiltering. International Journal of Computer Science and Network Security (IJCSNS),12(2), 66.

[3]. Chen, Y., Jain, S., Adhikari, V. K., Zhang, Z. L., &Xu,K. (2011, April). A first look at inter-data center traffic characteristics via yahoo! Datasets. InINFOCOM,2011 Proceedings IEEE(pp. 1620-1628). IEEE.

[4]. Barlow, K., & Lane, J. (2007, October). Like technology from an advanced alien culture: Googleapps for education at ASU. In Proceedings of the35th annual ACM SIGUCCS fall conference(pp. 8-10). ACM.

[5]. Fisher, D., Brush, A. J., Gleave, E., & Smith, M. A.(2006, November). Revisiting Whittaker &Sidner'semail overload ten years later. In Proceedings of the2006 20th anniversary conference on Computer supported cooperative work (pp. 309-312). ACM.

[6]. Blanzieri, E., &Bryl, A. (2008). A survey of learning based techniques of email spam filtering. Artificial Intelligence Review,29(1), 63-92.

[7]. Cunningham, P., Nowlan, N., Delany, S. J., &Haahr,M. (2003, May). A case-based approach to spam filtering that can track concept drift. In The ICCBR (Vol. 3, pp. 03-2003).

[8]. Drucker, H., Wu, D., & Vapnik, V. N. (1999). Support vector machines for spam categorization. IEEE Transactions on Neural networks, 10(5), 1048-1054.

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

[10]. Wang, Q., Guan, Y., & Wang, X. (2006). SVM-Based Spam Filter with Active and Online Learning. In TREC.

[11]. Mohammed, S., Mohammed, O., Fiaidhi, J., Fong,S. J., & Kim, T. H. (2013). Classifying Unsolicited Bulk Email (UBE) using Python Machine Learning Techniques.

[12]. Harisinghaney, A., Dixit, A., Gupta, S., & Arora, A.(2014, February). Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm. In Optimization, Reliability, and Information Technology (ICROIT),2014 International Conference on (pp. 153-155).IEEE.

[13]. Scholar, M. (2010). Supervised learning approach for spam classification analysis using data mining tools. organization, 2(8), 2760-2766

[14]. Christina, V., Karpagavalli, S., & Suganya, G.(2010). A study on email spam filtering

techniques. International Journal of Computer Applications,12(1), 0975-8887.

[15]. Metsis, V., And routsopoulos, I., & Paliouras, G.(2006, July). Spam filtering with naive bayes-which naive bayes? In CEAS (Vol. 17, pp. 28-69).

[16]. Androutsopoulos, I., Koutsias, J., Chandrinos, K. V.,Paliouras, G., & Spyropoulos, C. D. (2000). An evaluation of naive bayesian anti-spam filtering. arXiv preprint cs/0006013.

[17]. Hovold, J. (2005, July). Naive Bayes Spam Filtering Using Word-Position-Based Attributes. In CEAS(pp. 41-48).

[18]. Hidalgo, J. M. G. (2002, March). Evaluating cost sensitive unsolicited bulk email categorization In Proceedings of the 2002 ACM symposium onApplied computing(pp. 615-620). ACM.

[19]. Androutsopoulos, I., Paliouras, G., Karkaletsis, V.,Sakkis, G., Spyropoulos, C. D., & Stamatopoulos, P.(2000). Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach. arXiv preprint cs/0009009.

[20]. Fawcett, T. (2003). In vivo spam filtering: achallenge problem for KDD.ACM SIGKDDExplorations Newsletter, 5(2), 140-148

[21]. Wu, C. T., Cheng, K. T., Zhu, Q., & Wu, Y. L. (2005,September). Using visual features for anti-spam filtering. In Image Processing, 2005. ICIP 2005. IEEE International Conference on(Vol. 3, pp. III-509).IEEE.

[22]. Cormack, G. V., Gómez Hidalgo. M., &Sánz, E.P. (2007, November). Spam filtering for short messages. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (pp. 313-320). ACM.

[23]. Subramanian, T., Jalab, H. A., &Taqa, A. Y. (2010).Overview of textual anti-spam filtering techniques. International Journal of PhysicalSciences,5(12), 1869-1882

[24]. Sharma, S., & Arora, A. (2013). Adaptive approach for spam detection. International Journal of Computer Science Issues,10(4), 23-26.

[25]. Banday, M. T.,& Jan, T. R. (2009). Effectiveness and limitations of statistical spam filters.arXivpreprint arXiv:0910.2540

[26]. Chhabra, P., Wadhvani, R., & Shukla, S. (2010).Spam filtering using support vector machine. Special Issue IJCCT,1(2), 3

[27]. Tretyakov, K. (2004, May). Machine learning techniques in spam filtering. InData Mining Problem-oriented Seminar, MTAT(Vol. 3, No. 177,pp. 60-79).

[28]. Shahi, T. B., & Yadav, A. (2013). Mobile SMS spam filtering for Nepali text using naïve bayesian and support vector machine. International Journal of Intelligence Science, 4(01), 24.

[29]. Drucker, H., Wu, D., & Vapnik, V. N. (1999). Support vector machines for spam categorization. IEEE Transactions on Neural networks, 10(5), 1048-1054.

[30]. Suganya, T., Sridevi, K., & ArulPrakash, M. Detection of Spam in Online Social Networks (OSN) Through Rule-based System

[31]. Rahane, U., Lande, A., Bavikar, O., Chavan, S., & Shedge, K. N. International Journal of Engineering Sciences & Research Technology Advanced Filtering System to Protect OSN user Wall From Unwanted Messages.

[32]. Moody, J., & Darken, C. J. (1989). Fast learning in networks of locally-tuned processing units. Neural computation, 1(2), 281-294.

[33]. Rathi, M., & Pareek, V. (2013). Spam mail detection through data mining-A comparative performance analysis. International Journal of Modern Education and Computer Science, 5(12), 31.

[34]. Graham, P. (2002). A plan for spam (http://www. paulgraham.com/spam.html).

[35]. Kang, N., Domeniconi, C., & Barbará, D. (2005, November). Categorization and keyword identification of unlabeled documents. In Data Mining, Fifth IEEE International Conference on (pp. 4-pp). IEEE

[36]. Singh, V. K., & Bhardwaj, S. (2018). Spam Mail Detection Using Classification Techniques and Global Training Set. In Intelligent Computing and Information and Communication (pp. 623-632). Springer, Singapore.

[37]. Shafi'i Muhammad Abdul amid, M. S., Osho, O., Ismaila, I., & Alhassan, J. K. (2018). Comparative Analysis of Classification Algorithms for Email Spam Detection.

[38]. Sah, U. K., & Parmar, N. (2017). An approach for Malicious Spam Detection in Email with comparison of different classifiers

[39]. Verma, T. (2017). E-Mail Spam Detection and Classification Using SVM and Feature Extraction.

[40]. Rusland, N. F., Wahid, N., Kasim, S., & Hafit, H. (2017, August). Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets. In IOP Conference Series: Materials Science and Engineering (Vol. 226, No. 1, p. 012091). IOP Publishing.

[41]. Yüksel, A. S., Cankaya, S. F., & Üncü, İ. S. (2017). Design of Machine Learning Based Predictive Analytics System for Spam Problem. Acta Physica Polonica, A., 132(3).

[42]. Choudhary, N., & Jain, A. K. (2017). Towards Filtering of SMS Spam Messages Using Machine Learning Based Technique. In Advanced Informatics for Computing Research (pp. 18-30). Springer, Singapore.

[43]. DeBarr, D., & Wechsler, H. (2009, July). Spam detection using clustering, random forests, and active learning. In Sixth Conference on Email and Anti-Spam. Mountain View, California (pp. 1-6).

[44]. Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. Expert Systems with Applications, 36(7), 10206-10222

[45]. Mavroeidis, D., Chaidos, K., Pirillos, S., Christopoulos, D., & Vazirgiannis, M. (2006). Using tri-training and support vector machines for addressing the ECML/PKDD 2006 discovery challenge. In Proceedings of ECMLPKDD 2006 Discovery Challenge Workshop (pp. 39-47).

[46]. Klimt, B., & Yang, Y. (2004, July). Introducing the Enron Corpus. In CEAS.

[47]. Bratko, A., & Filipic, B. (2005, November). Spam Filtering Using Character-Level Markov Models: Experiments for the TREC 2005 Spam Track. In TREC.