International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835 Volume-4, Issue-3, Mar.-2017
http://iraj.in

A SURVEY OF VARIOUS MACHINE LEARNING ALGORITHMS ON EMAIL SPAMMING

[1]ESHA BANSAL, [2]PRADEEP KUMAR BHATIA

[1]Research Scholar, Department of Computer Science Dravidian University, Kuppam
[2]Chairperson, Dept. of Computer Sc. & Eng. GJUS&T, Hisar
E-mail: [1]eshabansal2006@gmail.com , [2]pkbhatia.gju@gmail.com

Abstract- One of the greatest accepted methods of communication involves the use of e-mail for personal messages or for business purpose. One of the considerable concerns of using the email is the problem of e-mail spam. The worst part of the spam emails is that, these are invading the users beyond their consent and bombarding of these spam mails fills up the whole email space of the user along with that, the issue of the wasting the network capacity and time consumption in checking and deleting the spam mails makes it even more concerning issue. Spam is a leading headache that attacks the purpose of electronic mails. So, there is appropriate substantial to distinguish ham emails from spam emails; many methods have been proposed for classification of email as spam or ham. Spam filtering is a technique which discovers nonessential, unsolicited, junk emails such as spam emails, and prevents them from getting into the user's inbox. With the increasing demand of removing the spam mails the area has become magnetic to the researchers. The filter classification can be categorized into two techniques - based on machine learning technique and those based on non-machine learning techniques. Machine learning techniques include Naïve Bayes, Support Vector Machine, AdaBoost, and Decision Tree etc. whereas Non-Machine Learning techniques are Black/White List, Signatures, Mail Header Checking etc. This paper intends to present the Comparative Analysis of performance of various pre-existing classification techniques.

Keywords- Classification, E-mail Threats, Spam Filtering, Efficiency.

I. INTRODUCTION

With the most preferred communication method e-mails have become part of regular life. Spams which are also called unwanted, junk, unsolicited mails are one of the considerable problems in utilizing the emails [2]. There are basically two things that are confused with each other - Paper Junk Mail and Spam Mail. Let's clear this concept that in the Paper Junk Mail Junk mailers pay for distribution of the material while in case of E-mail spamming the recipient has to pay in terms of bandwidth, disk space, server resources as well as lost productivity [2]. The issue of e-mail spamming can become a headache if not managed properly [1]. There are many issues that arise from the bombardment of the spam emails like filling up of the user's mailboxes, flooding important e-mails, wastage of memory along with bandwidth and time.

Email becomes the major source of communication these days. The email is mostly used by human for their personal or professional use as it is powerful, speedy and inexpensive way of communication. Globally, the number of email account is increasing day by day it is expected that the total number of email accounts will increase from 3.3 billion email accounts in 2012 to 4.3 billion by the end of year 2016[email statistic report 2012]. Now days, everyone in the world have an email account. The attention and usage for the email is growing day by day all over the world. It is an affordable means to smoothly transfer information worldwide with the help of internet.

Spam is an unwanted, junk, unsolicited bulk message which is used to spread virus, Trojans, malicious code, advertisement or to achieve profit on negligible cost. They are various kind of spams based on the way of transmission i.e. email spam, social networking spam, web spam, blogs or review platform spam, instant message spam, text message spam and comment spam. Spam message can contain text, image, video, voice data etc. Spam can be sent via web, fax, telephonically, sms (text messages) etc. [19].

As the use of Email is increasing day by day because of effective, fast and cheap way of exchanging information with each other and so is the problem of Email Spamming. According to the observation, it has been noted that a user receives more spam or irrelevant mails as compared to ham or relevant mails. The amount of sending spam mail per day is about 120 billion and the sending cost is approximately zero. According to a survey report the spam sending rate is 53.1 percent in December, 2015. Spam not only wastes the users time, energy, resources, storage space, computation power, and bandwidth but also irritates the user with enormous unwanted messages [19].For example, if you received 100 emails in a day out of which spam mails are 70 and ham includes only 30 emails. So, it takes time to identify the ham or important emails from it, which irritated the user. Email user receives hundreds of spam emails per day with a unique address or identification and new content which are generated automatically by robot software.

Email is a spam email if it meets the following criteria:

1. **Unsolicited email:** - The email which is not requested by recipient.

2. **Bulk mailing/mass mailing: -** The email send to large number of people are known as mass or bulk mailing.
3. **Nameless emails: -** The nameless emails are those email in which the identity and address of sender is not acknowledged.

The cost of the spam emails is billions of dollars per year to the Internet Service Provider because of the loss of bandwidth. Spam Emails deliberately creates problem for intended user, internet service provider and an entire internet backbone network. One of the examples is Denial of Service attacks(DOS) where the spammers send a large amount of emails to the server thus delaying relevant email to reach to the intended recipient. Spam is a major problem that attacks the existence of electronic mails. So, the researcher gives the major concern to distinguish ham emails from spam emails. To overcome this problem many methods have been proposed for classification of emails as spam or ham. Spam filters are basically used to detect unwanted, unsolicited, junk emails. The filter classification techniques are basically categorized into two parts:

1. Based on Machine Learning Technique.
2. Based on Non-Machine Learning Techniques.

The machine learning techniques include Naïve Bayes, Support Vector Machine, Neural Network, Decision Tree etc. whereas the Non- Machine Learning techniques include Heuristics, Black/White List, Signatures, Mail Header Checking etc. It is found that classification based on machine learning techniques has higher success ratio as compared to classification based on non-machine learning techniques.

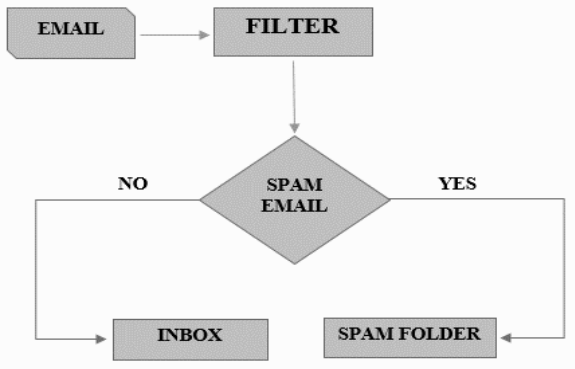

Fig 1. Flowchart of Spam Mail Filter

A. Spam Filtering

Spam filtering is a process that is used to detect unsolicited and unwanted emails and prevent them from getting to user's inbox. There are two levels at which the Spam filtering in the emails can operate that will involve a user level or an enterprise level. Individual Users refers to the single specific person that is working at home and who has been receiving and sending the e-mails via ISP, these users if wish to identify and filter the spam mails simply install the spam filtering system. In the Enterprise level spam filtering mails are filtered during entering time in the internal network of an Enterprise. In the Enterprise level spam filtering, spam filtering software is installed on the main mail server and it is meant to interact with the mail transfer agent (MTA) that classifies the message at the moment they are received [1]. Most of current spam sifting frameworks use principle based scoring systems. An arrangement of tenets is connected to a message and a score gathers in light of the guidelines that are valid for the message. Frameworks commonly incorporate several guidelines and these standards should be redesigned frequently so that spammers can't modify substance and conduct, so as to maintain a strategic distance from the channels. The engineering of spam separating is shown in Fig.2. Initially the model will collect the client messages which can be spam mail and non-spam mail. Then the underlying change procedure will start. The model states starting change, the user identification, highlight extraction, email information order, analyzer area. Machine learning techniques are used to classify the spam mails.

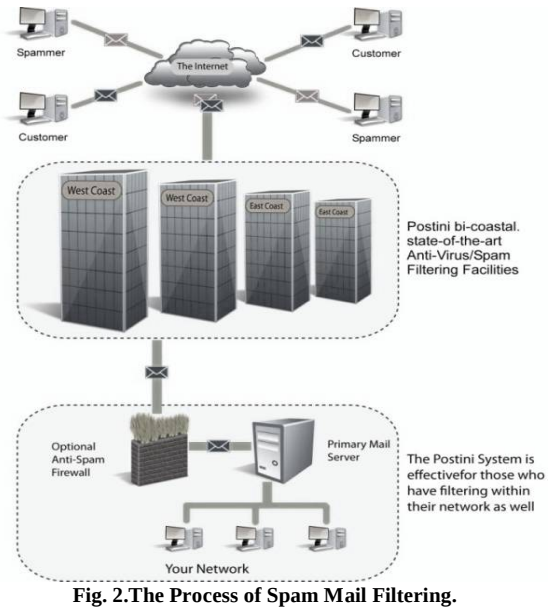

Fig. 2.The Process of Spam Mail Filtering.

B. Classifiers in Spam Mail Filtering

There are many types of classifiers that are meant for the purpose of classifying the e-mails as spam or hams and these are basically classified into two categories mainly those being: Content based classifiers and Non-content based classifiers.

1. Content based Classifiers -These classifiers are also famous by the name of hand crafted spam classifiers and these are the types in which the spams are categorized on the basis of the content it holds or information it stores. It checks for text in body of the Email, then URL. It also considers the mail header like subject for classification of text. It performs text classification task by employing

preprocessing on text in terms of HTML tags removal, Tokenizing and Word frequency calculation for determining word probability to find out whether a given mail is spam or not.

2. Non-Content based Classifiers - In this type of the classifier the automated filter is installed and in this the classification depends upon the human recipient. In this the classification occurs from the judgment of the sender's name, address etc.

This paper discusses about spam mails in section (1), In section (2) Literature survey of various classification techniques concept in spam filtering has been elaborated, In section (3) existing algorithms for classification are discussed and are compared in a tabulated form with respect to various parameters and lastly section (4) concludes the paper giving brief summary of the work.

II. RELATED WORK

Masurah Mohamad et al. (2015) In this paper, authors presented a hybrid feature selection method. In which they integrate the rough set theory and term frequency inverse document frequency (TF-IDF) to enhance the efficiency decisions in email filters. This paper also explain about the Feature Selection Methods such as Information Gain (IG), Gini Index, X2-Statistic, Fuzzy Adaptive Particle Swarm Optimization (FAPSO) and Term Frequency Inverse Document Frequency (TF-IDF)and Machine Learning Approaches such as Naïve Bayes and Rough set theory. They use header section and spam behaviors which are non-content based keywords. They use dataset corresponding of text messages and images. Then they explain their proposed spam filtering framework. In experimental work they show that rough set theory and TF-IDF were able to work together in order to achieve concise and more accurate results. But the combination of decision tree and TF-IDF gives the best accuracy among others i.e. 89.4%

IzzatAlsmadi et al. (2015)In this paper, the authors demonstrate various research papers based on spam detection, ontology classification on email content and other research ambition. The authors used general statistical data set about the email that can be provided by Google to Gmail account user. They distinguish the dataset in two methods such as based on Classification on WordNet class and second are Clustering and Classification evaluation. In second method they use k-Means algorithm and for classification they use Support Vector Machine. They also work to evaluate the SVM models. They evaluated three SVM models. In first case they analyze those Top 100 words-VS-emails before removing stop words, in second case Top 100 words-VS- email after removing stop words, and in last they evaluated that NGram terms-VS-email. They concluded that the True Positive(TP) rate is shown to be very high in each case but the False Positive (FP) rate is shown to be best in case of NGram based clustering and classification.

Savita Pundalik Teli et al. (2014) In this paper, the author study three different classification techniques KNN, Support Vector Machine and Naïve Bayes. In which, the authors exhibited that Naïve Bayes gives maximum accuracy as compared to other algorithms that is 94.2%. The authors also proposed a method to enhance the efficiency of Naïve Bayes. The proposed method is divided into three phases. In first phase the user creates rule for classification, second phase trains the classifier with training set by extracting the tokens, and in third phase based on maximum token matches, the email is classified as spam or ham. They also concluded that the accuracy of classifier algorithm is dependent on training phase. The performance of Naïve Bayes is improved by using this Algorithm.

Anirudh Harisinghaney et al. (2014) The objective of their work is to detect text as well as spam emails. For this purpose, they use Naïve Bayes, KNN and a new proposed method Reverse DBSCAN (Density-based spatial clustering of application with noise). They use enron cropus dataset of text as well as images. They extract words from image by using Google's open source library, Tasseract. In which, they use pre-processing of data and also illustrates that pre-processing gives 50% better accuracy results as compared to other three algorithms when they don't use pre-processing. At last they concluded that Naïve Bayes with pre-processing gives the best accuracy among other algorithms.

Rushdi Shams et al. (2013) In this paper, the authors exhibited a novel spam classification method based on features selection. This classification is based on email content language and readability merged with the previously used content based task features. The features are extracted from four benchmark datasets such as CSDMC2010, Spam Assassin, Ling Spam, and Enron-spam. They also divided the features in three categories i.e. traditional features, test features, and readability features. The proposed work is able to classify emails in any language because the features are language independent. They use five well-known machine learning algorithms to introduce spam classifier: Random Forest (RF),

Megha Rathi et al. (2013) In this paper the author exhibited the data mining techniques and also explained the classification algorithms. They evaluated various classification algorithms such as Naïve Bayes, Bayesian Net, Random Forest, Random Tree, SVM etc. without feature selection first. Then they evaluated all these classification algorithms with feature selection by best first algorithm. The author analyzed that the Random Tree has 90.43% accuracy, which is very low. But with feature selection it

reaches to 99.71% which is very high i.e. close to 100%. Therefore, they concluded that random tree is the best classification algorithm for email classification with feature selection.

D. KarthikaRenukaet al. (2011) In this paper, the authors compared three classification algorithms Naïve Bayes, J48 and Multilayer perceptron (MLP) classifier. They evaluated that MLP accuracy rate is higher among others but takes maximum time to classify. And Naïve Bayes takes minimum time but its accuracy is very less. They use filtered Bayesian Learning algorithm with Naïve Bayes to enhance the performance of Naïve Bayes. The FBL is used for feature selection. After using FBL, the accuracy rate of Naïve Bayes increased to 91%.

III. TYPES OF CLASSIFICATION ALGORITHMS

There are many algorithms that are designed for the purpose of email classification and some of them are discussed below:

3.1. Naive Bayes Algorithm

It is one of the famous machine learning algorithm working on the principle of Bayes theorem. Bayes theorem calculate the posterior probability. It is the technique that is widely used for the purpose of email classifications for spam and non-spam. It is defined as:

$$P(H/K) = P(K/H)\, P(H) / (P(K)) \quad \ldots (1)$$

where,

P (H/K) is the posterior probability of class(H) for given predicator(K).

P (K/H) is the likelihood which is probability of predicator for given class.

P (H) is the prior probability of class.

P (K) is the prior probability of the predicator.

Some common words are used in both spam and non-spam mails. It is not like that filters know the words previously, but there has to be a training process built up for them and after that these word probabilities are utilized for the purpose of email classification. In this case, each word or the most interesting words contribute to the email spamming. And there is a threshold that has been set for determining the spam and if the probability is increased above that threshold, then the email is considered as the spam. [9] [10][11]

3.2 Support Vector Machine Algorithm

SVM is a supervised machine learning technique which is used for both classification and regression. In this we plot each data item as a point in n-dimensional space where, n= number of features. In this technique original data is transformed into higher dimensionality and hence searches for the optimized hyperplane (decision boundary) which separates the tuples of one class from the other by an apparent gap that is as ample as possible.[20]

3.3 k-Nearest- Neighbor Algorithm

The k-Nearest Neighbor (kNN for short) is a non-parametric instance based learning technique or lazy learning. It is used for make decision based on complete training data set. The input consists of the k- closest data items in the feature space. The output is a class membership function. An object is classified by majority vote of its neighbors. The object will be assigned to the class which is most common among k- nearest Neighbors. [14]

3.4 Decision Tree Induction Algorithm

Decision tree consist of the root node, branches and leaf nodes. In this, the tree is created in a top-down, recursive and divide and conquer way. It works like a greedy technique. The internal node defines the condition on the attribute, each branch defines the output of the condition and each leaf node defines the class label. [15]

3.5 AdaBoost Classification Algorithm

Machine learning algorithm proposed by Freund and Robert Schapiro. It is a Meta algorithm which can be used in aggregation with some other learning algorithms to improve the performance of the algorithm. AdaBoost classifier uses Confidence based label sampling that works with the concept of active learning. Classifier is trained by the variance and obtains a scoring function which is used to classify the mail as spam or ham. The labeled data is used to train the data. The trained classifier generated the required functions which classify the message as spam. This algorithm improves training process.

3.6 Rule Based Classification Algorithm

In this algorithm classifier is represented as a set of IF-THEN rules. IF-THEN rule is of the form IF condition THEN conclusion. The “IF” part is known as rule antecedent and “THEN” part is known as rule consequent. The condition performs the test on one or more attributes. The class prediction is specified by rule consequent.

3.7 Backpropagation Algorithm

It is a neural network learning algorithm. It trains the feed forward multilayer neural network for given data samples. When each entry of the sample data item is presented to the network, the network checks the output response to the sample data item. The output response is then compared with known and desired output and error value is found out. Based on error value network weights are adjusted. The weights are adjusted by finding mean square error of output response with input sample. [16]

TABLE1: THEORATICAL FINDINGS OF CLASSIFICATION TECHNIQUES

Sr.no	Algorithm	Classification Principle	Findings	Disadvantages
1	Naive Bayes Algorithm	Works on Bayes Theorem.	It has high accuracy and speed when used for large data sets.	Assumption is made that events occurring are mutually exclusive.
2	Support Vector Machine Algorithm	Non- Linear Mapping.	Highly Efficient and accurate classifier. Less prone to overfitting.	Complex algorithm -difficult to understand. Training time is more.
3	k-Nearest-Neighbor Algorithm	Learning by analogy and distance based comparison.	Less work on training data sets but more work on classification.	Computationally expensive. Require efficient storage techniques.
4	Decision Trees Induction Algorithm	Top down, recursive, divide and conquer based.	Can handle high dimensional data. It is simple and fast and have good accuracy.	Branches may contain outliers in the training data sets.
5	AdaBoost algorithm	Based on error rate and boosting	powerful classifier that works well on both basic and more complex recognition problems	AdaBoost could be sensitive to noisy data
6	Rule Based Classification	Based on IF-THEN rules.	Rules are efficient technique for the representation of knowledge. Rules are specified by using coverage and accuracy.	What if more than one rule is fired specifying different classes. And if no rule is fired.
7	Classification by Backpropagation	Based on neural network learning algorithm.	Can deal with noisy data and have capability to classify data sets for which they are not trained.	Require more training times. Suffers from Poor interpretability.

In this table, classification principles of each classification technique are highlighted with their advantages and disadvantages. In modern era, there is a need to use feature selection technique to reduce training time and ensemble based techniques i.e. Bagging and Boosting to improve the accuracy. So, we need to combine feature selection algorithm with ensemble based techniques to achieve high performance.

CONCLUSIONS

Efficiency of spam mail filtering is depending on classification algorithm used. In this paper, a number of existing algorithms for spam mail filtering are discussed, compared with each other and tabulated with their findings. It helps to understand the wide variety of classification techniques in order to select one.

ACKNOWLEDGMENTS

The paper has been composed with kind assistance, guidance and support of my department who have helped me in this work. We would like to thank all the people whose encouragement and support has made the fulfillment of this work conceivable.

REFERENCES

[1] Omar Saad, Ashraf Darwish and Ramadan Faraj: "Asurvey of machine learning techniques for Spam filtering",IJCSNS ,International Journal of Computer Science andNetwork Security, VOL.12 No.2, February 2012.

[2] MasurahMohamad and Ali Selamat, "An evaluation on the efficiency of hybrid feature selection in spam email classification," IEEE International Conference on Computer Communication, and Control Technology (14CT 2015), pp. 657 – 666, 2015.

[3] IzzatAlsmadi and IkdamAlhami, "Clustering and Classification of email contents," Journal of King Saud University-Computer and Information Sciences, pp. 46-57, 2015.

[4] SavitaPundalikTeli and Santosh Kumar Biradar, "Effective Email Classification for Spam and Non-spam," International Journal of Advanced Research in Computer and software Engineering, vol. 4, pp.273-278, 2014.

[5] AnirudhHarisinghaney, Aman Dixit, Saurabh Gupta, and AnujaArora, "Text and image based spam email classification using KNN, Naïve Bayes and reverse DBSCAN Algorithm," International Conference on Reliability, Optimization and Information Technology - ICROIT 2014, pp. 153-155, 2014.

[6] Rushdi Shams and Robert E. Mercer, "Classification spam emails using text and readability features," IEEE 13th International Conference on Data Mining, pp. 657-666, 2013.

[7] MeghaRathi and VikasPareek, "Spam Email Detection through Data Mining-A Comparative Performance Analysis," I.J. Modern Education and Computer Science, vol. 12, pp. 31-39, 2013.

[8] Ms.D.KarthikaRenuka, Dr.T.Hamsapriya, Mr. M.Raja Chakkaravarthi,Ms.P.Lakshmisurya,"Spam Classification based on Supervised Learning using Machine Learning Techniques," IEEE, pp.1-7, 2011.

[9] I. Androutsopoulos, J. Koutsias, "An evaluation of naïveBayesian anti-spam filtering", 11thEuropean Conference on Machine Learning (ECML 2000),pp9–17, 2000.

[10] Androutsopoulos, G. Paliouras, "Learning to filter spam E-mail: A comparison of a naïve Bayesian and a memory based approach", 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp 1–13, 2000.

[11] K. Schneider, "A comparison of event models for naive bayes anti-spam e-mail filtering", 10th Conference of the European Chapter of the Association for Computational Linguistics, pp.307-314, 2003.

[12] N. Cristianini, B. Schoelkopf, "Support vector machines and kernel methods, the new generation of learning machines". Artificial Intelligence Magazine, pp31–41, 2002

[13] S. Amari, S. Wu, "Improving support vector machine classifiers by modifying kernel functions". Neural Networks, pp 783–789, 1999.
[14] C. Miller, "Neural Network-based Antispam Heuristics", Symantec Enterprise Security (2011), www.symantec.com Retrieved December 28, 2011
[15] AnirudhHarisinghaney, Aman Dixit, Saurabh Gupta, and AnujaArora , "Text and image based spam email classification using KNN, Naïve Bayes and reverse DBSCAN Algorithm, " ICROIT 2014, India, Feb 6-8 2014.
[16] MasurahMohamad and Ali Selamat, "An evaluation on the efficiency of hybrid feature selection in spam email classification," IEEE International Conference on Computer Communication, and Control Technology (14CT 2015), April. 2015.
[17] Rushdi Shams and Robert E. Mercer, "Classification spam emails using text and readability features," IEEE 13th International Conference on Data Mining, pp. 657-666, 2013.
[18] MeghaRathi and VikasPareek, "Spam Email Detection through Data Mining-A Comparative Performance Analysis," I.J. Modern Education and Computer Science, pp. 31-39, 2013.
[19] G. Kaur and R. K. Gurm, "A Survey on Various Classification Techniques in Email Spamming," International Journal of Technology and Computing (IJTC) vol. 5, no. 3, pp. 589–593, 2016.
[20] Esha Bansal and Anupam Bhatia, " Support Vector Machine for Multiclass Handwritten Digits" International conference on Advanced Information Communication Technology in Engineering (ICAICTE- 2K13), pp. 239-241.

★★★