I.J. Modern Education and Computer Science, **2017, 9, 42-52**
Published Online September 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2017.09.05

E-Mail Spam Detection Using Refined MLP with Feature Selection

Harjot Kaur
CT Group of Institution/CSE, Jalandhar, 144041, India
Email: Harjotkaur844@gmail.com

Er. Prince Verma
CT Group of Institution/CSE, Jalandhar, 144041, India
Email: prince.researchwork@gmail.com

Received: 28 March 2016; Accepted: 22 August 2017; Published: 08 September 2017

***Abstract*—**Electronic Mail (E-mail) has established a significant place in information user's life. E-Mails are used as a major and important mode of information sharing because emails are faster and effective way of communication. Email plays its important role of communication in both personal and professional aspects of one's life. The rapid increase in the number of account holders from last few decades and the increase in the volume of emails have generated various serious issues too. Emails are categorised into ham and spam emails. From past decades spam emails are spreading at a tremendous rate. These spam emails are illegitimate and unwanted emails that may contain junk, viruses, malicious codes, advertisements or threat messages to the authenticated account holders. This serious issue has generated a need for efficient and effective anti-spam filters that filter the email into spam or ham email. Spam filters prevent the spam emails from getting into user's inbox. Email spam filters can filter emails on content base or on header base. Various spam filters are labelled into two categories learning and non-machine learning techniques. This paper will discuss the process of filtering the emails into spam and ham using various techniques.

***Index Terms*—**Data Mining, Knowledge Discovery (KDD) Process, E-Mail, Spam, Ham, Spam Filter, N-Gram based feature selection, Multi-Layer Perceptron Neural Network (MLP-NN) and Support Vector Machine (SVM) classification algorithms.

I. Introduction

The generation in the growth of data from past few decades is increasing tremendously. Various sources like commercial sites, engineering field, Facebook and other social links like Twitter, Youtube contributes to the size and complexity of data. To handle and to extract relevance among the data, various tool and techniques are available that ensure relevant extraction of data from irrelevant content. Data Mining is a process used for extracting hidden and unknown information from the databases for seeking knowledge. Data can vary in size, complexity to structure. Data can be in the form of audio, video or simply a text data. To handle and to extract the desirable properties from the data, mining is carried out.

A. Knowledge Discovery Process

Knowledge form data can be achieved by undergoing various steps as mentioned below. Data mining term is also labelled as Knowledge discovery technique, which means a procedure of extracting useful information from a set of raw data. Data mining is a part of knowledge discovery [13], [14].

- *Collection of Raw Data*: Dataset can be collected from various sources like online and offline, social media sources, banks, retail sector etc.
- *Data Selection:* Relevant data that is of use is selected for analysis.
- *Data Pre-Processing:* Cleansing of the data to remove any sort of noise, bogus or missing value from the data is carried out.
- *Transformation:* The data is transformed into appropriate form so that mining operation can be carried out.
- *Data Mining:* Extraction of relevant patterns from the data by using various data mining techniques.
- *Evaluation:* Extracted patterns are analysed for the truth-worthiness of the patterns and its relevance.
- *Knowledge:* The above procedure mines the relevant knowledge from the raw dataset. Knowledge can be represented by various techniques.

B. Various Techniques of Data Mining

The various techniques followed in data mining are classified as mentioned below [13]. The following steps are executed on irrelevant data to gain and access relevant information.

- *Anomaly Detection:* Data information that is bogus or irrelevant is mined. Anomaly detection detects the information without any facts.

- *Association Rule Mining (ARM):* It is a process of identifying the relationship among the attributes present in the datasets.
- *Clustering:* It is a process that groups the similar data in one cluster without using any predefined model. Clustering defines its own model and is a descriptive procedure of grouping the data.

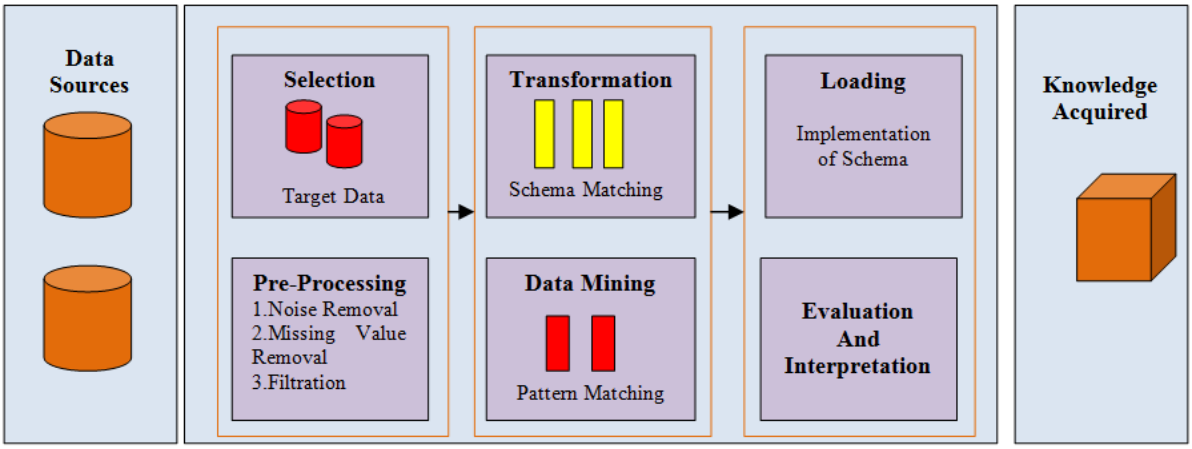

Fig.1. Knowledge Discovery Process in data mining.

- *Classification:* It is a process that has a predefined model and generalises the data into known predetermined classes. Classification is a predictive model.
- *Summarization:* A process of representing the data in a precise form for visualisation.

C. Email Spam

The Internet has become an important and essential part of human life. The increase in the utilisation of internet has increased the number of account holders over various social sites. Email is the simplest and fastest mode of communication over the internet that is used both personally and professionally. Due to the increase in the number of account holders and an increase in the rate of transmission of emails a serious issue of spam emails had aroused. From a survey, it was analysed that over 294 billion emails are sent and received every day. Over 90% emails are reported to be spam emails [15].

Emails are labelled into two categories Spam emails and Ham emails. Spam emails are the junk emails received from illegitimate users that might contain advertisement, malicious code, Virus or to gain personal profit from the user. Spam can be transmitted from any source like Web, Text messages, Fax etc., depending upon the mode of transmission spam can be categorised into various categories like email spam, web spam, text spam, social networking spam [3].

The rate at which email spamming is spreading is increasing tremendously because of the fast and immodest way of sharing information. It was reported that user receives more spam emails than ham emails. Spam filtration is important because of spam waste time, energy, bandwidth, storage and consume other resources [12].

Email can be categorised as a spam email if it shows following characteristics:

- *Unsolicited Email:* Email received from unknown contact or illegitimate contact.
- *Bulk Mailing:* The type of email which is sent in bulk to many users.
- *Nameless Mails:* The type of emails in which the identity of the user is not shown or is hidden.

Spamming is a major issue and causes serious loss of bandwidth and cost billion of dollars to the service providers. It is essential for distinguishing between the spam mail and ham mail. Many algorithms are so far used to successfully characterise the emails on their behaviour but because of the changing technologies, hackers are becoming more intelligent. So, better algorithms with high accuracy are needed that successfully label an email as spam or ham Email. Spam filter technique is used to label the email as a junk and unwanted email and prevents it from entering the authenticated account holder's inbox. Filters can be grouped into two categories [12]:

- *Machine Learning Based Technique*: These techniques are Support Vector Machine, Multi-Layer Perceptron, Naïve Bayes Algorithm, Decision Tree Based etc.
- *Non-Machine Learning Based Technique:* These techniques are signature based, heuristic scanning, blacklist/whitelist, sandboxing and mail header scanning etc.

The success ratio of machine learning algorithms over non-machine learning algorithms is more. These techniques work by selecting the best features from the

data to group the emails as spam or ham. Feature selection can be carried out in two ways:

- *Header Based Selection:* Selecting the best feature from the header of the mail. It contains sender's address, BCC (Blind Carbon Copy), CC (Carbon Copy), To, From, Date and Subject.
- *Content Based Selection:* Selecting the best feature from the content in the mail. It contains the main message either in the form of text, audio or video, attachments etc.

Content Based Feature Selection is proven as the most authenticated feature selection as compared to Header Based as Header Based Feature Selection can be easily tempered by the hackers or spammers.

The survey paper is outlined in different sections. Section 2 represents the related work introducing various papers in the field of email spam detection. Section 3 describes the best feature selection technique to label the email as spam email or ham email by using n-gram based feature selection technique. Section 4 and 5 describes various machine and non-machine learning algorithms for classification of emails and Section 6 concludes the paper and prescribes the best algorithm with high accuracy for spam detection. Section 7 illustrates the conclusion and future steps to be taken to boost the performance of the algorithms.

II. Related Work

This section describes various papers about the work carried so far for detection of spam emails.

Bo Yu and Zong-ben Xu (2008) performed a comparative analysis on content-based spam classification using four different machine learning algorithms. This paper classified spam emails using four different machine learning algorithms viz. Naıve Bayesian, Neural Network, Support Vector Machine and Relevance Vector Machine. The analysis was performed on the different training dataset and feature selection. Analysis results demonstrated that NN algorithm is no good enough algorithm to be used as a tool for spam rejection. SVM and RVM machine learning algorithms are better algorithms than NB classifier. Instead of slow learning, RVM is still better algorithm than SVM for spam classification with less execution time and fewer relevance vectors [1].

Tiago A. Almeida and Akebo Yamakami (2010) performed a comparative analysis using content-based filtering for spam. This paper discussed seven different modified versions of Naïve Bayes Classifier and compared those results with Linear Support Vector Machine on six different open and large datasets. The results demonstrated that SVM, Boolean NB and Basic NB are the best algorithms for spam detection. However, SVM executed the accuracy rate higher than 90% for almost all the datasets utilised [2].

Loredana Firte, Camelia Lemnaru and Rodica Potolea (2010) performed a comparative analysis on spam detection filter using KNN Algorithm and Resampling approach. This paper makes use of the K-NN algorithm for classification of spam emails on the predefined dataset using feature's selected from the content and emails properties. Resampling of the datasets to appropriate set and positive distribution was carried out to make the algorithm efficient for feature selection [3].

Ms.D.Karthika Renuka, Dr.T.Hamsapriya, Mr.M.Raja Chakkaravarthi and Ms.P.Lakshmisurya (2011) performed a comparative analysis of spam classification based on supervised learning using several machine learning techniques. In this analysis, the comparison was done using three different machine learning classification algorithms viz. Naïve Bayes, J48 and Multilayer perceptron (MLP) classifier. Results demonstrated high accuracy for MLP but high time consumption. While Naïve Bayes accuracy was low than MLP but was fast enough in execution and learning. The accuracy of Naïve Bayes was enhanced using FBL feature selection and used filtered Bayesian Learning with Naïve Bayes. The modified Naïve Bayes showed the accuracy of 91% [4].

Rushdi Shams and Robert E. Mercer (2013) performed a comparative analysis of the classification of spam emails by using text and readability features. This paper proposed an efficient spam classification method along with feature selection using the content of emails and readability. This paper used four data sets such as CSDMC2010, Spam Assassin, Ling-Spam, and Enron-spam. Features are categorised into three categories i.e. traditional features, test features and readability features. The proposed approach is able to classify emails of any language because the features are kept independent of the languages. This paper used five classification based algorithms for spam detection viz. Random Forest (RF), Bagging, Adaboostm 1, Support Vector Machine (SVM) and Naïve Bayes (NB). Results comparison among different classifiers predicted Bagging algorithm to be the best for spam detection [5].

Megha Rathi and Vikas Pareek(2013) performed an analysis on spam email detection through Data Mining by performing analysis on classifiers by selecting and without selecting the features [6].

Anirudh Harisinghaney, Aman Dixit, Saurabh Gupta and Anuja Arora (2014) performed a comparative analysis of text and images by using KNN, Naïve Bayes and Reverse-DBSCAN Algorithm for email spam detection. This analysis paper proposed a methodology for detecting text and spam emails. They used Naïve Bayes, K-NN and a modified Reverse DBSCAN (Density- Based Spatial Clustering of Application with Noise) algorithm. Authors used Enron dataset for text and image spam classification. They used Google's open source library, Tesseract for extracting words from images. Results show that these three machine learning algorithms give better results without pre-processing among which Naïve Bayes algorithm is highly accurate than other algorithms [7].

Savita Pundalik Teli and Santosh Kumar Biradar (2014) performed an analysis of effective email classification for spam and non-spam emails [8].

Izzat Alsmadi and Ikdam Alhami (2015) performed an analysis on clustering and classification of email contents for the detection of spam. This paper collected a large dataset of personal emails for the spam detection of emails based on folder and subject classification. Supervised approach viz. classification alongside unsupervised approach viz. clustering was performed on the personal data set. This paper used SVM classification algorithm for classifying the data obtained from K-means clustering algorithm. This paper performed three types of classification viz. without removing stop words, removing stop words and using N-gram based classification. The results clearly illustrated that N-gram based classification for spam detection is the best approach for large and Bi-language text [9].

Ali Shafigh Aski and Navid Khalilzadeh Sourati (2016) performed an analysis using Machine Learning". This paper utilised three machine learning algorithms viz. Multi-Layer Neural Network, J48 and Naïve Bayes Classifier for detection of spam emails from ham emails using 23 rules. The model demonstrated high accuracy in case of MLP with high time for execution while Naïve Bayes showed slightly less accuracy than MLP and also low execution time [10].

III. N-Gram-Based Feature Selection

Feature selection is a dimensionality reduction method which is used for better classification results by selecting the most desirable feature from the pre-processed data. In this survey N-Gram based feature selection is discussed. N-Gram is a prediction based algorithm used for predicting the chance of occurrence of next word after making observations of N-1 words in a sentence or text corpus. N-Grams use probability-based methods for the prediction of next word. N-Gram is used in text mining and natural language processing. N-grams are the group of co-occurring words that move one or X (number of words in a corpus) steps or words ahead while computing N-Grams.

Let X, be the number of words in a given text corpus T, the number of N-Grams can be calculated by:

$$N_{Gram}(T) = X - (N - 1) \quad (1)$$

N-Grams are collected from a text corpus and vary according to the size of N. In the above equation for the calculation of N-Gram, T represents the text, X represents the number of words in the text corpus, and N represents the size of the text.

- *Uni-Gram:* The N-Gram of size one is termed as Uni-Gram. For example, the word "FOOD" in Uni-Gram can be represented by moving one step ahead viz. "F to O", "O to O", "O to D".
- *Bi-Gram or Di-Gram:* The N-Gram of size two is termed as Di-Gram. For example, "FOOD" in Bi-Gram can be represented by moving two steps ahead in the string of data viz. "FO to OO", "OO to OD".
- *Tri-Gram:* The N-Gram of size three is termed as Tri-Gram and so on for N= 4, N=5 etc.

N-Gram for a text corpus "Children are enjoying the sunny weather" using Bi-Gram (N=2) will be

"Children are"
"Are enjoying"
"Enjoying the"
"The sunny"
"Sunny weather"

In the example of text corpus containing 6 N-grams were listed as we move two steps ahead for generating the possibility of occurrence of next word viz. "Children are", "Are enjoying", "Enjoying the", "The Sunny" and "Sunny weather".

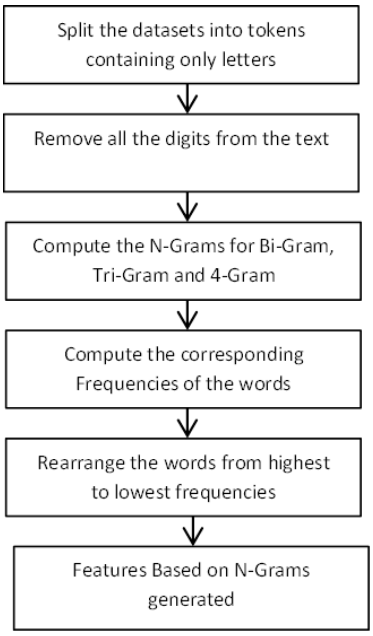

Fig.2. N-Gram Process Diagram.

N-Gram has a wide range of utilisation in various application like spelling correction, plagiarism detection, summarization of words, feature selection, breaking the words in text etc. N-grams are the N number of characters in a text or string and are widely used in various domains because of its ability to deal with noisy ASCII inputs and low error rate. The probability of N-Grams can be calculated by:

- *Simple N-Gram:* Simple N-Gram is a non-smoothed N-Gram that predicts the probability of next word by labelling equal probability to all the number of words present in the corpus. Suppose if there are N words in a text their probability would be 1/N without taking into consideration about the frequency of occurrence of words.
- *Markov Assumption:* The probability of occurrence of next word depends only on the previous word and this model was labelled as Markov Assumption. Markov Assumption calculates the probability by only considering the last word without taking into

consideration about the history of words that occurred in the past. Markov assumption (Bi-Gram) can be upgraded to tri-gram by considering last 2 words in the past and bi-gram can be upgraded to N-Gram by looking N-1 words in the past history. The simplest way to calculate the probability of occurrence of next word is by using Maximum Likelihood Estimation (MLE) which takes counts from the text and normalises the counts to lie in the range interval of [0,1].

- *Smoothing:* The limitation of MLE is that it shows a poor result for the words with low frequencies and for zero probability some N-Gram evaluation metrics does not work. Therefore, smoothing is used along with MLE for making N-Gram efficient for those sequences of words with low frequency by borrowing the probability from higher frequencies.

A. Advantages of N-Gram

- Words alone cannot provide information but using N-Grams provides an informative combination of words which help in an easy understanding of the meaning of the text.
- N-Grams can automatically capture the frequencies of words in the text that are repeated usually.
- N-Gram is independent of the language used in the document. Also, N-Gram can efficiently work with languages like Chinese and Urdu, where the words are not properly distinguished by borders.
- N-Grams do not require any initial partitioning of text into bag-of-words.
- N-Gram is highly tolerable towards any kind of words or spelling mistakes. For example, if a word "Table" is written as "Talbe" N-gram can easily recognise the correct existence of the word "Table".
- N-Gram effectively considers words and its ordering too.
- Learning rate is fast in N-Gram as compared to other feature selection techniques.

B. Dis-Advantages of N-Gram

- History of the text corpus is not taken into account. N-Gram calculates the probability of the occurrence of next words depending only on the word occurred in the past.
- Metrics Evaluation for low-frequency and zero-probability of the data is not possible which affects the efficiency of the N-Gram model.

IV. Non-Machine Learning Technique

Various algorithms are present for detecting spam in the email system. Some of the algorithms for spam detection can be based on machine learning and some are based on non-machine learning as defined below:

Non-Machine Learning is a technique of establishing a relationship among the variables using some self-proclaimed rules without relying on the data for knowledge. Non-machine learning is a non-efficient technique for spam filtration and detection. Various Non-Machine Learning algorithms can be categorised [11], [12].

- *Signatures:* Signatures contains the information taken from the documents. Signatures detect the spam or threats by generating a unique value called a hash value for each spam message. Signatures can be generated in two ways firstly by fragmenting the words into pairs and secondly by random generation of numbers. Signature uses the hash value with the new email value to compare and to analyse if the email is spam or ham.
- *Blacklist and Whitelist:* A blacklist is a list of spammers or any illegitimate contact that tries to send a spam or malicious email while whitelist is a list that contains legitimate users or contacts that are known to an individual account holder.
- *Heuristic Scanning:* This technique uses rules to detect malicious contents and threats. Heuristic scanning is a faster and efficient technique that detects the spam or threats without executing the file and works by understanding the behaviour. Heuristic scanning allows the user to change the rules.
- *Mail Header Checking***:** In this technique set of rules are developed that are matched with the email header to detect if the email is spam or ham. If the header of the email matches the rules, then it invokes the server and directs the emails that contain an empty field of "From", confliction in "To", confliction in "Subject" etc.

Table 1. Comparison of Different Non- Machine Learning Techniques [12].

Techniques	Advantages	Disadvantages
Blacklist/Whitelist	Simple	Can be easily attacked by spammers
Signatures	Low rate of False positive	Unable to detect a spam until unique hash is not distributed.
Heuristic Scanning	Users can change rules and is fast technique	Information limited to threats name.
Header Checking	Easiest technique	Low accuracy for spam detection.

V. Machine Learning Technique

Machine Learning Techniques enables the computer to learn by itself without being programmed. Machine learning algorithms are more efficient in contrast to those of non-machine learning. Machine learning work in a similar way like data mining, both acquire knowledge from data and find relevance in the data. Machine learning algorithms can be categorised into supervised and unsupervised algorithms. Some of the supervised

machine learning algorithms for classification of data is listed below:

A. Multi-Layer Perceptron Neural Network

Artificial neural networks are the part of artificial intelligence and MLP is a type of neural network. Multi-Layer Perceptron (MLP) is a feed-forward network that maps the group of inputs to their corresponding outputs. Fig. 2 demonstrates a feed-forward multilayer perceptron neural network [10], [18]. MLP is made up of simple neurons termed as a perceptron. Neural network generates information by enabling input perceptron's consisting the values labelled on them. The activation function of neurons is calculated by the formula mentioned below in the output layer [10]:

$$\alpha_i = \sigma(\sum_j W_{ij} O_j) \quad (2)$$

Where α_i represent the level of activation for i^{th} neurons; j is the set of neurons of the previous layer; W_{ij} is the weight of the connection between neurons i and j; O_j represents the output of j^{th} neuron and $\sigma(x)$ is the transfer function.

$$\sigma(x) = \frac{1}{1+e^x} \quad (3)$$

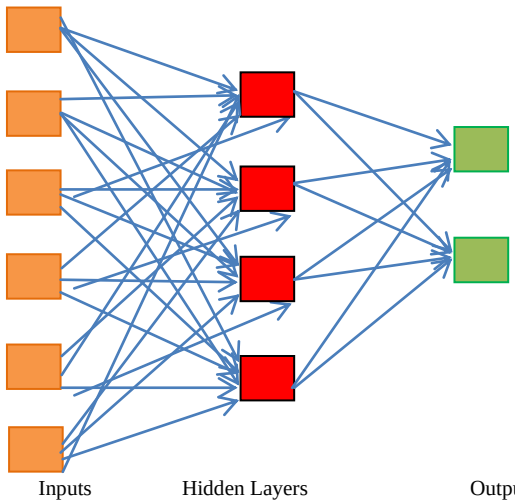

Fig.3. Multi-Layer Neural Network Architecture.

A back-propagation strategy based on delta rule is used to train multilayer perceptron network. MLP consist of various neurons divided into various layers, as follows:

- *Input Layer*: This layer generates the input for the network. The number of neurons depends upon the number of input given to the network.
- *Hidden Layer:* The layer that maps the input to the corresponding output is named as a hidden layer. Hidden layers vary in number.
- *Output Layer:* The layer from where the resultant can be seen. The number of neurons in output layer depends upon the learning of the kind of problem.

The type of relationship between the input and output vectors in MLP is a non-linear relationship. This is done by interconnecting the neurons in the antecedent and succeeding layers. Outputs are achieved by multiplying them with weight coefficients. In the training phase, the neural network is given the information regarding training only. Later on, the weights of the network are tuned between [-0.5, 0.5] to minimise the error rate between the expected and observed outputs and to enhance the frequency of training to a predicted level. A sequence untrained inputs are applied to the input to formalise the training. These input set are different from the inputs that are used for the training of the network. Training of the neural network is highly complex due to a large number of variables. MLP holds lots of advantages over other algorithms even if a correct relationship is not induced between the input and output or if the essential and exact information is not achieved [10]. Non-Linear Activation Function of MLP makes MLP different from other networks. Algorithmic steps for MLP- Neural Network can be modelled as:

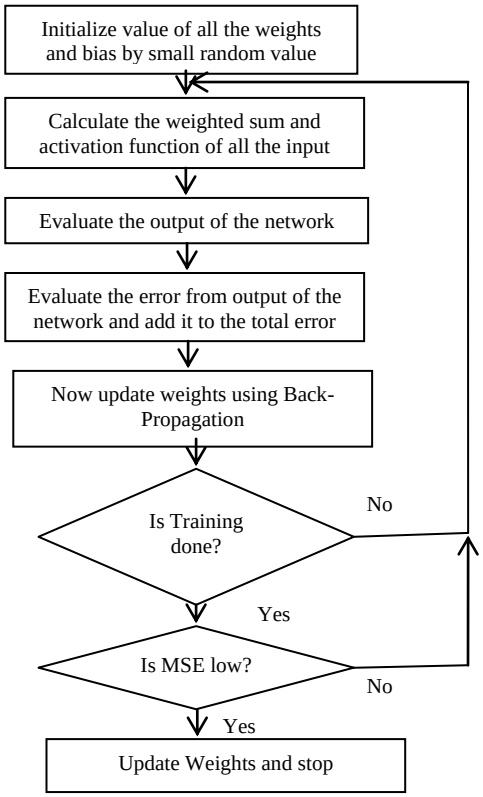

Fig.4. Multi-Layer Neural Network Process diagram.

Let a dataset D, consist of training samples and their target values, L be the rate of learning by the network to generate a trained network:

1. Initialize the weights and the biases of the layers using small random values.
2. Compute the weighted sum of the inputs, where

$$O_j = I_j \quad (4)$$

The output of the inputs is the true input values.

3. Compute the activation functions of the hidden layers, where

$$I_j = \sum_j W_{ij} O_j + \theta_j \quad (5)$$

Compute the net input of j with respect to 'i' (previous layer).

4. Compute the output of the layers, where

$$O_j = \frac{1}{1 + e_j^{-1}} \quad (6)$$

5. Compute the error rate by Back-Propagation, Error for output layer:

$$Error_j = O_j(1 - O_j)(T_j - O_j) \quad (7)$$

Error calculation of next hidden layer, h:

$$Error_j = O_j(1 - O_j)\sum_h Error_h W_{jh} \quad (8)$$

Weight update:

$$w_{ij} = w_{ij} + \Delta w_{ij} \quad (9)$$

Bias update:

$$\theta_j = \theta_j + \Delta\theta_j \quad (10)$$

Where Δw_{ij} and $\Delta\theta_j$ are the change in weight and bias.

B. *Support Vector Machine*

Support vector machine (SVM) is a supervised approach for machine learning. The main idea used in SVM is constructing a hyperplane that is optimal for the classification of patterns that can be linearly separated [20]. This algorithm work by plotting each information point in an n-dimensional workspace, where n represents the number of features which are equal to the coordinates in the workspace. The optimal hyper -plane differentiates the classes at this point [21].

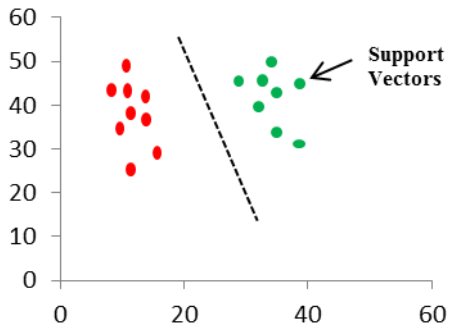

Fig.5. SVM representing the difference between two classes using hyper-plane.

In email spam detection, the aim is to divide the email into two categories spam or ham email by using an optimal hyperplane. The idea is to distinguish the two classes to achieve maximum marginal difference between two classes, viz. spam and ham. SVM represents the information points in the workspace, mapped so that the information points of the other categories are partitioned by a maximum marginal difference. New information points are labelled to that same workspace and predictions are conducted to analyse the category of the new information point. SVM can efficiently perform non-linear classification by kernel trick (similarity function). Algorithmic steps of SVM for the classification process are as follows [19].

1. Train the initial SVM using all the training data to have support vectors decision functions.
2. Eliminate those support vectors generated from the training of initial SVM whose projections have greatest curvatures on the hypersurface by: finding the projection of the support vectors along the gradient of decision function used, calculate the notion of curvature for every support vector on the hyperplane, lastly sort the support vectors in the decreasing order and deduct the top N-percentage of the vectors of support.
3. Retrain the SVM by left over vectors for best decision.
4. Use the group of information point to finally train the SVM, generating support vectors.

SVM classifiers are grouped into linear and non-linear classifiers, as follows:

- *Linear Classifiers:* Separating the data points in linear order by using a hyperplane is classified as linear classifiers. There are different hyperplanes but the best way to separate the data using hyperplane is by maximum margin difference viz. the distance of hyperplane and the closest information point of any class.
- *Non-Linear Classifiers:* Sometimes the data is not separated properly or linearly in the high dimensional plane for such separation non-linear classifiers are used that correctly classify the information points and label them to their exact class by using kernel tricks. Some mostly used kernel tricks are as follows:

1. *Homogenous kernels:* Polynomial kernels that are used for analysing the similarity of vectors are represented by the expression below:

$$k(\vec{\alpha_i}, \vec{\alpha_j}) = (\vec{\alpha_i}, \vec{\alpha_j})^d \quad (11)$$

Where k is the kernel function and $(\vec{\alpha_i}, \vec{\alpha_j})$ are the vectors of the workspace with d as the degree of the polynomial.

2. *Non-Homogenous kernels:* In Non-homogenous kernels, a free parameter is added that leverage the group of features combined together.

$$k(a,b) = (a^T b + c)^d \quad (12)$$

VI. Results and Discussion

Results are evaluated on Enron dataset with 6000 emails having 50% spam and ham emails collected from UCI repository. Results comparison ensures which algorithm can correctly remove vague information from the dataset. It was analysed that MLP classification algorithm can correctly classify instances and remove vague information but a randomised approach was followed for removing vague information. Randomization of the MLP algorithm means it start selecting the initial information by starting arbitrarily from any initial points; this degrades the performance of the algorithm and leads to increase in the model building time of the algorithm. MLP serves several advantages over SVM such as fault tolerance and generalisation. In the future research, effective steps to solve the randomization problem of the MLP algorithm will be considered. For feature selection, n-gram based feature selection technique is implemented on both classification algorithms. Initially pre-processing of the dataset was done to reduce the dimensionality of the dataset. Dimensionality reduction allows the algorithms to efficiently work on large datasets. By n-gram based feature selection 100 best features are considered. Results of n-gram with both the classification algorithms are labelled in Table 3.

Table 2. Comparison of SVM and MLP classification algorithm on Enron Dataset.

Percentage Split	Parameters / Algorithms	SVM	MLP
66%	**Correctly Classified Instances (Accuracy)**	64.66 %	78.09%
66%	**In-Correctly Classified Instances**	35.34%	21.91%
66%	**Sensitivity**	0.65	0.781
66%	**Specificity**	0.489	0.786
66%	**Precision**	0.722	0.789
66%	**F-Measure**	0.563	0.783
66%	**Root Mean Square Error (RMSE)**	0.594	0.386

Pre-processing of the dataset, eliminates the bogus, missing and incomplete values from the dataset. Secondly, Enron-dataset is a text dataset so it is essential to convert text corpus into words, so as to avoid the performance degradation of the algorithms. The main issue with the MLP classification algorithm is the randomization of the algorithm, that makes the algorithm highly time-consuming and degrades the performance measure of the MLP classification algorithm. In the research work, the main focus is to uplift the performance the MLP algorithm for the detection of email spamming and to remove any kind of vague information by avoiding the randomization of the classification algorithm.

In Table 2, the pre-processing results can clarify that MLP classification algorithm is a better approach for the detection of spam emails with high accuracy of 78.09%. Sensitivity and Specificity rate for the detection of the spam and ham emails for the MLP is high as considered to the SVM technique with 0.781and 0.786 rate respectively.

MLP algorithm shows better accuracy over SVM algorithm for Enron dataset. Fig. 6 illustrates the comparison between both the classification algorithms, where SVM correctly classified 64.6643% instances with 35.3357 % incorrectly labelled instances while MLP performs better with 78.091% accuracy for correctly labelling the instances and 21.9081% for incorrectly labelled instances. Fig. 7 shows the comparison for root mean square error, for an algorithm it is desirable to have a low root mean square error. In this case, MLP showed a low error rate of 0.3867 while SVM demonstrated 0.5944 error rate.

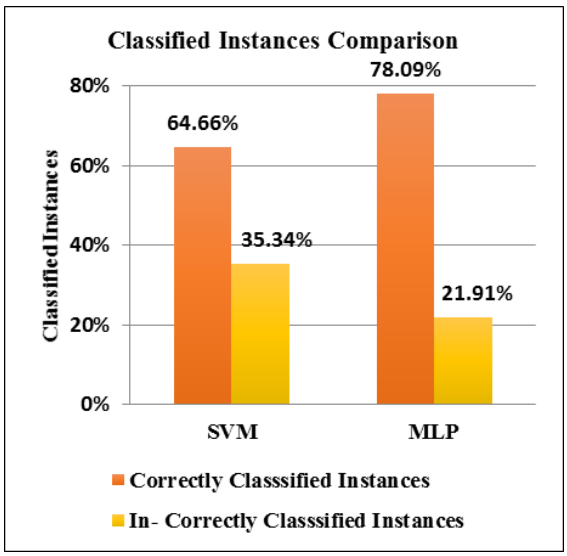

Fig.6. Classified instances comparison for SVM and MLP Algorithms.

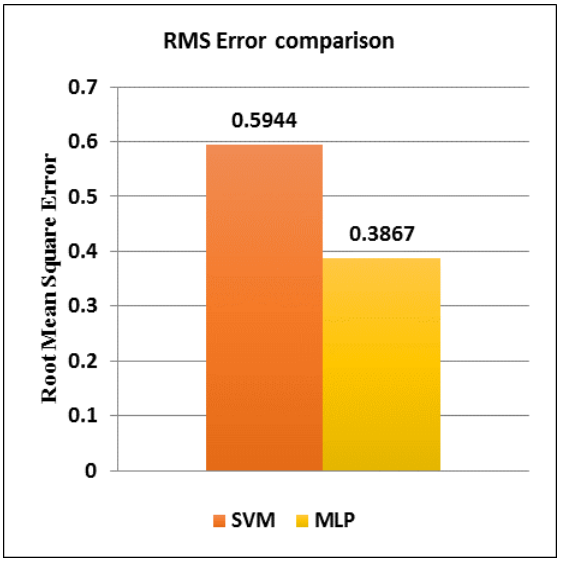

Fig. 7: RMSE comparison for SVM and MLP Algorithms.

In Table 3, N-gram based feature selection for choosing the best features from the dataset is performed by selecting 100 best and optimal features. Result comparison is performed by choosing two consecutive words for Bi-gram, three words for Trigram and four words for Four-gram techniques. After pre-processing and feature selection still MLP guarantees better results over SVM, because of the limitation of the SVM algorithm to choose information point because of kernel trick it fails to accommodate all the information point lying far from density function.

In Table 3, we can check the fluctuations in the values of the MLP and SVM algorithm, the values are increasing and decreasing for Bi-gram, Tri-gram and Four-gram. The main reason for the fluctuations in the values of the N-Gram is due to the problem of randomization of the MLP classification algorithm. The main work of the N-Gram based feature selection technique is to select the best features from the text dataset. After selecting the best 100 features from the Enron dataset the improvement in the algorithms is by 1% only. The randomization problem of the MLP algorithm still prevails to boost the performance and accuracy of the algorithm.

Our main focus is to enhance the performance of the MLP algorithm to eliminate bogus data from the dataset and to make MLP an efficient algorithm for email spam detection by removing the disadvantage of randomization of the MLP algorithm. In our research work, we performed N-gram analysis up to four words to avoid the formation of sentence corpus, as increasing the value of N (number of words) slower the performance of the algorithm.

Table 3. Comparison of the N-Gram results on MLP and SVM algorithms for Bi-gram, Trigram and Four-gram.

Percentage Split	Algorithms	SVM ALGORITHM			MLP ALGORITHM		
66%	Parameters	Bi-Gram-SVM	Tri-Gram-SVM	Four-Gram-SVM	Bi-Gram-MLP	Tri-Gram-MLP	Four-Gram-MLP
66%	Correctly Classified Instances (Accuracy)	65.01%	63.95%	63.95%	79.15%	81.62%	79.85%
66%	In-Correctly Classified Instances	34.98%	36.04%	36.04%	20.84%	18.37%	20.14%
66%	Sensitivity	0.622	0.64	0.64	0.792	0.816	0.799
66%	Specificity	0.37	0.360	0.360	0.807	0.838	0.802
66%	Precision	0.387	0.409	0.409	0.806	0.833	0.817
66%	Recall	0.622	0.64	0.64	0.792	0.816	0.799
66%	F-Measure	0.477	0.499	0.499	0.793	0.818	0.8
66%	Root Mean Square Error (RMSE)	0.614	0.600	0.600	0.407	0.392	0.399

VII. Conclusion

In this paper, author illustrated various machine learning and non-machine learning algorithms. From last few decades, the number of account holder has increased and this increased the amount of data and its complexity too. Various illegitimate sources spread its existence over the internet. The major problem user hold is of spam emails from unknown and illegitimate contacts. Various techniques to detect spam emails has discussed by the author in this paper. From various studies conducted so far by various authors it has been concluded that no algorithm guarantees 100% results in spam detection but still there are some algorithms that provide high accuracy for detection of spam emails when used with feature selection technique like MLP neural network but MLP has a limitation of selecting initial information point using a randomized approach which increases the execution and model building time of the MLP algorithm and degrades the performance of the algorithm, so effective and efficient approach to solving the drawback of MLP will be considered and corresponding solution will be carried out in future research which will ensure high accuracy for the detection of spam emails with low execution time along with the n-gram based feature selection technique for removing any noise and outliers in the dataset and for selecting the best possible features from the corpus of irrelevant features by selecting 100 best features.

Acknowledgment

The author expresses its humble thanks to CT group of Engineering, Management, and Technology for their motivational participation and encouragement in the research field. The author also presents its gratitude towards computer science research group for the support. The author is thankful to its mentor for guidance throughout the research work.

References

[1] B. Yu and Z. Xu, "A comparative study for content-based dynamic spam classification using four machine learning algorithms*," Knowledge-Based System-Elsevier*, vol. 21, pp. 355–362, May 2008.

[2] T. A. Almeida and A. Yamakami, "Content-Based Spam Filtering," in International Joint Conference on Neural Networks (IJCNN) - IEEE, pp. 1-7, 2010.

[3] L. Firte, C. Lemnaru, R. Potolea, "Spam Detection Filter using KNN Algorithm and Resampling," in *6th International Conference on Intelligent Computer Communication and Processing- IEEE*, pp.27-33, 2010.

[4] D. K. Renuka, T. Hamsapriya, M. R. Chakkaravarthi and P. L. Surya, "Spam Classification Based on Supervised Learning Using Machine Learning Techniques," in *2011 International Conference on Process Automation, Control and Computing- IEEE*, pp. 1–7, 2011.

[5] R. Shams and R. E. Mercer, "Classifying Spam Emails using Text and Readability Features," in *International Conference on Data Mining (ICDM) -IEEE*, pp. 657–666, 2013.

[6] M. Rathi and V. Pareek, "Spam Email Detection through Data Mining - A Comparative Performance Analysis," *International Journal of Modern Education and Computer Science (IJMECS)*, vol. 12, pp. 31-39, December 2013.

[7] A. Harisinghaney, A. Dixit, S. Gupta, and Anuja Arora, "Text and Image based Spam Email Classification using KNN, Naïve Bayes and Reverse DBSCAN Algorithm," in *International Conference on Reliability, Optimization and Information Technology (ICROIT)-IEEE*, pp.153-155, 2014.

[8] S. P. Teli and S. K. Biradar, "Effective Email Classification for Spam and Non- spam," International Journal of Advanced Research in Computer and Software Engineering, vol. 4, June 2014.

[9] I. Alsmadi and I. Alhami, "Clustering and classification of email contents," *Journal of King Saud University - Computer and Information Science -Elsevier*, vol. 27, no. 1, pp. 46–57, January 2015.

[10] A. S. Aski and N. K. Sourati, "Proposed efficient algorithm to filter spam using machine learning techniques," *Pacific Science Review- A Natural Science Engineering- Elsevier*., vol. 18, no. 2, pp. 145–149, July 2016.

[11] M. Prilepok, and P. Berek, "Spam Detection Using Data Compression And Signatures And Signatures," *Cybernetics and Systems: An International Journal*, vol. 44, pp. 533–549, August 2014.

[12] G. Kaur, R. K. Gurm, "A Survey on Classification Techniques in Internet Environment", *International Journal of Advance Research in Computer and Communication Engineering*, vol. 5, no. 3, pp. 589–593, March 2016.

[13] P. Verma and D. Kumar, "Association Rule Mining Algorithm's Variant Analysis," *International Journal of Computer Application (IJCA)*, vol. 78, no. 14, pp. 26–34, September 2013.

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

[15] Z. Elberrichi and B. Aljohar, "N-grams in Texts Categorization," *Scientific Journal of King Faisal University (Basic and Applied Sciences)*, vol. 8, no. 2, pp. 25–39, 2007.

[16] D. Jurafsky and J. H. Martin, "N-Gram," *Speech and Language Processing*, 2014.

[17] J. Clark, I. Koprinska and J.Poon, "A Neural Network-Based Approach to Automated email classification," *in WIC International Conference on Web Intelligence –IEEE*, 2003.

[18] S. Karamizadeh, S. M. Abdullah, M. Halimi, J. Shayan, and M. J. Rajabi, "Advantage and drawback of support vector machine functionality," in *1st International Conference on Computer Communication and Control Technology- IEEE*, pp. 63–65, June 2014.

[19] A. Fatima, N. Nazir, and M. G. Khan, "Data Cleaning in Data Warehouse: A Survey of Data Pre-processing Techniques and Tools", *International Journal of Information Technology and Computer Science (IJITCS)*, Vol.9, No.3, pp.50-61, 2017. DOI: 10.5815/ijitcs.2017.03.06

[20] M. Iqbal, M. M. Abid, M. Ahmad, and F. Khurshid,"Study on the Effectiveness of Spam Detection Technologies", *International Journal of Information Technology and Computer Science (IJITCS)*, Vol.8, No.1, pp.11-21, 2016. DOI: 10.5815/ijitcs.2016.01.02.

[21] A. Naik, “Density Based Clustering Algorithm,” 06-Dec-2010.[Online].Available:https://sites.google.com/site/dataclusteringalgorithms/density-based-clustering-algorithm. [Accessed: 15-Jan-2017].
[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software,” *SIGKDD Exploration Newsletter.*, vol. 11, no. 1, pp. 1-10, 2009.
[23] R. Ng and J. Han,” Efficient and Effective Clustering Method for Spatial Data Mining,” *In - 20th VLDB Conference*, pp. 144-155, Santiago, Chile,1994.
[24] Cios, K. J., W. Pedrycz, et al., *Data Mining Methods for Knowledge Discovery*, vol. 458, Springer Science & Business Media, 2012.
[25] S. Dixit, and N. Gwal, "An Implementation of Data Pre-Processing for Small Dataset," *International Journal of Computer Application (IJCA)*, vol. 10, no. 6, pp. 28-3, Oct. 2014.
[26] S. Singhal and M. Jena, “A Study on WEKA Tool for Data Pre-processing, Classification and Clustering,” *International Journal of Innovative Technology and Exploration Engineering*, vol. 2, no. 6, pp. 250–253, May 2013.
[27] O. Y. Alshamesti, and I. M. Romi, “Optimal Clustering Algorithms for Data Mining*” Int. Journal of Info. Eng. and Electron. Bus. (IJIEEB)*, vol. 5, no. 2, pp. 22-27, Aug 2013. “DOI: 10.5815/ijieeb.2013.02.04’’.

Authors’ Profiles

Harjot Kaur was born in Jalandhar, Punjab, India in 1992. She received the B.Tech degree in Computer Science and Engineering from C.T. Group of Institution, Jalandhar, India, in 2014. She is currently a student of M.Tech in Computer Science and Engineering from C.T. Group of Institution, Jalandhar, India. The M.Tech degree will be completed in 2017. Her main areas of research interests are Data Mining, Data Warehousing.

Prince Verma, he received the B.Tech degree in Computer Science from MIMIT, Malout (Pb), India in 2008 and M.Tech degree in Computer Science in 2013 from DAVIET, Jalandhar (Pb), India. Currently, he is Assistant Professor in Computer Science Department of CTIEMT, Jalandhar (Pb), India. His research focuses on Data Mining, Algorithm optimisation techniques.

How to cite this paper: Harjot Kaur, Er. Prince Verma," E-Mail Spam Detection Using Refined MLP with Feature Selection ", International Journal of Modern Education and Computer Science(IJMECS), Vol.9, No.9, pp. 42-52, 2017.DOI: 10.5815/ijmecs.2017.09.05