Chapter II

2. LITERATURE REVIEW

2.1. Literature Review

MD. Rafiqul Islam et al. (2005) discussed about different machine learning algorithms for spam filtering and presented a comparative study of spam filters. Their research includes a study of automated filtering and machine learning techniques like rule based, content based, personalized, collaborative, support vector machine and kernel based algorithms for filtering spam. They presented a comparative analysis on different filtering techniques and its advantages.

Ni Zhang et al. (2006) developed a method for filtering spam mails from the Internet service providers in its heavy traffic. They applied their method to email traffic data captured at one of the largest commercial internet service providers in China. They achieved a result of 70.4% reduction of junk mail traffic. Finger print method is used to detect the similar earlier mails and sets a parameter for the email category. Mail database and finger print database are used to store information. By simply adding the entry in the MD and delete the unimportant mails. They explained about the three advantages of BMTC. They are automatic hand-free deployment and online update mechanism, high accuracy in identifying emails, and handling a large amount of data with small memory and reasonable CPU time.

Seongwook Youn et al. (2007) proposed a comparative study for email classification. Neural Network, SVM, Naive Bayesian and J48 classifiers are used to filter spam from the datasets of emails. Neural network consists of data preprocessing, data training and testing. Feature selection extracts more informative and removing irrelevant and redundant features. The preprocessed features are fed into the NN and email classifier is generated through the NN. In the third step of testing the email classifier is used to verify the efficiency of NN. An error back propagation algorithm is used for the experiment. SVM are very successful in learning and generalization tasks. SVMs learn by examples and each example consists of a number of data points followed by a label, which is in the two class classification. They are +1 represents one state and -1 represents another state. The optimum hyper plane separates the two
classes. Support Vector minimizes the distance between the closest +1 and -1 points. It divides two separate classes which are generated from training examples. SVM introduced a separate hyper plane which maximizes the margin between two classes for obtaining a well generalized test data. Naive Bayesian classifier is an effective classifier based on Bayesian theorem and theorem of total probability. J48 is a decision tree creates a binary tree used for classification of legitimate and spam. They evaluated the four classifiers on different datasets and different features.

\[
\text{Correctly-Classified\_Emails} \\
\text{Accuracy (\%) = } \frac{\text{Correctly-Classified\_Emails}}{\text{Total\_Emails}} * 100 \ (2.1)
\]

For evaluating the performance precision and recall were used as metrics for email classification. They suggested J48 and NB classifiers obtained a better result and accuracy than SVM and NN classifiers.

Enrico Blanzieri et al. (2008) proposed a survey on learning based techniques of spam filtering. This Paper discussed about the learning based methods of spam filtering like keyword filtering, image based filtering, language based filtering, filters based on non-content features, collaborative filtering and hybrid approaches. They presented the evaluation and comparison of the results obtained from the various filtering methods.

Ali Ciltik et al. (2008) proposed a method of spam email filtering methods with high accuracies and low time complexities. They took Turkish mails for their research. They used PC-KIMMO system, a morphological analyzer to extract root forms of words as input and produce parse of words as output. This method is based on the n-gram approach and a heuristics. They developed two models, a class general model and an e-mail specific model. The general model classifies the mail as spam or legitimate by using bayes rule. The second model determines the correct class of a message by comparing it with the similar previous message for matching. The third model is a combined perception refined model. It is a combination of above two models. Free word order is used for ordering the word in fixed order for n gram
model. This spam filtering method is based on classifying text contents and raw contents of emails obtaining results from the categorization of data sets. They faced the increase of time complexity problem when handling the larger number of words. Adaboost ensemble algorithm is used to compare with its previous work. They performed extensive tests on various number datasets sizes and initial words. They have obtained a result of high success rates in both Turkish language and English.

A.G.Lopez-Herrera et al. (2008) developed a multiobjective evolutionary algorithm for filtering spam. They evaluated the concepts of dominance and paretoset. SPAM-NSGA-II-GP is used for filtering spam mails. MOEA is used to learn a set of queries with good precision and recall. PUI datasets are used for spam filtering. SPAM-NSGA-II-GP with very strong filtering rules are (high recall and low precision) used to block all the legitimate emails and labeled as spam. They used the weak filtering rules (high precision and low recall) for labeling a minimum portion of spam emails.

Liu Pei-yu et al. (2009) suggested the method of improved bayesian algorithm for filtering spam. KNN algorithm, SVM, decision tree, and improved bayesian algorithm are used for classifying texts. KNN algorithm is a simple and accurate method for spam filtering by using the k nearest neighbor. SVM is also used for filtering spam and finds hyper plane to classify the legitimate and spam mails. It works with smaller training set. Decision tree is used for faster and simple classification which gives higher accuracy of judgment. Bayesian algorithm is a base and simple classification method classifies the mail as C_legal and spam C_rubbish. In the bayesian method one feature is treated as independent of other. Improved naive bayesian algorithm is a combination of bayesian algorithm with boosting method, developed to reduce the rate of misjudgment and improve the accuracy of classification. Boosting is a universal learning algorithm. They treated the naive bayesian algorithm as weaker learning algorithm and made it stronger by boosting it with boosting algorithm. They obtained better result by applying this boosted naive bayesian algorithm for filtering spam.

Alaa El-Halees et al. (2009) developed to filter spam messages in mixed Arabic and English. Six classifiers are used for filtering spam messages and compared the results obtained from these classifiers. Maximum entropy, decision trees, artificial
neural sets, naive bayes, support system machines and k-nearest neighbor are used for spam filtering. Recall and precision are the two ways for presenting the system performance. SVM is used as the best classifier for English and ME performed better than NB in Arabic messages. They suggested increasing the parameter will improve the performance.

Jan Gobel et al. (2009) proposed a method to filter spam mail in a proactive way which intercepts the communication held between spambot and the intended server and redirects its communication with local mail server at the gateway. They collect spam messages at the gateway and obtain the current spam messages sent by spambotnet. They clean the machine system using a software based restoring system to execute next spambot. They have collected the spam messages by resetting the honeypot. The next process is filtering the message in a proactive way. Longest common string algorithm is used for the extraction of emails and single raw template. They generate templates taking subject, xmailer and complete body of the message for consideration. Longer emails are processed first as they took the first email from the list sorted based on text length and considered it as first raw template named α. They took second email from the list and merge it with α to form a second raw template named β which was more specific than previous one. Then they compared both α and β and determine the amount of text that was replaced by the placeholders. If the removed text percentage is below a predefined threshold Ø then they are treated β as their new α. The email used to form β was removed from the list and they continued with third step. If the changed text percentage is above Ø the current β is too generic and is therefore discarded. They used template generation process for detecting the spam rate.

M. Basavaraju et al. (2010) proposed the text based clustering method for spam detection. Preprocessing of data, methodology of classification, vector space model, and data reduction are the methodologies used for spam filtering. The Porters stemming and stopping algorithm are used for preprocessing of data. Hierarchical and partition clustering algorithms are used for partitioning and clustering. They used BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) method to clustering the documents. NNC and K-NNC are the two classifiers used to classify the neighbors. K-NNC classifier is used to classify the patterns. The vector space model
is used to calculate the inverse document frequency of each word i.e. tf-idf test patterns. After clustering of training patterns the non-spam data are stored in the centroids. Test patterns and centroids are passed in to the classification module for spam and non-spam detection.

Alireza Nemaney Pour et al. (2012) proposed for Minimizing the time of spam detection by relocating the filter to the sender messages. They used DSPAM and TREC anti-spam software for filtering spam. They used four steps to detect spam. The first two steps are used to check the IP validity of both the sender and receiver. The Sender IP validity is checked by the mail server and receiver IP validity is checked by the DNS server respectively. The third step checks the category of mail belongs to white list or black list or grey list. In the Final step they applied rule and content based filters for detection of spam. This research helps them to preserve the network resources such as bandwidth, time and memory and also minimize the time.

Dhananjay et al. (2012) developed an adaptive neural fuzzy inference system classifier which includes both the neural networking and fuzzy logic concept to detect the spam message on social networking websites. ANFIS classifier is used to identify the spam from input vector. They identified five input vector parameters. Number of associated user pages, number of times marked as spam, text priority, presence of URL or Hyperlink and the number of common timestamps are the parameters used to classify the spam. They developed the fuzzy inference system with three parts. They are input member function, output member function and the rule set linking the two member functions. Input member function has five parameters like the number of associated user pages, number of times marked as spam by user, presence of Hyperlink or URL, the number of instances of common timestamps and the priority of text in the message. Seven fuzzy rules are used and the output member functions produce a result which equals the rules in the fuzzy rule set. The researcher suggested increasing the parameters would decrease the false positives and improve the detection rate of spam. This system is not designed for a particular specialized social networking website.

Sudhakar. P et al. (2011) developed fuzzy logic concept for spam detection. They applied five fuzzy rules on five fuzzy parameters. The 5 fuzzy parameters are sender address, sender IP, subject words, content words, and attachments. All the five
parameters are compared against the black list and white list. If match was found they considered the parameter as spam or ham. This approach consumes large amount of time to identify spam words.

Subhodini gupta et al. (2012) suggested a fuzzy filtration module for spam detection. They developed two modules. The first module applied stemming, stop-word elimination and tokenization process on the extracted email words. Fuzzy rules are applied on the document set to verify it for spam or ham. In this method five fuzzy parameters are used. They are sender address, sender IP, subject words, content words and attachments. These extracted parameters are passed through the fuzzy rules for detecting spam. This method applied only for plain text used in subject and body content.

Dr. Sonia (2010) developed a vector space model to classify the mail. It converts the mail into matrix and inverse frequency is calculated. They calculated the similarity coefficient by using term frequency and inverse message frequency. The fuzzy decision maker used to take the sc as input for fuzzification. Fuzzification classifies the input as legitimate or spam mail.

Jitendra Nath Shrivastava et al. (2012) discussed about the trends, issues and challenges concerning the spam. They present the role of botnets in spreading spam and predicts the statistical figures of spam mails from 2011 - 2015. It covers the area of the anti spam filtering methods, its classification and consequences on stopping the spam mails.

M. Muztaba Fuad et al. (2004) proposed a method of trainable fuzzy filters for filtering spam mails automatically. A trainable fuzzy classification module consists of a set of fuzzy rules and fuzzy inference system used for the classification of spam. The messages from the corpus parsed and features are extracted. It extracts the features from, to, cc, subject, and header fields. In the fuzzification five fuzzy sets are used in which two sets are used for feature extraction from the header part and others for the body features. In the fuzzification it determines the degree of input that belongs to which fuzzy set. Then the rule antecedents are evaluated by fuzzy AND operation and the consequences are combined by OR operation and passed the output for defuzzification process. It will produce a crisp output and it is compared with
threshold value and predicts the output value as a spam or ham. They have also constructed a confusion matrix (contingency table) to evaluate the classifier’s performance. True positives (TP) are treated as the correct classifications of positive that is spam examples. True negatives (TN) are treated as the correct classifications of negative that is ham examples. False positives (FP) are represented as the incorrect classifications of ham examples into class spam and False negative (FN) are the spam examples incorrectly classified into class of ham. Classifier accuracy \( A_{cc} \) is used to measure the proportion of correctly classified instances.

\[
A_{cc} = \frac{TP + TN}{TP + TN + FP + FN}
\]

(2.2)

Spam precision denotes the percentage of messages in the test data which is truly classified as spam.

\[
S_p = \frac{TP}{TP + FP}
\]

(2.3)

Spam recall denotes the proportion of actual spam messages in the test set that are categorized as spam by the classifier.

\[
S_r = \frac{TP}{TP + FN}
\]

(2.4)

They obtained 90% of spam accuracy in their results. They suggested that this method can eliminate a large amount of spam from inbox of the user.

Mehdi Samiei yeganeh et al. (2012) developed a model for fuzzy logic based machine learning approach for filtering spam. They discussed the methods of automatic spam filters like naive bayes classifier, artificial immune classifier and fuzzy logic. They built a classification model from a set of pre-classified email instances by using fuzzy similarity approach. They used three stages for filtering spam. In the preprocessing stage the html tags are stripped off and stop word are
removed from the mail. In the second stage of training they built a model based on the characteristics of each category in a pre-classified set of email messages. Each Sample message is labeled with a specific category and the message is then classified by comparing its fuzzy similarity measures. They have enhanced the functionality of the model and also enhanced the feature identification of emails and deletion of spam mails on its own. They suggested that the fuzzy logic is adaptable for spammer tactics.

Begol et al. (2011) proposed a fuzzy system method to detect the edges of image. This method is used to process each pixel by means of vicinity. Each pixel value is passed to the fuzzy system as input. The triangular output membership function produces the edge pixel as output. Two types of pixel vicinities used for edge detection. They are four and eight pixel vicinities. In an image grey pixels are treated as noise. The noises are omitted from the image. Canny and sobel method did not detect the edges properly. But fuzzy technique is used to detect the original image edges correctly and eliminates the noises in the image.

Sun park et al. (2012) proposed a method on Email categorization using inherent features and fuzzy theory. In preprocessing phase the keywords are extracted from the subject and body of an email. Word stopping or removing and word stemming are performed by Rijsbergen’s and Porter’s stemming algorithm. In the email category construction the keywords are extracted and constructed the keyword matrix $E$ and this matrix is decomposed into the non-negative semantic feature matrix $W$ and the nonnegative semantic variable matrix $H$ by using NMF. The category labels are generated by this phase by using semantic features of NMF. By using fuzzy association the emails are classified into category labels. When the user is dissatisfied with the result of email category construction by using fuzzy product the email is reorganized. They implemented 6 email classification methods. They evaluated the performance NMF-FA results 10.45% higher than other classification methods.

Christina.V et al. (2010) proposed a study on email spam filtering methods. They discussed about various spam identification methods and spam filtering techniques. In spam identification methods Whitelist/Blacklist, Bayesian analysis, Mail header analysis, and Keyword checking are discussed. In spam filtering techniques they discussed about the naïve Bayesian classifier, SVM, rule based,
content based filters, K nearest neighbors, distributed adaptive blacklists, the multilayer networks, technique of search engines and technique of artificial immune system. They concluded that there is a need to develop a method to provide an ideal solution with 0% false positive and 0% false negative.

Sivakumar (2012) proposed a paper on A Fuzzy Similarity Approach for Automated Spam Filtering and Naïve Bayes Classifier is a near-duplicate phenomenon of spams. SAG is focused on email layout and used html content in email. Structure abstraction generation process composed of three types. They are tag extraction, tag reordering and appending types. SAG captures the near-duplicate phenomenon of spams. They used SpTable and SpTrees to store large amounts of the email abstractions in reported spams. The values assigned to tags by Bayes theorem helps to find the tag as spam or not.

P.Divya et al. (2012) proposed this approach to extract email abstractions from HTML content by using SAG procedure (Structure Abstraction Generation). Preprocessing approach includes two objectives, one is abstractions are extracted by removing unwanted html tags and the other prevents malicious tag insertion attack. Sp trees and Sp table are used to store reported spam abstractions from emails. Sp trees are used for efficient near duplicate matching. Hash function is used to map each subsequence in to integer and this subsequence should be matched. System pattern of Cosdes contains three modules. In the first module Email abstractions are extracted from abstraction generation module with SAG procedure. Database maintenance module used to collect reported spams. Action, insertion and deletion handler functions are involved in this second module. Reputation mechanism is used to block subsequent near-duplicate spams by the collection of human judgment. Five modules are described to find and block spam. They are User registration, mail composer, spam detection module, block list, and report module. In this module the each and every user must register in the domain and send mail to register user only. If they found any spam mail that is reported to the administrator then the administrator will block the spam and also reported to the database. This system works only for registered users but it would not work for whole.

Suganya M. et al. (2012) proposed this fuzzy similarity approach for detecting Spam. The preprocessing Stage includes the removal of html tag and all the stop
words. The messages are tokenized (words, Phrase, or key words). All mixed tokens are converted in to lowercase. The process of stemming is done to their roots means it takes the base form of the word. The base form token appears less than three times than it is removed. Preprocessing is done to extract tokens and determine the number of occurrences in each category and this process is divided by the number of occurrences in all categories. If the token occurs only in one category and the result value is obtained in the membership for the category as 1 otherwise it is 0. Then they determine membership degree to the token in a message for finding the fuzzy similarity measures. They applied various methods to compute fuzzy conjunction and disjunction. They used performance evaluated metrics for evaluating the effectiveness of various fuzzy similarity methods. TP (True Positive- the number of spam messages correctly classified) TN (True Negative- the number of legitimate messages are correctly classified) FP (False Positive- the number of Legitimate messages are classified as spam) FN (False Negative- the number of spam messages are classified as Legitimate). They have obtained above 97% low false positive rate achieved than the naive Bayesian classifier.

Shalini puri et al. (2012) proposed a study on different fuzzy similarity related algorithms and methodologies. They concluded that by using fuzzy logic and fuzzy sets which results good effect on text mining and text classification and paved a better way for text categorization.

Duncan Cook et al. (2006) proposed this approach to stop the spam in the network gateway to reduce the load on mail server. They used domain specific dynamic blacklist (DSDBL) to black list the IP address of the spammers who attack the domain. It has two phases. In the first phase amalgamated audit log information collected from all gateways and mail servers into a single file. In the second stage all the log lines placed in the hash table according to the IP address. The second process includes the reduction of IP address with no email activity and with email activity. The final stage was to output the remaining hash entries. From, msgid, relay are the three fields used in phase1 for classification of spam. Phase2 is used to recognize the pre- cursor activity from spammers in real time helps to identify the spam sending IP address. There are two hypotheses developed by the researchers a simple and alternative. The IP address is classified as malicious or benign. If it is classified as
malicious it is communicated to all the gateways and rules will be added to the gateways to deny the traffic entry to the network from that IP address.

Rafiqul Islam et al. (2010) presented a paper on instance selected method for classification of spam. The ISM is used to select subset of instances. The new dataset helps to reduce the spam rate. Instance selection method divided into three. They are noise filters, condensation algorithm, and prototype construction methods. The first two methods are considered as prototype selection method. This method decides which instances in the training data to include in the reduced dataset, using either incremental or decremental methods. The third type is called Prototype construction method and this method finds new instances that can represent the whole dataset through data squashing or data clustering methods. They extracted feature from incoming emails and preprocessing the email messages in to a format that could be recognized by the classifier. They used tokenization and domain specific feature selection methods for feature extraction. They included the behavioural features like the frequency of sending/receiving emails, email attachment, type of attachment, size of attachment and length of the email for improving performance and reducing false positive problems. They extracted broad set of features from the email and passed into the classification system. In the first initial process started with transformation or preprocessing of email sets and split the entire corpora into different sets. They constructed the training and test set and used the weka data mining tool for classification and validation. They have tested five base classifiers Naive Bayes, SVM, IB1, Decision Table and Random Forest. They also included adaptive boosting as meta-classifier for their test. They concluded that their classification method achieved 97% of accuracy.

Tiago A. Almeida et al. (2010) presented a comparative study on content based filters. They studied seven different versions of naïve baysian classifiers and linear SVM methods employed automatic spam filtering. They used large six databases for their experiment. They found that SVM method’s performance is better than others. To provide fairer comparison they used Matthews’s correlation coefficient (MCC) as the evaluation measurement.
Mehdi Samiei yeganeh et al. (2012) this paper discussed about the machine learning methods they are Naïve Bayes, Artificial Neural Networks, Artificial Immune System Classifier methods, and fuzzy logic method to filter the spam mails. They built a classification model from a set of pre-classified e-mail instances by using fuzzy similarity approach. Their method consists of preprocessing, training and classification stages. They concluded that their model is an enhanced fuzzy model for feature identification and deletion of spam mails.

N.T.Mohammad et al. (2011) developed a method of spam filter by using the Fuzzy C-Mean Clustering algorithm consists of feature extraction, training and testing. They collected the datasets from SpamAssassin Spam Corpus. They used the Heterogeneous Value Difference Metric (HVDM) for normalization of features. They tested this algorithm with the normalized features. They evaluated the use of fuzzy clustering and text mining for spam filtering.

Jens Huhn et al. (2006) developed a method on FURIA: An Algorithm for unordered fuzzy rule induction is an extension of RIPPER algorithm. FURIA is a fuzzy rule based algorithm. A new novel rule stretching technique is used by the FURIA algorithm. RIPPER algorithm differs from FURIA. RIPPER learns conventional rules and rule lists but FURIA learns fuzzy rules and unordered rules. FURIA includes a number of modifications and extensions. They concluded that FURIA performed well than other classifiers like ripper and C4.5.

Han, Weili et al. (2008) proposed this method of automated individual white list approach is a tool used to build white list and automatically maintained by the naïve Bayesian classifier protects user’s web digital identities and also recognize the successful login process. AIWL is an efficient automated tool specializing in detecting phishing and pharming. AIWL checks the LUI information and recognize the phishing and pharming. If AIWL is installed in a machine it is difficult to fight against the Trojan horse and viruses. It has a synchronization problem when the user has many machines.

Qian Xu et al. (2012) who proposed a method used a combination of temporal and network features for spam detection. Feature is based on three steps. They are based on static features, temporal and network. Static feature includes the number of
messages and message size. Temporal includes the features such as size of messages during a day and on each day of week. Finally network feature includes number of recipients and clustering coefficient. SVM and KNN classifiers are used for the spam filtering purpose. They obtained a better result from SVM classifier than KNN. AUC (area under curve) is used to show the performance of these classifiers. These classifiers yield high performance in spam detection in SMS.

Omar Al-Jarrah et al. (2012) proposed a paper on Identifying Potentially Useful Email Header Features for Email Spam Filtering. They identified email header features by analyzed the publicly available datasets. They extracted features from received field, Sender Address Legality, Number of Receivers, Date of Reception, Mail User Agent (MUA), Message-ID, and Email Subject fields. They used CEAS2008 and CSDMC2010 datasets for their experiments. They found that RF classifier performed well than other classifiers.

Hasan shojaa alkahtani et al. (2011) proposed a paper on taxonomy of spam filters. In this paper they classified the spam filters into three techniques. They are origin based techniques, social based techniques and traffic analyzing. Origin based techniques based on network information, such as source IP and email addresses. Here they discussed black lists, white lists, challenge-response systems and origin diversity analysis. In social filtering techniques they classified it as implicit techniques and explicit techniques. They are used to analyze the fields of emails headers like ‘To’, ‘Cc’, ‘Bcc’. They build a graph of social relations of users and classify the new mails based on this graph in the implicit techniques. Whereas in the explicit techniques they build the social network through user interaction and also they utilized user-supplied or automatically computed reputation ratings. The traffic analysis technique is a common analysis used to detect the spam is to identify when a host or network issues an abnormally large amount of emails. They also discussed about the content based filtering techniques, which includes heuristic filters, machine learning and finger printing. And also they addressed the problems to circumvent in these techniques.

Ahmed Khorsi et al. (2007) proposed a method on an overview of Content-Based spam Filtering Techniques. This paper distinguished two methods of machine learning classification. One is based on the rules defined manually and the second one
is on using machine learning techniques. They discussed the Bayesian classifier, k nearest neighbors, support vector machine, maximum entropy, neural networks, search engines, genetic programming, and artificial immune system. They concluded that the hierarchical filter is the only solution for the generalized classification problem. And the available classifier of multimedia documents has time and space complexities remain far from the requirements of a real time computation.

2.2. Objective of the thesis

The present work has been developed for the following objective:

Ø To identify the spam mails using spam words and spammer’s address ranking and Classification by using fuzzy rules.

2.3. Organization of the thesis

The present study is organized into seven chapters.

Chapter 1 presents the general introduction with the definition of Spam, characteristics of spam, categories of spam, spam techniques, spam filtering techniques, disadvantages of spam, non-technological solutions, email address harvesting by spammers and impacts of spam mail.

Chapter 2 it provides literature survey of previous works. Spam filtering techniques have been discussed in this chapter.

Chapter 3 describes the proposed work of Classification of spam. In this chapter the architecture of the spam classification which includes ranking and classification process. And also it describes the spam filtering steps.

Chapter 4 discusses about the Implementation of this work. This chapter discusses about the configuration of outlook for email accounts, functioning of spam report generator, structure fuzzy inference system, input and output membership functions, fuzzy rules, and surface viewer.

Chapter 5 this chapter discusses about the results.
Chapter 6 discusses about the Summary.

Chapter 7 draws the conclusions from the research work and discusses the possibilities of extension of the work is also mentioned.