Performance Evaluation of Spam Filtering Using Bayesian Approach

Archana Sahu, Amit Mishra, Shiv Kumar Sahu

Abstract: ***Spam filtering is the technique to find out spams. This field is important aspect of text classification. Spam filtering technique is used with email servers, and population of spam is usually more than genuine emails, this is why spam filtering has become important technique. Most of existing spams filtering techniques are unable to detect spam because spammers know how to make spam to reach the destined email account without being filtered. In such situation, naïve bayes spam filter is proved to be a great technique, because several aspects are there to improve the performance of spam filter. Hence, it is an important research field in detecting spams. In this dissertation, technique for spam detection and filtering has been proposed based on Naïve Bayes classification technique, which is the existing spam filtering technique. Some enhancements are made in making it adaptive to new kind of spams. In existing spam filtering techniques, static filtering technique has been used, but we proposed dynamic and enhanced filtering technique, which helps in fast and accurate spam detection. Regular training of classifier should be done, database of spam should be updated all the time, and also a particular word should not be always behaved as spam word or a genuine word. Experimental results show that proposed enhancements improves accuracy of spam filtering.***

Keywords: ***Spam filtering, detecting, field, accuracy proposed enhancements, classifier Regular, proposed, spam***

I. INTRODUCTION

There are numerous text documents available in electronic form. More are becoming available constantly. The Web itself contains over a billion documents. Millions of people send e-mail every day. Academic publications and journals are becoming available in electronic form. These collections and many others represent a massive amount of information that is easily accessible. However, seeking value in this huge collection requires organization. Many web sites offer a hierarchically-organized view of the Web. E-mail clients offer a system for filtering e-mail. Academic communities often have a Web site that allows searching on papers and shows an organization of papers. However, organizing documents by hand or creating rules for filtering is labor-intensive. This can be greatly aided by automated classifier systems. The accuracy and our understanding of such systems greatly influence their usefulness.

We aim to advance the understanding of commonly used text classification techniques and through that understanding, to improve upon the tools that are available for text classification. Text classification is the task of increasing importance and the key technique of Web content mining. It refers to the automated assigning of natural language texts to predefined classes based on their contents. The familiar methods include Bayes, K Nearest Neighbor (KNN), Support Vector Machine (SVM), decision tree method and neural network algorithm. These classical methods usually don't process words at the initial stage instead only adopt simple method of word frequency statistics to index based on word forms themselves. Its disadvantages are skin-deep and formal. As it is short of sentential analysis, it can't truly reflect the signification of text and query.

As the volume of on-line text documents grows continuously in networked resources such as the Internet, digital libraries, news sources and company-wide intranets, (automated) text classification becomes highly important technique not only from an academic but also from an industrial point of view. In real-world operational environment, text classification systems should handle the problem of incomplete training set and no prior knowledge of feature space. In this regard, the most appropriate algorithm for operational text classification is the Naive Bayes since it is easy to incrementally update its pre-learned classification model and feature space.

The first step in text categorization is to transform documents, which typically are strings of characters, into a representation suitable for the learning algorithm and the classification task.

Information Retrieval research suggests that word stems work well as representation units and that their ordering in a document is of minor importance for many tasks. This leads to an attribute-value representation of text. Each distinct word corresponds to a feature, with the number of times word occurs in the document as its value. To avoid unnecessarily large feature vectors, words are considered as features only if they occur in the training data more than fixed number of times and if they are not “stop-words” (like ‘and’ ‘or’ etc.). Millions of people send e-mail every day, but main problem of email users is spams. Spams are undesired emails, which we don’t want to be in our email account, so filtering of spam is becoming very necessary. E-mail servers offer a system for filtering e-mails and save our time and bandwidth.

Manuscript published on 30 June 2016.
***** Correspondence Author (s)

Archana Sahu, M.Tech. Scholar, Department of Information Technology, Technocrats Institute of Technology, Bhopal (M.P.), India.
Amit Mishra, Asst. Professor, Department of Information Technology, Technocrats Institute of Technology, Bhopal (M.P.), India.
Dr. Shiv K Sahu, Professor & Head, Department of Information Technology, Technocrats Institute of Technology, Bhopal (M.P.), India.

Spam Filtering is the text classification technique which proved to be a great technique for dealing with spams. It refers to the automated assigning of emails to predefined classes as Spam or Genuine email based on their contents.

The familiar methods include Bayesian filter, Support Vector Machine (SVM), instance based classifiers, neural network classifiers etc. These methods usually don't process words at the initial stage instead only make use of simple method of word frequency. So these classifiers don’t show desired results. Bayesian approach is the statistical-based spam filter method [1] which is strong algorithm for classification. But it is not so good in self-learning and self-adaptability [2]. Dynamic training and classification has shown impressive performance of filtering spam. So it feasible to combine the advantages of this method with Bayesian approach into a single model. In this dissertation, we attempt to combine the mechanism of Naive Bayes [3] and dynamic nature into a single algorithm [4]. Our experiment results proved that this algorithm is effective to filter spam.

Spam or unsolicited e-mail has become a major problem for companies and private users. This dissertation explores the problems associated with spam and some different approaches attempting to deal with it. The most appealing methods are those that are easy to maintain and prove to have a satisfactory performance. Statistical classifiers are such a group of methods as their ability to filter spam is based upon the previous knowledge gathered through collected and classified e-mails. A learning algorithm which uses the Naive Bayesian classifier has shown promising results in separating spam from legitimate mail. Tokenization, probability estimation and feature selection are processes performed prior to classification and all have a significant influence upon the performance of spam filtering. The main objective of this work is to examine and empirically test the currently known techniques used for each of these processes and to investigate the possibilities for improving the classifier performance. Firstly, how a filter and wrapper approach can be used to find tokenization delimiter subsets that improve classification is shown. After this, many probability estimators are tested and compared in order to demonstrate which of them improve the performance. Finally a survey of commonly used methods for the feature selection process is performed and recommendations for their use are presented.

A. *Motivation*

Spam filtering task is becoming more challenging. People use email services but most of the emails are spam. Percentage of spams in emails is more than genuine emails. Because of spams there is difficulty in checking emails, and also spams take our useful time and bandwidth. So a proper filter is required to ensure that our time and bandwidth is not wasted in checking spam. Also sender of spam are getting smarter, they also are adapting there spams to the existing spam detectors. Spammers use several techniques so that is not being filtered. They use words similar to spam words, thus these words are not caught easily. These tricks by spammers are making spam filtering task more challenging.

II. GENERAL SPAM DETECTION METHODS

A. *Neural network*

In text classification task, neural networks [5] are based on two approaches. One is analytical learning and other is inductive learning. Some neural networks use combination of these two.

B. *Analytical versus Inductive Learning:*

Inductive learning methods learn a general hypothesis by finding empirical regularities over training examples. Neural networks fall under this category. Purely inductive methods have an advantage that they do not require explicit prior knowledge and learn solely using the training data. They use statistical inference to learn beyond observed data. When given insufficient training data they can be misled by implicit inductive bias that they use to generalize beyond observed data. Analytical methods use prior knowledge to derive general hypothesis that fits domain theory. These have an advantage of learning from scarce data, but they can be misled if the prior knowledge is imperfect. Naive Bayesian Classifier computes posterior probability of a hypothesis from observed prior probabilities in given training data. So accuracy depends upon the probability distribution of the training data. Learning problems vary by availability of training data and prior knowledge. At one end, a large volume of training data is available but no prior knowledge. At another extreme, strong prior knowledge is available but little training data. Most practical learning problems lie somewhere between these two extremes. Hence we are interested in systems that take prior knowledge as an explicit input and learn using a combination of inductive as well as analytical learning that achieves better results. There are many types of neural networks. In feed forward neural networks, there are three or more layers of such neurons: input, middle (or hidden) and output layer. The number of nodes in the output layers is equal to the number of target classes. The number of input layer neurons depends on the number of attributes defining an instance of an object. The number of neurons in hidden layers is not fixing. It is normally decided by studying the problem.

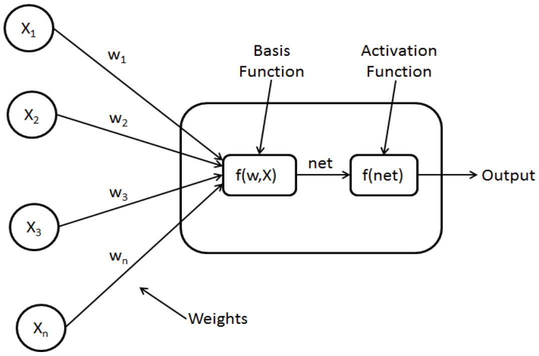


Fig2.1: Artificial Neural Network

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

All consecutive layers are fully connected. Learning and classification is done as follows:

1. Select a basis and activation function. Usually linear is taken as basis and sigmoidal is taken as activation.
2. Design the neural network according to the number of features and classes.
3. Initialize all weights to small random values.
4. Use "back propagation algorithm" to learn the weights of networks.
5. Use learned network to find the target category.

They use statistical inference to learn beyond

C. K-nearest neighbor classifier

K-nearest neighbor classifier (KNNC) [6] is widely used because of its simplicity. It includes k-nearest neighbors search (KNNS) and classification. Existing centralized KNNS does not scale up to large volume of data, and the classification still suffers from inductive biases that result from its assumptions, such as the presumption that training data are evenly distributed Text categorization is one important task of text mining, for automated classification of large numbers of documents. Many useful supervised learning methods have been introduced to the field of text classification.

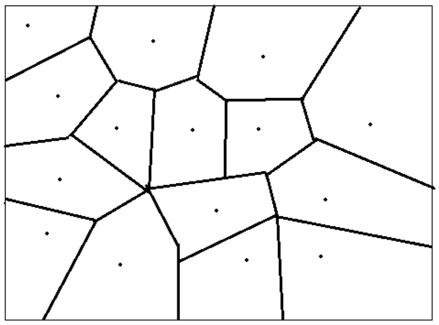


Fig2.2: 1-Nearest Neighbor Classifier

Among these useful methods, K-Nearest Neighbor (KNN) algorithm is a widely used method and one of the best text classifiers for its simplicity and efficiency. For text categorization, one document is often represented as a vector composed of a series of selected words called as feature items and this method is called the vector space model. KNN is one of the algorithms based on the vector space model. However, traditional KNN algorithm holds that the weight of each feature item in various categories is identical. Obviously, this is not reasonable. For each feature item may have different importance and distribution in different categories. Considering this disadvantage of traditional KNN algorithm, we put forward a refined weighted KNN algorithm based on the idea of variance. Experimental results show that the refined weighted KNN makes a significant improvement on the performance of traditional KNN classifier.

D. Support vector machine

The following paper support vector machine algorithm [7] based on Naive Bayes adopted in this paper is given as follows:

- Select i samples from data sample set S, which never takes any categorization labels and mark its categorization correctly, then construct and initialize training sample T, so that T at least includes a sample whose output are 1 and -1.

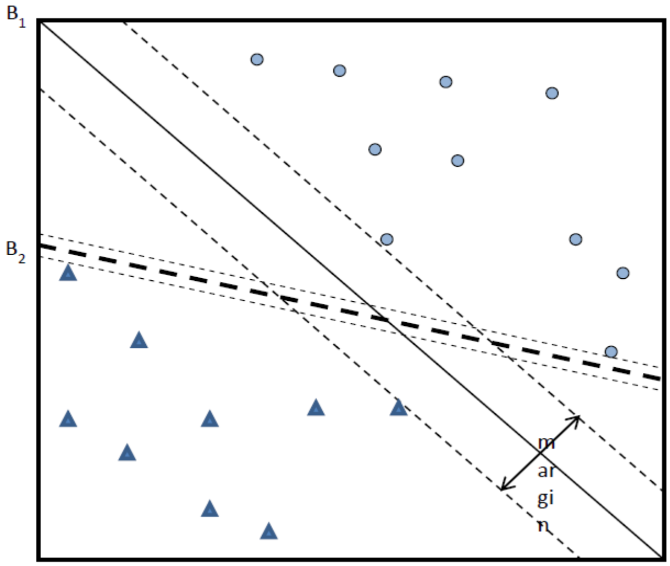


Fig2.3: Support Vector Machine

- Construct the optimal classifier f according training data T, namely, use Naive Bayesian principle to classify unknown sample X and calculate each class until X is assigned into the largest class Ci.
- Randomly select sample M in S and add to training set.
- Test M in classifier f.
- Test and evaluate, if testing accuracy rate comes up to a certain value, the algorithm is terminated. Otherwise, repeat step 2

E. Bayesian Classifiers

Bayesian classification theory [8] was originally derived from the Bayes theorem in probability theory. The theorem says something about the future probability of occurrence can be calculated it has occurred to estimate the frequency. It applies to the message category (such as spam and legitimate e-mail); the text belongs to a category by calculating the probability of the text fall into the category to the maximum probability to determine the message type, in the calculation, the use of Bayesian probability formula.

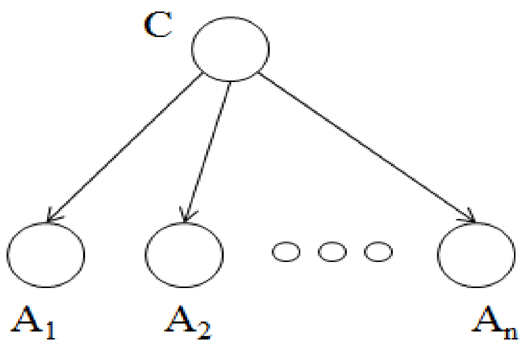


Fig 2.4: Bayesian Classifier

Naive Bayes classification model is the use of the word categories of prior probability and the distribution of the types of conditional probability to calculate the unknown text belongs to a class of probability. It is based on "Bayesian hypothesis" based on the assumption that independent of each other among all the features. Applied to the message classification, it is the first to provide a certain number of users via e-mail spam and non-spam e-mail the training set as automatic classification model training, the training results as the main basis for determining unknown e-mail, applied to the appropriate classification algorithm.

F. Why Bayesian filtering is better

1. The Bayesian method considers the whole message and find out keywords that are able to find spam, but it also recognizes words that denote genuine emails. For example: not every email that contains the word “money” and “job” is spam. The advantage of the Bayesian method is that it considers the most interesting words. Interesting words are those which are useful in classification task. After finding these words, probability of message being a spam is calculated. The Bayesian method would find the words “money” and “job” interesting but it would also check for the name of the business contact that sent the message and thus classify the message as legitimate, for instance; it allows words to “balance” each other out. In other words, Bayesian filtering is a much more intelligent approach because it examines all aspects of a message, as opposed to keyword checking that classifies a mail as spam on the basis of a single word.
2. A Bayesian filter is constantly self-adapting - By learning from new spam and new valid outbound mails, the Bayesian filter evolves and adapts to new spam techniques. For example, when spammers started using “m-o-n-e-y” instead of “money” they succeeded in evading keyword checking until “m-o-n-e-y” was also included in the keyword database. On the other hand, the Bayesian filter automatically notices such tactics; in fact if the word “m-o-n-e-y” is found, it is an even better spam indicator, since it’s unlikely to occur in a genuine emails. Another example would be using the word “5urprising” instead of “Surprising“. You would probably not have a word 5urprising in a genuine mail, and therefore the likelihood that it is spam increases.
3. The Bayesian technique is sensitive to the user. It learns the email habits of the company and understands that, for example, the word ‘mortgage’ might indicate spam if the company running the filter is, say, a car dealership, whereas it would not indicate it as spam if the company is a financial institution dealing with mortgages.
4. The Bayesian method is multi-lingual and international – A Bayesian anti-spam filter, being adaptive, can be used for any language required. Most keyword lists are available in English only and are therefore quite useless in non English-speaking regions. The Bayesian filter also takes into account certain languages deviations or the diverse usage of certain words in different areas, even if the same language is spoken. This intelligence enables such a filter to catch more spam.
5. A Bayesian filter is difficult to fool, as opposed to a keyword filter – An advanced spammer who wants to trick a Bayesian filter can either use fewer words that usually indicate spam, or more words that generally indicate valid mail is impossible because the spammer would have to know the email profile of each recipient - and a spammer can never hope to gather this kind of information from every intended recipient.
6. Using neutral words, for example the word “public”, would not work since these are disregarded in the final analysis. Breaking up words associated with spam, such as using “m-o-r-t-g-a-g-e“ instead of “mortgage“, will only increase the chance of the message being spam, since a legitimate user will rarely write the word “mortgage” as “m-o-r-t-g-a-g-e”.

G. Bayesian filters or updated keyword lists?

Some types of anti-spam software regularly download new keyword files. While this is, of course, better than not updating keyword lists, it is a rather patchy approach that is easily circumvented. Downloading updates makes it a little bit harder, but the principal system is flawed compared to a Bayesian filter.

H. Advantage of Bayesian classification

Compared to other classification filtering, Bayesian classification method has the following advantages

Better than the other algorithms in efficiency. Bayesian classification algorithm to scan all the training samples again, and statistics for each word in the normal email and spam in the number of occurrences of each Token after the query just once again, the last Token for each product or additive. The SVM method requires scanning multiple training samples.

- In storage, Bayesian classification algorithm only needs to store the number of words, rather than the actual message. Thus, very little storage space, but the resulting data can be shared between users without having to consider the privacy of the message.

- Bayesian classification methods continue to receive a single message with the incremental update, you can adapt to the evolution of forms of spam. Changes in the content of spam was more, Bayesian classification methods can be collected from users under the guidance of recently received spam features, effectively filtering.
- Bayesian classification method is suitable as a personal filter. Each user can customize the filter allows the filter more effective, you can customize your filtering accuracy, but also the content of feature selection can be defined, so that spammers are difficult to adjust the message through the filter.

III. NEW ALGORITHM FOR SPAM FILTERING

A. *Training Stages*

Collection of known emails – For training of spam filter, collection of emails is needed whose classification labels are known. This collection should be from several and different kind of sources so that it contain all kind of spam and genuine words. If it from few sources then it would not train spam filter well and when filter would face to a spam of new kind then it would be classified incorrectly. Also training of filter should be frequent thus it would have new spam words in its knowledge base.

Preprocessing of emails- In next step of training filter, some words like conjunction words, articles are removed from email body because those words are not useful in classification. Sender information is stored separately so that it would be used for new arrived emails for prediction purpose. Studying the email sending habit will be useful for filtering spam. So a record of senders is also maintained who send mostly spams or who sends mostly genuine emails.

Creating Hash map of words- After preprocessing task, a hash map of words is created and count of each word occurring in emails is also maintained with words. Suppose if a word exists in hash map then count related to word is increased on occurrence of word, but if word is not in hash map, then put the word in hash map with single count. Also treat different forms of words like singular-plural, different forms of verb as same.

Calculating probabilities- Probabilities of word occurring in spam and genuine emails is calculated. Then spam probabilities of words are calculated.

Spam probability of a word,

$$Sp = f1/(f1+f2), \qquad \ldots\ eq^n \qquad (4.1)$$

Here f1 is frequency in spams and f2 is frequency in genuine emails.

Sorting words in relevant order of probabilities- after calculating spam probabilities, all words are sorted on the basis of their spam probabilities. Words which have spam probability more than 0.85 make their probability1. Words which have very high spam probability or a very low spam probability are useful in classification task because they are either known for spam words or known for classification task easy. These words are also called interesting words. Words which have spam probability near 0.5 are not interesting words and those are not very much useful because they appear in both kinds of emails. So those words are ignored.

IV. CLASSIFICATION STAGES

Prepare a set of emails for testing- A separate collection of emails is also needed for testing our spam filter. Classification email set can be smaller than training email set because for making spam filter more accurate, we need a large data set for training.

Preprocessing of emails- As we have done in training stage, words like conjunction words and article words are removed because these words cannot predict about an email being a spam.

Generate interesting word list- a word list is generated which contains n words which are useful in classification task or that exist in hash map with very high spam probability or very low spam probability. Value of n might be taken as 10 to 15 that is enough for calculating spam probability

Finding overall spam probability- For finding out overall spam probability, spam probabilities of all words are multiplied, this will be overall spam probability. Now also check for sender in sender's record. If email is send by a habitual spam sender than add 0.2 to spam probability. This additional probability is added because if a sender is habitual spam sender then there is a great chance of email being spam otherwise if past record of sender is good, then reduce 0.2 from spam probability.

$$P(A_1,\, A_2,\cdots,A_n/C) = P(A_1/C)\ \times P(A_2/C)\ \times\cdots \times P(A_n/C) \qquad eq^n\ (4.2)$$

(Formula for calculating overall spam probability)

Here A_1, A_2, …, A_n are meant for words and C is meant for class Spam.

Classifying an email- if overall spam probability is more than 0.5 than an email can be classified as spam.

V. TRAINING AND PREDICTION

Spam messages are not hard to collect in large quantities, but negative ones are difficult to collect. As a popular classification algorithm, Naïve Bayes algorithm has been used in spam email detection. Naïve Bayes can be defined as Bayes Theorem with a conditional independency assumption that all variables in a given category C are conditional independent with each other. Spam filtering is becoming challenging task because existing techniques [26] are not adaptive to new kind of spams. And also, these days, senders of spams are getting smarter and using tricks for emails not to be filtered for e.g. they do not use words which make them being treated as spam by using alternative words, changing spelling of words, using different email ids every time.

So there is a requirement of a filter that is adaptive to these tricks. Overall spam filtering task is divided into two steps. One of them is training of spam filter and other is classification of emails. In first step training of filter is done by calculating probabilities and in classification step, an email is classified based on the calculated probabilities.

Overall spam filtering task is divided into two steps-

a. Training spam filter.

b. Classification of email or prediction of classification label.

Before starting classification large feature space should be reduced by applying some dimensionality reduction techniques.

A. *Dimensionality Reduction Or Feature Selection*

Data dimension reduction has many applications. Data visualization in 2D or 3D provides further insight into the data structure, which can be used for either interpretation or data model selection. Data dimension reduction can allow for extracting meaningful features from a very bulky representation. For example, in text document classification the bag of words model offers a vector representation of the relative word frequency over a dictionary. With a large dictionary, each document can be identified with a high dimensional vector. Other aspects of data dimension reduction involve de-noising or removal of redundant or irrelevant information. For example, when going after object orientation in video of a single object from multiple angles simultaneously, the relatively high volume information about the object shape can be discarded and only its orientation is retained. Feature selection is also used in combination with other techniques of classification. Text is converted into another reduced form. This reduced text is in form of vectors.

By combining feature selection and dimensionality reduction, better results of classification are obtained. Feature selection is to select a set of most discriminative features from feature space, so as to reduce the size of feature space, and to enhance the performance of text classifiers. There are several successful feature selection algorithms. Feature selection is the process of selecting a subset of the terms occurring in the training set and using this subset as features in text classification. Feature selection serves two main purposes. First, it makes training and applying a classifier more efficient by decreasing the size of the effective vocabulary. Second, feature selection often increases classification accuracy by eliminating noise features. A noise feature is one that, when added to the document representation, increases the classification error on new data[3].Currently, a selection method for feature subsets is usually built by appraisal functions to assess each feature respectively and independently. Consequently, each feature receives its assessment score, and then we list the features in order of assessment score. After that, we select the optimal feature as our subsets according to the predefined rules.

Frequency –based feature selection: Frequency usually can be either defined as document frequency or as collection frequency. Document frequency is more suitable for the Bernoulli model; collection frequency is more suitable for the multinomial model. There is an assumption for document frequency which is a low frequency when it is less than a threshold value. It could eliminate noise features and reduce dimensionality. That is the reason why text categorization must take a feature selection. We define w as word or term, c as category of training sample. If features are too many then frequency based feature selection method often does well. Frequency-based feature selection can be a good alternative to more complex methods, though it is the simplest one in many feature selection methods.

After preprocess, the dimension of feature vector is still very high, we need generate feature vector with a low dimension under the condition of not losing categorization information. One method is to select some of the most effective feature words from the original set to form a new feature vector, this process is called feature selection, another method is to map higher-dimensional feature vector into lower-dimensional space (generally linear mapping),and this process is called feature extraction.

VI. EXPERIMENT AND RESULTS

A. *Data Set Description*

A dataset of 1000 emails was used for training of our spam filter. This dataset consisted of 500 spam emails and other 500 genuine emails.

1000 more were used for classification or testing purpose. Test set consist of 500 spams and others as genuine emails.

B. *Spam Filtering Using Proposed Model*

1. The true positives (TP) and true negatives (TN) are correct classifications.
2. A false positive (FP) occurs when the outcome is incorrectly predicted as yes (or positive) when it is actually no (negative).
3. A false negative (FN) occurs when the outcome is incorrectly predicted as negative when it is actually positive.
4. Recall: The percentage of the total relevant documents in a database retrieved by your search. If you knew that there were 1000 relevant documents in a database and your search retrieved 100 of these relevant documents, your recall would be 10%. Recall =TP/ (TP+FN)
5. Precision: The percentage of relevant documents in relation to the number of documents retrieved. If your search retrieves 100 documents and 20 of these are relevant, your precision is 20%.

Precision=TP/ (TP+FP)

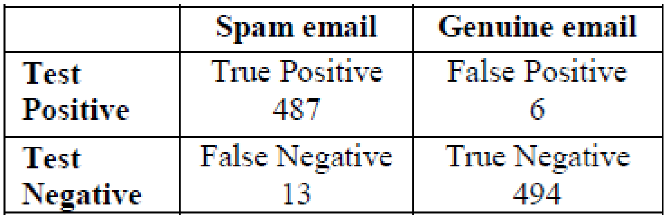
There are two measures, which give the accuracy of a spam filter. One of them is sensitivity and other is specificity. Sensitivity of a spam filter is the probability of positive test given that email is spam.

Sensitivity of spam filter, S= $p_t / (p_t + n_f)$ ... *eq^n (5.1)*

Here, pt is number of true positives and n_f is number of false negatives.
Calculated sensitivity = 487/500 = 0.974
negative test given that the email is genuine.

Specificity of spam filter, s= $n_t / (n_t + p_f)$ ... *eq^n (5.2)*

Here, nt is number of true negatives and pf is number of false positives.
Calculated specificity, s = 494/500 = 0.988



	Spam email	Genuine email
Test Positive	True Positive 487	False Positive 6
Test Negative	False Negative 13	True Negative 494

VII. RESULTS AND COMPARISON

Several spam filtering techniques are compared. Naïve bayes algorithm gave best results out of those. Here are some previous results of spam filtering techniques on different size of training set. Based on existing results, naïve bayes shows best accuracy of 92.4%

VIII. CONCLUSIONS

In this dissertation we showed that adaptive filtering targeting specific categories of spam messages improves the overall spam filtering when it is combined with other techniques as an additional layer. Using our filter, we were able to block additional spam messages that smart keyword filter failed to stop.

- Most of spam detection techniques are unable to find the spams effectively, because regular training of these classifiers is not done. However we have updated the database of emails all the time with more number of training emails. Thus new type of spams is also detected thus improving quality of the spam detection. Adding new emails for training the spam filter makes it adaptive in nature.
- Existing spam filters are static in nature, because of that these spam filter show false positive or false negative results. Regular training of spam filter improves spam filtering extremely. Also number of false positive and false negative is reduced.
- The results show that using adaptive technique with naïve bayes algorithm the overall spam filtering is improved. The words known for spam are given chance to be considered as genuine by checking the sender of the email. Thus genuine emails containing such type of words are prevented from being classified as spam. Similarly a spam containing genuine words is prevented from being classified as genuine.

REFERENCES

1. Meena, M.J.; Chandran, K.R.; , "Naïve Bayes text classification with positive features selected by statistical method,", 2009. ICAC 2009. First International Conference on Advanced Computing, vol., no., pp.28-33, 13-15 Dec. 2009
2. Yan Zhou; Mulekar, M.S.; Nerellapalli, P.; "Adaptive spam filtering using dynamic feature space," , 2005. ICTAI 05. 17th IEEE International Conference on Tools with Artificial Intelligence, vol., no., pp.8 pp.-309, 16-16 Nov. 2005
3. Haiyi Zhang; Di Li; , "Naïve Bayes Text Classifier", 2007. GRC 2007. IEEE International Conference on Granular Computing, vol., no., pp.708, 2-4 Nov. 2007
4. Pelletier, L.; Almhana, J.; Choulakian, V.;, "Adaptive filtering of spam," , 2004. Proceedings. Second Annual Conference on Communication Networks and Services Research, vol., no., pp. 218-224, 19-21 May 2004
5. Saha, D.; , "Web Text Classification Using a Neural Network," , 2011 Second International Conference on Emerging Applications of Information Technology (EAIT), vol., no., pp.57-60, 19-20 Feb. 2011
6. Lijuan Zhou; Linshuang Wang; XuebinGe; Qian Shi; , "A clustering-Based KNN improved algorithm CLKNN for text classification," 2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR), 2010, vol.3, no., pp.212-215, 6-7 March 2010
7. Amayri, O.; Bouguila, N.; , "Online spam filtering using support vector machines," IEEE Symposium on Computers and Communications, 2009. ISCC 2009., vol., no., pp.337-340, 5-8 July 2009
8. Yin; Zhang Chaoyang; , "An Improved Bayesian Algorithm for Filtering Spam E-Mail," 2nd International Symposium
9. on Intelligence Information Processing and Trusted Computing (IPTC), 2011, vol., no., pp.87-90, 22-23 Oct. 2011
10. Sang-Bum Kim; Kyoung-Soo Han; Hae-Chang Rim; Sung HyonMyaeng; , "Some Effective Techniques for Naive Bayes Text Classification," IEEE Transactions on Knowledge and Data Engineering, vol.18, no.11, pp.1457-1466, Nov. 2006
11. Zhang Yang; Zhang Lijun; Yan Jianfeng; Li Zhanhuai; , "Using association features to enhance the performance of Naive Bayes text classifier," Fifth International Conference on Computational Intelligence and Multimedia Applications, 2003. ICCIMA 2003.Proceedings, vol., no., pp. 336- 341, 27-30 Sept. 2003
12. M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. "A bayesian approach to filtering junk e-mail". In Learning for Text Categorization: Papers from the 1998 Workshop, Madison, Wisconsin, 1998.
13. Tarek M Mahmoud, alaa Ismail EI Nashar, Tarek Abd - EI - Hafeez ans Marwa Khairy "*En Efficient Three Phase Email spam Filtering Technique*" British Journal of Managnent & Computer Science 4(9), 1184-1201, 2014