IJCSN International Journal of Computer Science and Network, Volume 3, Issue 3,June 2014 **Impact Factor – 0.274**
ISSN (Online) : 2277-5420
www.IJCSN.org



# Empirical Study of Different Classifiers with Feature Extraction for E-Mail Spam Filtering

**Himadri Sekhar Atta**

M.Tech Final Year Department of Computer Science and Engineering
Institute of Engineering and Management, West Bengal, India

**Abstract-** E-mail or electronic mail is a principal mode of communication for quite some time in both professional and personal use. But over the last few years email spam has rapidly increased. Several techniques have been adopted for spam filtering. Among the various approaches developed to eliminate spam, filtering is an important and popular one. In this paper, an empirical study is done using some email datasets. In the first step datasets were taken and various classifiers like naive bayes, SVM, *k*-NN and decision tree were implemented and the performances were observed. In the next level, the important features were extracted from the datasets and then performances of the classifiers were observed. The objective of this paper is to highlight the findings through the empirical study, which will also help us to determine a good classifier for spam filtering. It also illustrates the information regarding feature extraction and different classifiers.

**Keywords-** *Classifiers, Feature Extraction, Filtering, Spam, Spam Filtering.*

## 1. Introduction

With the rise of technologies, use of computers and computer networks, internet in this era made our life more hassle free. Sending electronic mails or e-mail today is another quick way of communicating with each other. E-mail or electronic mail is generally an electronic messaging system that transmits messages across computer networks or from one location to another. In this system, users simply type in the message to be delivered, then add the recipient's e-mail address or addresses and then finally click on the send button. Yahoo mail, Gmail, Hotmail, etc are some free e-mail service provider that can be accessed by users or they can simply register with ISPs (Internet Service Providers) in order to obtain an e-mail account. Besides this facility, e-mail can be also received almost immediately by the recipient once it is sent out [1].

With the growing of these e-mailing facilities, the growth of illegal actions(e.g., fraud, money laundering etc.) is also increasing[2][3]. There are lots of unwanted commercial mails that contain spam, which is a considerable threat to the users. Nowadays a typical user gets up to 10-20 unwanted or spam mails. Spam mails which we generally know as unsolicited mails are the unwanted and undesirable mails that may be any sort of commercial mails [4]. The problem of undesired e-mails is nowadays a serious issue, as spam mails constitute up to 75–80% of the total amount of email messages [5]. Spam causes several problems, spam refers to bulk unsolicited commercial e-mail sent indiscriminately to users, some of them resulting in direct financial losses [10]. More precisely, spam causes misuse of traffic, allocate storage space and use computational power [6][7][9].

Several anti-spam legal measures are gradually being adopted, but still they have had a very limited effect so far. Many anti-spam filters, software tools are built up that attempt to block automatic spam messages. Apart from the blacklists of frequent spammers and lists of trusted users, which can be incorporated into any anti-spam strategy, these filters have so far relied mostly on manually constructed keyword patterns.(e.g. Blocking messages whose bodies contain "It's all Free", "Call Me for Free", etc.). It is, therefore, required to develop anti-spam filters that will learn automatically how to block spam messages by simply processing previously received spam and legitimate messages [8]. Sometimes it is also found that some spam mails enter into our inbox and sometimes some of our useful mails are being identified as spam by our mailing system spam filters. This is a major problem in today's life that needs to be checked quickly to make our life spam free. In this paper a detailed study of the characteristics of some spam mails is provided. This paper also depicts the study of some classifiers like naive bayes, *k*-nn, SVM and decision tree which helps in classifying the mails as spam or legitimate after feature extraction.

The remainder of this paper is organized as follows. Few related works in this field are discussed in section 2. In Section 3 we discuss about some characteristics of spam mails, feature extraction, various classifier algorithms used and information about the datasets. The proposed methodology is shown in section 4. In section 5 we discuss the detail analysis of our experiment result. The conclusion is presented in section 6.

## 2. Related Work

Spam filtering in e-mails is an open research problem and a lot of work and researches are going on to build a

better spam filter. An illustrated idea about how the organizations are facing problems and the impact of spam on the e-mail marketing system is discussed in [1][9]. Several machine learning tools and algorithms have been implemented for categorization [16][17][18]. These algorithms help to classify documents into fixed categories or class, based on their content, after being trained on manually categorized documents. And now algorithms of this kind have also been used and developed to thread e-mail, and classify e-mail into spam and legitimate. A work on unlabeled e-mail based data is also done and then it operates an algorithm "Co-training on email and increase the performance of a classifier" [11]. Nowadays a lot of work is done by using Naive bayes classifier to solve the problem of spam [12][13][14][15].

But in this case, the performance of Naïve Bayes classifier was very poor, thus creating an open field to work for improving the performance of the classifier to label the emails [20]. An investigation of gender attribution mining from e-mail text documents is in [19]. It deals with a different aspect of categorizing the emails on the basis of gender with Support Vector Machine learning algorithm. Experiments were done using a corpus of e-mail documents which were generated by a large number of authors of both genders and it gave promising results for gender categorization. A work on categorizing spam and non-spam emails using the vector space model and feature extraction is done by using *k*-NN classifier [21]. Feature selection is also a major aspect in the classification of the mails and a recent work is depicted in [22]. This works shows that the need of a better spam filter is growing day by day and extracting the important features of the mails will surely help us to classify our mail as spam or non-spam.

## 3. Characteristics of Some Spam Mails

A detailed study of spam mails and non-spam mails of some Gmail accounts were done and it was found that a lot of important (some job related, IRCTC based and some other trustworthy) mails were sometimes classified as junks. Actually Gmail spam filter checks them and if the pattern of the mail was found auspicious considering some elements of the mail, then it classifies them as spam.

Some features which were observed during the study were:-

1. **The mail address of the sender**- If a job site (like monster.com) uses an absurd user name, then it is usually classified as spam. Not only user name, it also checks the domain name also, such as, if there is a job offer from a particular company then its email address must belong to the company's domain, but when it doesn't happen the spam filter classify it as spam.

2. **The content of the mail**- Gmail spam filter generally checks some of the writings in the content of the email like "Wowwww!!! Get some bonus prizes!!!" or "great discount offer" etc and due to this the mails are transferred in the spam box. There are also some emails where we find "Get richer now!!", "Hurrayy!!!! You have won" and many more similar contents which are also classified as spam.

3. **The subject line**- If a subject line consists of some symbols or if in the content the same thing is written in the same way as it is written in the subject line and no other sentence is written in the content then it is usually classified as spam. There are also some absurd subject lines like "I Will Need Your Help in This Situation I Find My Self", ":) Diwali Special HAPPY HOURS (:", "xxxxxx@gmail.com has indicated you're a friend. Accept?", etc.

4. **The language used**- If the message of the mail is in some other language which is not preferred by my Gmail language preference, then the mail is also classified as spam. Also, sometimes if we write something of other languages in English, then also it may be classified as spam.

5. **Users identified spam**- There were some mails from job sites for job offers which were detected as spam because many other users marked it as spam.

6. **Links provided in the message content**- Some of the mails contain only a link in the content which is written as "click here". Google spam filter finds it as spam since the link may redirect us to some malicious sites.

7. **The Content of the mail is written more than once**- There was one mail where the content of the mail is written twice in the body. It may be a factor responsible to be determined as spam.

8. **Empty content**- Some emails were classified as spam since the content of the mail was empty.

9. **Wrong reply to option**- In some of the mails, when we try to reply to the mail provided for replying a system generated mail occurs. The reason may be a wrong replying email address. The Gmail spam filter also checks it.

10. **Unknown Sender and Attachments**- In some of the mail sender's name or email address is not provided which is also a major point to classify the mail as spam. Some of the mail also contains some attachments which contain some unsafe contents, so the mail gets classified as spam.

### 3.1 Feature Extraction

Feature extraction or feature selection is generally applied to extract the important features or attributes from a large dataset. It is the method of selecting a subset of relevant features, which helps to construct a model for prediction. There are different feature extraction metrics but in our experiment we used **Chi Squared Statistics ($\chi^2$).** It is used to calculate the lack of independence between term and class, and compared to the $\chi^2$distribution with one degree of freedom. Its expression is given in Eq. (1).

$$\chi^2 = \frac{D \times (P_c E - NQ)^2}{(P_c + N) \times (Q + M) \times (P_c + Q) \times (N + M)} \quad (1)$$

Where D = total number of documents
$P_c$ = the number of document of a class containing a term
Q = the number of document containing term occurs without class
N= the number of document of a class occurs without a term.
M=the number of document of other class without a term.

### 3.2 Classifiers

- **Naive Bayes_**:- Naive Bayes classification is a supervised learning technique which is based on the Bayes theorem of probability. It says that, for a document D and a class C, the probability of a document D being in class C is examined as

$$P(C \mid D) = \frac{P(D|C)P(C)}{P(D)} \quad (2)$$

P(*D* | C) = probability of getting document *D* given class *C*,

*P(C)*= probability of occurrence of class *C*,

*P(D)*= probability of occurring document *D*. This can actually be ignored, since it is same for all classes.

- **Support Vector Machines** :- Support Vector Machines(SVM) is also a supervised machine learning algorithm that analyze data and recognize patterns, used for classification and regression analysis. It is a discriminative classifier which is defined by a separating hyper plane. Suppose for a given labelled training data, the algorithm produces an optimal hyper plane which categorizes the testing data.

- ***k*-Nearest Neighbors** :- *k*-Nearest Neighbors(*k*-NN) algorithms is one of the top data mining algorithm, which is used for classification and regression. It is a non-parametric lazy learning algorithm where the input consists of *k* closest training examples in the feature space [23][24]. The output of this algorithm is a class membership. An entity is classified by a majority vote of its neighbors, with the entity being assigned to the class which is most common among its k nearest neighbors.

- **Decision Tree** :- Decision tree is a classifier algorithm, which is expressed in the form of a tree structure. Where decision node specifies a test on a single attribute, leaf node indicates the value of the target attribute, edge is the split of one attribute and path indicates a disjunction of test to make the final decision. Decision trees classify instances by starting at the root node of the tree and moving through it until a leaf node occurs.

### 3.3 Dataset Information

The following table summarises the characteristics of the dataset used in our experiment.

Table 1: Characteristics of the datasets

| DATASETS→ | SMS spam collection (SPAMHAM) | Spambase Data Set | MyMail Data Set* |
|---|---|---|---|
| Number of Instances | 5572 | 4601 | 194 |
| Number of Terms | 6632 | 57 | 4466 |

***MyMail DataSet** :- This data set is a self created data set which was created by collecting 97 legitimate mails from inbox and 97 spam mails from spam box of my Gmail account. The inbox mails and the spam mails were collected from period of August 2013 to November 2013.

## 4. Proposed Methodology

The steps implemented for the experiment are explained below:-

(i) A spam based dataset is taken for implementation.

(ii) The dataset is then divided into two subsets. Training set, which contains 70% of the data and Testing set, which contains the remaining 30%.

(iii) In the first case, I applied different classifier on the training set and the prediction was done through the testing set without feature extraction.

(iv) The performances of different classifiers on the datasets were observed and are given in the result section.

(v) In the second level, we applied a feature extraction mechanism(Chi Squared metrics in our experiments) to extract the important features.

(vi) This feature selection mechanism reduces the training data by selecting only the non-zero attributes.

(vii) A formula is created which includes only those selected features.

(viii) Then we applied the classifiers for prediction on the extracted features of the training set and then the accuracy performance of the training set is calculated through a confusion matrix.

(ix) Then the prediction for the training set was applied in the testing set to finally predict the accurate class.

(x) The performance of this prediction model for all the dataset is given in the result section.

## 5. Result and Discussion

The accuracy rates obtained by applying different classification algorithms on the data sets without feature extraction are discussed in table 2.

Table 2: Accuracy rate of classification algorithms without feature extraction

| **Classifiers ↓** | **SPAMHAM** | | **Spambase** | | **MyMail** | |
|---|---|---|---|---|---|---|
| | Training | Testing | Training | Testing | Training | Testing |
| Decision Tree | 88% | 87% | 90% | 87% | 66% | 70% |
| Naive Bayes | 14% | 13% | 70% | 72% | 50% | 49% |
| Support Vector Machines | 91% | 90% | 93% | 90% | 90% | 91% |
| k-Nearest Neighbors | 92% | 90% | 97% | 89% | 94% | 91% |

The performances of different classification algorithms for different data sets without feature extraction are depicted in the following figure:

Figure 1: Performance rate of different classifiers without feature extraction

After feature extraction and then applying the classification algorithms on different data sets, the accuracy rate obtained is discussed in table 3 below.

Table 3: Accuracy rate of classification algorithms after feature extraction

| **Classifiers ↓** | **SPAMHAM** | | **Spambase** | | **MyMail** | |
|---|---|---|---|---|---|---|
| | Training | Testing | Training | Testing | Training | Testing |
| Decision Tree | 94% | 93% | 92% | 90% | 92% | 90% |
| Naive Bayes | 93% | 92% | 74% | 73% | 84% | 82% |
| Support Vector Machines | 99% | 95% | 96% | 93% | 98% | 95% |
| k-Nearest Neighbors | 99% | 96% | 100% | 90% | 100% | 97% |

The performances of different classification algorithms for different data sets after feature extraction are illustrated in the following figure.

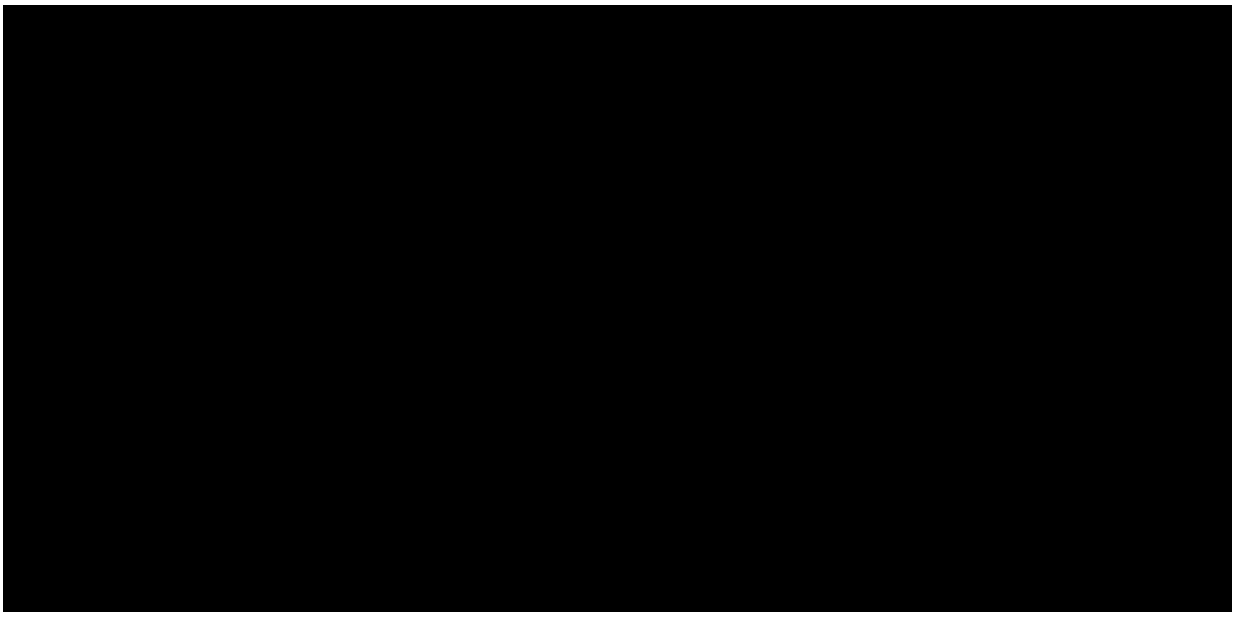

Figure 2: Performance rate of different classifiers with feature extraction

### 5.1 Result Analysis

The data sets which were used in the experiment are detailed in table 1. As described in the methodology, the experiment was conduct in two phases to analyse the performance rate. The result obtained by applying the classifiers on the datasets without feature extraction is shown in a tabular form in table 2 and a performance chart is also shown in figure 1. In the second phase, the result is obtained by extracting the important features of the data set by implementing Chi Squared($\chi^2$) feature selection method and then applying different classifiers to predict the accurate class. The prediction analysis in a table form is shown in table 3 and prediction performance chart is also shown in figure 2. From the results, it is clearly found that the use of a feature selection method enhanced the performance of the classifiers and the accuracy rate also increased.

## 6. Conclusion

The experiment results clearly show the effect of different classifiers for classifying a mail as spam or legitimate. The use of feature extraction also helped to identify the important attributes or features for classifying. It also increased the performance rate of the classifiers for the prediction of the class. So it is a better approach to build a spam filter using feature extraction. From the results it can also be concluded that the performance of SVM and k-NN classifiers were good enough than all other classifiers. So these classifiers will surely help us to implement a good spam filter. We could work out a spam filter which will directly access

an incoming mail online, remove the unnecessary URLs(if present) or features and determine it as a spam mail or legitimate mail.

## References

[1] Subramaniam, Thamarai, Hamid A. Jalab, and Alaa Y. Taqa. "Overview of textual anti-spam filtering techniques." *International Journal of the Physical Sciences* 5.12 (2010): 1869-1882.

[2] Blanzieri, Enrico, and Anton Bryl. "A survey of learning-based techniques of email spam filtering." *Artificial Intelligence Review* 29.1 (2008): 63-92.

[3] Wang, Xiao-lin. "Learning to classify email: a survey." *2005 International Conference on Machine Learning and Cybernetics*. Vol. 9. 2005.

[4] Wang, Zi-Qiang, et al. "An efficient SVM-based spam filtering algorithm."*Machine Learning and Cybernetics, 2006 International Conference on*. IEEE, 2006.

[5] MAAWG. Messaging anti-abuse working group. Email metrics report. Third & fourth quarter 2006. Available at http://www.maawg.org/about/ MAAWGMetric 2006 3 4 report.pdf Accessed: 04.06.07, 2006.

[6] Siponen, Mikko, and Carl Stucke. "Effective anti-spam strategies in companies: An international study." *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*. Vol. 6. IEEE, 2006.

[7] Moustakas, Evangelos, Chandrasekaran Ranganathan, and Penny Duquenoy. "Combating Spam through Legislation: A Comparative Analysis of US and European Approaches." *CEAS*. 2005.

[8] Kiritchenko, Svetlana, and Stan Matwin. "Email classification with co-training."*Proceedings of the 2011 Conference of the Center for Advanced Studies on Collaborative Research*. IBM Corp., 2011.

[9] Raad, Mostafa, et al. "Impact of spam advertisement through e-mail: A study to assess the influence of the anti-spam on the e-mail marketing." *Afr. J. Bus. Manage* 4.11 (2010): 2362-2367.

[10] Ying, K. O. N. G., and Z. H. A. O. Jie. "Learning to Filter Unsolicited Commercial E-Mail." *International Proceedings of Computer Science & Information Technology* 49 (2012).

[11] Kiritchenko, Svetlana, and Stan Matwin. "Email classification with co-training."*Proceedings of the 2011 Conference of the Center for Advanced Studies on Collaborative Research*. IBM Corp., 2011.

[12] Androutsopoulos, Ion, et al. "An evaluation of naive bayesian anti-spam filtering." *arXiv preprint cs/0006013* (2000).

[13] Metsis, Vangelis, Ion Androutsopoulos, and Georgios Paliouras. "Spam filtering with naive bayes-which naive bayes?." *CEAS*. 2006.

[14] Almeida, Tiago A., Jurandy Almeida, and Akebo Yamakami. "Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers." *Journal of Internet Services and Applications* 1.3 (2011): 183-200.

[15] Cormack, Gordon V., Mark D. Smucker, and Charles LA Clarke. "Efficient and effective spam filtering and re-ranking for large web datasets." *Information retrieval* 14.5 (2011): 441-465.

[16] Lewis, David D., et al. "Training algorithms for linear text classifiers."*Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1996.

[17] Sebastiani, Fabrizio. "Machine learning in automated text categorization." *ACM computing surveys (CSUR)* 34.1 (2002): 1-47.

[18] Dagan, Ido, Yael Karov, and Dan Roth. "Mistake-driven learning in text categorization." *Proceedings of the second conference on empirical methods in NLP*. 1997.

[19] Androutsopoulos, Ion, et al. "An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages."*Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2000.

[20] Yang, Zhen, et al. "An approach to spam detection by naive Bayes ensemble based on decision induction." *Intelligent Systems Design and Applications, 2006. ISDA'06. Sixth International Conference on*. Vol. 2. IEEE, 2006.

[21] Hamsapriya, T., and Ms D. Karthika Renuka. "Email classification for Spam Detection using Word Stemming." (2010).

[22] Elssied, Nadir Omer Fadl, Othman Ibrahim, and Ahmed Hamza Osman. "A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification." (2014).

[23] Zhang, Min-Ling, and Zhi-Hua Zhou. "ML-KNN: A lazy learning approach to multi-label learning." *Pattern recognition* 40.7 (2007): 2038-2048.

[24] Fdez-Riverola, Florentino, et al. "Applying lazy learning algorithms to tackle concept drift in spam filtering." *Expert Systems with Applications* 33.1 (2007): 36-48.

**Himadri Sekhar Atta** is from final year in Master of Technology in Computer Science and Engineering department of Institute of Engineering and Management, Kolkata, West Bengal. He passed his Bachelor of Technology degree in 2012 in Computer Science and Engineering department from Kanad Institute of Engineering and Management, Burdwan, West Bengal.