# Classification of Unstructured Data Using Machine Learning Algorithm

**Miss Sujata S. Damawale [1], Prof. Vipul V. Bag[2]**

[1]N.K.Orchid College of Engineering And Technology, Solapur.

[2]Associate Professor, N.K.Orchid College of Engineering and Technology, Solapur.

## Abstract

***Email is an essential thing in today's era, as emails are main source of communication. Emails are used on personal and corporate levels. Emails are used by number of people to interact between each other. Many email users receives legitimate emails and also unwanted emails. It becomes necessary to classify legitimate emails (HAM) from unwanted emails (SPAM). There are many Machine learning techniques now a day's used to automatically filter the spam e-mail. This paper proposes and reviews the classification technique. The purpose of proposed algorithm is to automatically classify mails into spam and legitimate message. The algorithm used for classification is support vector machine. The mails are classified on the basis of email text. The proposed algorithm is effective and reasonable method for email classification.***

**Keywords:** Ham, Spam, SVM, E-mail classification, Machine learning algorithms

## 1. INTRODUCTION

Email is essential thing in today's age. Recently unsolicited commercial / bulk e-mail also known as spam, become a big trouble over the internet. Spam is waste of time, storage space and communication bandwidth. The problem of spam e-mail has been increasing for years. In recent statistics, 40% of all emails are spam which about 15.4 billion email per day and that cost internet users about $355 million per year. Automatic e-mail filtering seems to be the most effective method for countering spam at the moment. Some published studies have examined spam detectors using Naive Bayesian approaches and large feature sets of binary attributes that determine the existence of common keywords in spam, and many commercial applications also use Naive Bayesian techniques.

Spammers began to use several tricky methods to overcome the filtering methods like using random sender addresses and/or append random characters to the beginning or the end of the message subject line [1]. But these evasive tactics are themselves patterns that human readers can often identify quickly. This work had the objectives of developing an alternative approach using a neural network (NN) classifier brained on a corpus of e-mail messages from several users. The features selection used, because the feature set uses descriptive characteristics of words and messages similar to those that a human reader would use to identify spam.

The paper proposes the classification model to classify mails into ham and spam mails. Basically, Classification is the process of finding model that describe and distinguishes data classes or concepts. The models are derived based on the analysis of set of objects for which the class label is unknown. Classification is a type of data analysis that extracts models describing important data classes. The paper proposes a system, which is used for classification of unstructured data in proper format. The Support Vector Machine Algorithm (SVM algorithm) is used in this system for classification of mails into spam and ham. The proposed system architecture contains mainly two phases, phase 1 for training of system with training email dataset using SVM algorithm and phase 2 for testing such that classification of real time mails.

The paper is organized as follows: section 1 is the paper introduction, section 2 summarize the related work done using machine learning algorithm, section 3 gives a general theoretical description, section 4 present detailed steps of the experiment implementation and finally closed the paper with the conclusion in section 5.

## 2. RELATED WORK

There are some research work that apply machine learning methods in e-mail classification.Muhammad N. Marsono, M. Watheq El-Kharashi, Fayez Gebali[2]demonstrated that the naïve Bayes e-mail content classification could be adapted for layer-3 processing, without the need for reassembly. Y. Tang, S. Krasser, Y. He, W. Yang, D. Alperovitch [3] proposed a system that used the SVM for classification purpose. This system extract email sender behavior data based on global sending distribution, analyze them and assign a value of trust to each IP address sending email message. The SVM classification is effective but still this system fails as spammer does not send spam mails from fixed IP address.

They always use different IP addresses for sending spam mails.

Yoo, S., Yang, Y., Lin, F., and Moon [1]developed personalized email prioritization (PEP) method that specially focus on analysis of personal social networks to capture user groups and to obtain rich features that represent the social roles from the viewpoint of particular user, as well as they developed a supervised classification framework for modeling personal priorities over email messages, and for predicting importance levels for new messages. Rasim M et al. [4] proposes a new method for clustering of spam messages collected in bases of antispam system is o□ered. The genetic algorithm is developed for solving clustering problems. The objective function is a maximization of similarity between messages in clusters, which is defined by k-nearest neighbor algorithm. Application of genetic algorithm for solving constrained problems faces the problem of constant support of chromosomes which reduces convergence process.

## 3. PROPOSED SYSTEM AND DESIGN

### 3.1 System Design

The proposed system is designed mainly to classify incoming mails into ham and spam class. Spam has serious negative on the usability of email and network resources. Spam is flooding the internet with many copies of the same message, in an attempt to force the message on people who would not otherwise choose to receive it. Hence, the filtration system is designed to filter ham mails from spam and making easy access of user to genuine mails.

The proposed system is designed and executed in two different phases. One phase is of one time training and second is of testing. The first phase is basically used to train the system with training dataset. In proposed system, “lingspam_public” mail dataset downloaded from csmining website [5] is used for training and testing. The used dataset contains mainly four directories, corresponding to four versions of the corpus, bare: Lemmatiser disabled, stop-list disabled. lemm: Lemmatiser enabled, stop-list disabled. lemm_stop: Lemmatiser enabled, stop-list enabled. stop: Lemmatiser disabled, stop-list enabled. Each one of these 4 directories contains 10 subdirectories (part1... part10). These correspond to the 10 partitions of the corpus that were used in the 10-fold experiments. In each repetition, one part was reserved for testing and the other 9 were used for training.

Each one of the 10 subdirectories contains both spam and legitimate messages, one message in each file. Files whose names have the formspmsg*.txt are spam messages. All other files are legitimate messages.

The training phase generates the word data with probability of ham or spam. The generated word data is stored in a database. The ham-spam word generated database is later on used in testing of incoming mail and accordingly classify them into ham and spam class. The testing phase is used to test both mail, the mail from “lingspam_public” mail dataset and live incoming mails in real time. The following architecture diagram represents the working structure of phases, training phase and testing phase.

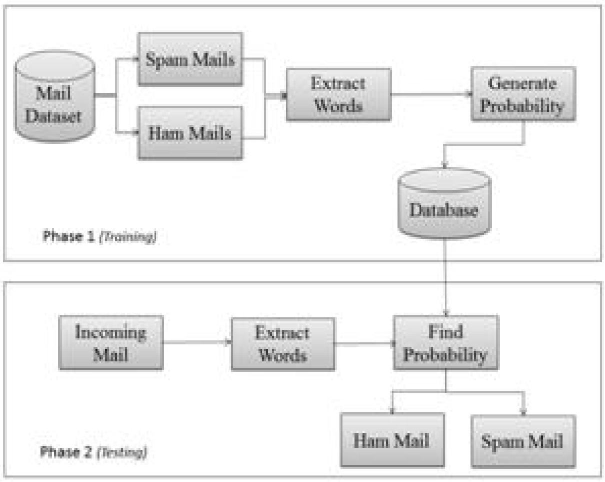


**Figure 1:** System Architecture

### 3.2 Implementation Methodology

As shown in architecture diagram, the proposed system is designed and executed in two different phases. First phase is of one time training of an application and second is one of testing of mail in real time. The training phase will train the system using “lingspam_public” mail dataset downloaded from csmining website for identifying ham and spam mails. While the testing phase will use spam-ham word database generated during training of an application to classify the real time mail into ham and spam class.

Consider subdirectorypart1 from each directory of dataset for testing while subdirectoriespart2 to part10 for training. First offal, consider training phase. Consider the mail database file from folders Part 2 to Part 10.Find out mail type as ham or spam from file name. Read mail file for each single word from file. If read out is a stop word, then ignore it such that remove word. As given in SVM, find out the occurrences (repetition) of word (feature dimension) in a mail document. Then set word probability as per mail type (+ve for ham and –ve for spam). Positive probability means word to be considered as ham and negative means to be consider as spam. From this analysis, the probability of this particular word(feature) as spam or ham word is finalized. Save this analysis result along with word and its corresponding probability to be use during testing of mail classification into ham / spam mails. Repeating the process for each word and then for each mail generates the training database for mail classification.

Now, consider the testing mail.Read each word from mail and remove if it is stop word. If extracted word is not a stop word then check this word against trained database of mail classification. The checking result gives the probability of word as ham or spam. Similarly analysis each words in mail database. Now, the average probabilities of mail document classify the mail document in either ham or spam.

### 3.3 Equations

The implemented system is worked in two different phase. First phase is of training and other is of testing live mails for ham and spam mail classifications.

For training phase, the methodology of occurrences count for ham and spam word is described using following equations.

The type of mail T[mail] is found out according to file type of input file from training folder. Then stop words [STOP_WORD] are removed from mail file and only remaining words [WORD] are considered for further training.

$$S(WORD)^{'} = \sum_{k=0}^{n} E[WORD] - [STOP\_WORD] \text{ ---- } (1)$$

Where, n = total number of words in a training mail file
S(WORD) = Set of words from training mail file which are not stop words

$$\text{O(WORD)} = \sum_{k=0}^{n} \begin{cases} 1, & WORD = HAM \\ -1, & WORD = SPAM \end{cases} \text{ ------- } (2)$$

Where, O(WORD) = Total occurrences of a word
The O(WORD) is stored in a record with corresponding word.

The above equation, equation no (2) gives total occurrence of a single word from a training mail file. The total occurrence is either positive or negative.

For testing phase, the methodology of probability generation for ham and spam word is described using following equations.

The probability of mail P(mail) is given by,

$$\text{P(mail)} = \sum_{k=0}^{n} \begin{cases} O(WORD), & W[mail] = WORD \\ 0, & W[mail]! = WORD \end{cases} \text{ - } (3)$$

Where, w[mail] = word extracted from mail
WORD is a word stored in a record with corresponding occurrence.

The equation no (3) gives overall probability of a mail by summation of occurrence of each word in a mail matching with a word record generated in equation 2. If word from mail is not matched with word record from equation no (3), then it occurrence of that particular word is considered as 0.

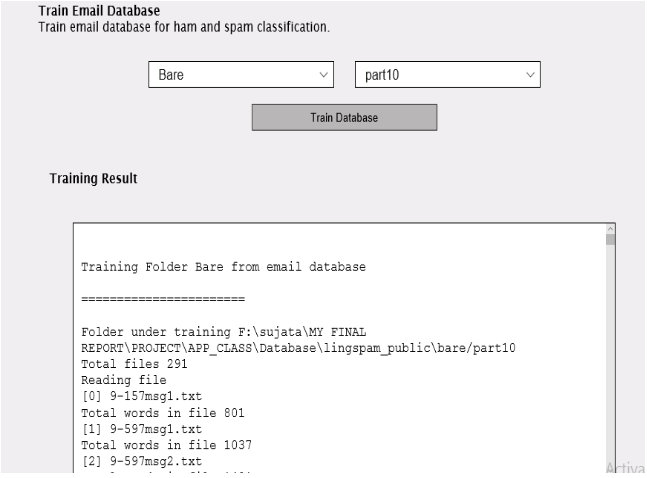
The type of mail T(mail) from probability of mail P(mail) is given by,

$$T(mail) = \begin{cases} spam, & P(mail) < 0 \\ ham, & P(mail) \geq 0 \end{cases} \text{ ------- } (4)$$

The equation no(4) gives mail type as per probability of mail. If P(mail) is positive then mail type is ham and if is negative then mail type is spam.

## 4. EXPERIMENTAL SETUP

The classification of unstructured data is done in proposed system. For classification the unstructured data i.e. emails are used, the dataset is downloaded from csmining website. These downloaded dataset is used for training of an application. For generating trained word database having probability of spam-ham words, the training phase is important. For training, user has to select the directory and corresponding subdirectory. Then system will read all mail files from selected subdirectory and generates corresponding word database. The following figure shows training of mails from subdirectory ‘part10’ from directory ‘Bare’. The subdirectory ‘part10’ overall contains 291 mail files including spam and ham mails.



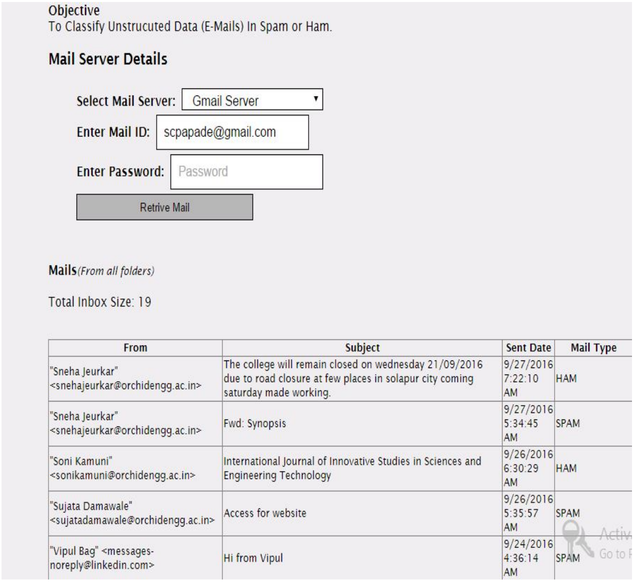
```
Training Folder Bare from email database

=======================

Folder under training F:\sujata\MY FINAL
REPORT\PROJECT\APP_CLASS\Database\lingspam_public\bare/part10
Total files 291
Reading file
[0] 9-157msg1.txt
Total words in file 801
[1] 9-597msg1.txt
Total words in file 1037
[2] 9-597msg2.txt
```

**Figure 2:** Training of an application

For testing of live mails in real time, user has to enter his/her mail login details. The login mail details contains, mail server, login mail id and password. The mail server is nothing but the POP3 server used for retrieving mail from user account. The application will retrieve mails from all folders such that, inbox, spam, trash, etc. Then application will analysis each downloaded mail from mail server for spam-ham classification. The real time mail classification such that testing phase uses the word database generated in training phase. Each word from incoming live mail is checked and removed if it is stop word. After that, checking remaining word against word database, each word is labeled as ham or spam word along with corresponding probability. At the end, such that after labeling all words, the overall probability gives the mail class as ham or spam. Following figure shows the live mail classification result.



Objective
To Classify Unstrucuted Data (E-Mails) In Spam or Ham.

Mail Server Details

Select Mail Server: Gmail Server
Enter Mail ID: scpapade@gmail.com
Enter Password: Password

Mails(From all folders)

Total Inbox Size: 19

| From | Subject | Sent Date | Mail Type |
|---|---|---|---|
| "Sneha Jeurkar" <snehajeurkar@orchidengg.ac.in> | The college will remain closed on wednesday 21/09/2016 due to road closure at few places in solapur city coming saturday made working. | 9/27/2016 7:22:10 AM | HAM |
| "Sneha Jeurkar" <snehajeurkar@orchidengg.ac.in> | Fwd: Synopsis | 9/27/2016 5:34:45 AM | SPAM |
| "Soni Kamuni" <sonikamuni@orchidengg.ac.in> | International Journal of Innovative Studies in Sciences and Engineering Technology | 9/26/2016 6:30:29 AM | HAM |
| "Sujata Damawale" <sujatadamawale@orchidengg.ac.in> | Access for website | 9/26/2016 5:35:57 AM | SPAM |
| "Vipul Bag" <messages-noreply@linkedin.com> | Hi from Vipul | 9/24/2016 4:36:14 AM | SPAM |

**Figure 3:** Classification of live mails into ham and spam class

## 5. PERFORMANCE EVALUATION

The performance of implemented system is calculated using confusion matrix. In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix[6], is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (or vice versa) [7].

The confusion matrix table shows original class and predicted class. The answer true positive (TP) is indicated by 1 if both original class and predicted classes are same such that ham. Similarly, answer is true negative (TN) when predicated class and original class, both is spam. The TN result is also indicated by 1. In case, predicted result is ham and original result is spam, then answer is false negative (FN) which indicated by 0. Similarly, false positive (FP) is indicated by 0 which comes when predicted result is spam and original result is ham [8].

For calculating, 9 test cases are considered. From these total 9 test cases, 8 case results are either TP or TN in combination while only 1 test case are comes as FP or FN in combination. Hence overall accuracy comes good. Following table shows predicated and original class along with confusion matrix result.

**Table 1:** Mail classification result along with TP or TN

| Subject | Org Class | Predicted Class | (TP) OR (TN) |
|---|---|---|---|
| 1,2,3& 4 BHK Apts in Bangalore starting at Rs.21 lacs . | SPAM | SPAM | 1 |
| Ride out of Town with Ola Outstation! | SPAM | SPAM | 1 |
| Pre Book the iPhone7 with full amount at Easy EMI and more | SPAM | SPAM | 1 |
| अपनी जन्म तारीख से जाने अपना भविष्य 2016 से 2020 तक | SPAM | SPAM | 1 |
| Upto 60% off on Car Insurance: Smart savings awaiting you! | SPAM | SPAM | 1 |
| Join 20,000+ peers and experts this October | SPAM | SPAM | 1 |

| | | | |
|---|---|---|---|
| 4 formal shirts @ Rs.1199 \| 6 Kurtas @ Rs.1499 \| 2 Free delivery coupons | SPAM | SPAM | 1 |
| Fw: HIOX INDIA : 10000 MB windows Hosting Package with Unlimited MB Bandwidth for scspcloud.com | HAM | SPAM | 0 |
| How to Identify Solid Stocks | SPAM | SPAM | 1 |

For table 1,
TP + TN = 8
Total = 9

Accuracy = (TP + TN) / TOTAL
Accuracy = 8 / 9 = 89
Hence, the analysis of implemented system using confusion matrix gives 89% accuracy.

## 6. CONCLUSION

The implemented system using SVM algorithm is successfully classified incoming mails in real time in two classes, ham or spam. The implemented algorithm SVM uses “lingspam_public” mail dataset downloaded from csmining website for one time training of system. The one time training is important as system does need to train again and again which helps in minimizing execution time of mail classification in real time and performance improvement. The accuracy of implemented system is checked using confusion matrix and its result is very good.

## References

[1] Cormack, Gordon. Smucker, Mark. Clarke, Charles " Efficient and effective spam filtering and re-ranking for large web datasets" Information Retrieval, Springer Netherlands. January 2011

[2] Muhammad N. Marsono, M. Watheq El-Kharashi, Fayez Gebali “Targeting spam control on middleboxes: Spam detection based on layer-3 e-mail content classification” Elsevier Computer

[3] Yuchun Tang, Sven Krasser, Yuanchen He, Weilai Yang, Dmitri Alperovitch ”Support Vector Machines and Random Forests Modeling for Spam Senders Behavior Analysis” IEEE

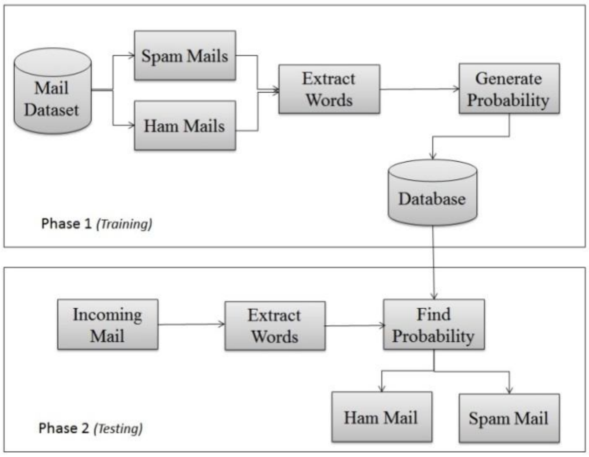
[4] Rasim M. Alguliev, Ramiz M. Aliguliyev, and Saadat A. Nazirova, “Classification of Textual E-Mail Spam Using Data Mining Techniques”, Hindawi Publishing Corporation Applied Computational Intelligence and Soft Computing Volume 2011, Article ID 416308, 8 pages doi:10.1155/2011/416308

[5] http://csmining.org/index.php/spam-email-datasets-.html

[6] Stehman, Stephen V. (1997). "Selecting and interpreting measures of thematic classification accuracy". *Remote Sensing of Environment*. **62** (1): 77–89. doi:10.1016/S0034-4257(97)00083-7.

[7] Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness& Correlation" (PDF). Journal of Machine Learning Technologies. **2**(1): 37–63.

[8] http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/



Architecture Diagram