Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3975 - 3992
Received 20 January 2021; Accepted 08 February 2021.

Spam filtering using Semantic and Rule Based model via supervised learning

S. Abiramasundari,
Assistant Professor, Department of Computer Science & Engineering, Srinivasa Ramanujan Centre,
SASTRA deemed university, Kumbakonam, 612001.
Dr. V. Ramaswamy,
Dean, Srinivasa Ramanujan Centre, SASTRA deemed university, Kumbakonam, 612001.
Dr. J. Sangeetha,
Assistant Professor, Department of Computer Science & Engineering, Srinivasa Ramanujan Centre,
SASTRA deemed university, Kumbakonam, 612001.
Corresponding author mail id.: abiraminathan_mcs81@src.sastra.edu

Abstract

Exchanging information over internet becomes a mandatory requirement for everyone. Electronic mail is an important medium through which all types of information can be shared. Despite the important role played by emails several challenges are also involved. Email carries not only legitimate messages but several times unwanted messages also. Spam email detection is challenging task since spammers use advanced technique through subject and email content. Emails with most fascinated words can be easily identified and classified as spam. But the professional way in which spammers manage to send messages makes it difficult for researchers to classify. Always, Spammers concentrate on either subject or content or both to persuade users to open the email. In this paper, Rule Based Subject Analysis (RBSA) and Semantic Based Feature Selection (SBFS) techniques are integrated with the Machine Learning algorithms. Various rules are outlined to check the subject field of the emails. Semantic based Feature selection technique is applied on the content of the email to reduce the features. RBSA and SBFS are integrated with four classifiers namely Support Vector Machine, Multinomial Naive Bayes, Gaussian Naive Bayes and Bernoulli Naive Bayes. The efficiency of the proposed techniques is tested on Enron dataset and it is observed that our proposed techniques with Support Vector Machine achieve lowest False Positive rate of 0.03.

Keywords: Spam detection, Email Classification, Rule based model, Semantic based learning, Machine Learning

1. Introduction

Internet is an important medium for users to share information. Email (Electronic Mail) is a powerful tool through which information can be exchanged. This innovation also introduces problems of mishandling. Hence the organization of emails is very essential so as to avoid such problems. Emails fall into two categories. Emails that are either ham (solicited) or spam (unsolicited). Ham emails are genuine or relevant emails originated from the known source. Spam emails on the other hand are unwanted mails which are created and transmitted to the users' inbox for various reasons. These spam emails are created and transmitted for marketing purposes, increasing website rank and stealing the personal information of users. These spam emails are flooding into the mail list of users and occupy more memory. This results in users spending their valuable time in deleting these unwanted emails. Hence it is essential that these spam emails are to be filtered and collected into Junk list. Numerous Spam filtering and Classification schemes are available today. Yet, the spam emails are drowning into users' inbox.

Email consists of header features and email body. Header includes From address, To address, Bcc, CC, Subject, Date and Time, Delivered To, Xmailer, Domain. Email body includes content, URL, Signature, Images, Attachment and other types of files. Spammers use any one of the above features to attract users in countless ways. But subject and content of the email plays an important role for describing the type of the email. Spammers do not always spread harmful emails. Sometime these emails are needed by the interested users. In most cases, users are not ready to spend their valuable time even to open these mails. Hence, these unwanted emails should be filtered into the junk box by the filtering system available in the server.

Many filtering systems have been designed based only on the header features. These systems only check the header features like From, To, Date and Subject in order to find the spam email; Researchers have developed filtering systems based on all the header features. But all the features are not useful in classification and results in memory wastage and more processing time. This has motivated the researchers to design systems with reduced feature set. Several Feature reduction techniques are available today. These techniques are used in various filtering systems to extract the most required features from the Email dataset.

Nowadays, various business associates such as Tourism packages, banking systems, Loan providers, Education providers etc., want to approach the users for their business development. These types of emails are created and sent using genuine sources. Hence a system which concentrates only on header features will not be able identify spam emails. Some filtering systems have been designed based on both header and email body. Even though this filtering system is able to identify spam emails, more time is consumed in processing header and email body.

In this paper, the classification system is designed based on the two main features namely subject and content of the email body. Subject reveals the purpose of emails that are sent by the spammers. Email body contains variety of information. Subject and email body are extracted and stored in a two dimensional array. Rule Based Subject Analysis (RBSA) is proposed for analyzing the spam terms in the subject field. Based on spam terms, every email is assigned a weight called spam_term_weight. Semantic Based Feature Selection (SBFS) is proposed to reduce the number of features required for classification process. The subject and content go through various functional units and get transformed into numerical values. These numerical values and weights are stored in a matrix and sent as an input to supervised learning algorithms. Email dataset is split into training and test dataset. The model is built using 80 percentage of training dataset. Same model is used for 20 percentage of test dataset. Four classification algorithms have been implemented for classifying the emails into spam and ham and their results are compared.

.

2. Related work

S. Venkatraman, B. Surendiran, P. Arun Raj Kumar(2019) have proposed Spam e-mail classification for the Internet of Things environment using semantic similarity approach. This paper addressed the problem of detecting spam emails based on the word similarity. Semantic based conceptual technique had been integrated with Naïve Bayes machine learning algorithm to detect the spam emails. This proposed algorithm is tested on Spambase, PU1, Enron corpus, and Lingspam email dataset[1].

SankeyVis: Visualizing active relationship from emails based on multiple dimensions and topic classification methods have been developed by Yong Fang , Cuirong Zhao , Cheng Huang , Liang Liu(2020) to discover the social relationships and to find the meaningful topics in emails. They have considered four parameters such as From, To, Date and Body of emails. Enron public email dataset is used in this study. First, the relationship between senders and receivers had been calculated. Then, Latent Dirichlet Allocation model is applied on the email body in order to generate the meaningful topics from the email. Once the topics are generated, forensic work had been conducted using Sankey diagram to visualize it. This paper observed that the Sankey Visualization helped in forensic analysis to acquire the power of evidence. It also improves efficiency and saves time[2].

Wenjuan Li, Weizhi Meng, Zhiyuan Tan, Yang Xiang (2019) have proposed The Design of multi-view based email classification for IoT systems via semi-supervised learning. Emails were classified using multi-view disagreement based semi supervised learning. The classification model is built in two phases. In the first phase, features are extracted from the email dataset and two attribute dataset were constructed based on the multi view. In the second phase, semi supervised learning is applied using labelled multi view instances for labelling unlabeled dataset. This methodology suggests that multi-view data construction improves efficiency compared to single view data[3].

Rushdi Shams and Robert E. Mercer (2013) classified emails using text and readability features. After extracting these types of features from the email dataset, various learning algorithms were applied. Random forest, Ensemble method, Bagging, Support Vector Machine and Naive Bayes algorithms were used in this paper. Various evaluation measures were used to evaluate the performance of the algorithms. This paper concluded that the combination of all features produced the optimum result [4].

Sunday OlusanyaOlatunji (2017) developed an improved email spam detection model based on Support Vector Machine. This paper used 57 features as predictor attributes and one feature as a target attribute that is a class label. Class label classified the emails as spam or ham. Implementation is done in MATLAB. Standard data base is used in this experiment. This paper concluded that the SVM model outperforms and provides an improvement of 3.11% over the NSA-PSO hybrid model [5].

N. Arulanand, K. Premalatha (2014) developed the Bin Bloom Filter using Heuristic Optimization Techniques for spam detection. The traditional bloom filter classifies the emails with high false positive rate. This rate is reduced in this paper by applying several hash functions. BBF was an improved version of bloom filter that was used to store the spam words with different weights. This work applied the various heuristic optimization techniques in BBF. The experimental result was finally compared with the traditional Bloom filter. Genetic algorithm produced the better result than other approaches [6].

RushdiShams Robert E. Mercer (2016) developed anti-spam filter named SENTINEL which applied five classifiers on the Enron-spam and Ling spam dataset using Natural Language attributes. The real valued and natural language attributes are extracted from the email text in order to generate binary classifiers. The classifiers are explored using five learning algorithms and evaluated with non-personalized email dataset CSDMC2010, SpamAssassin, Enron-Spam and LingSpam dataset. This paper uses BORUTA algorithm to measure the importance of the attributes and also used to compute scores for spam attributes. This paper suggests that the readability attributes are less important than Natural Language attributes and concludes that the ADABOOSTM1 and BAGGING perform better than Random Forest [7].

Eric Jiang (2006) has proposed Junk email filtering model 2LSI-SF based on augmented category LSI (Latent Semantic Indexing) spaces. He has categorized the emails into spam or ham by using their content. Classifier model is built using some discriminative information in the training dataset. Feature selection and message classification algorithms have been used to classify the emails into spam or ham. Finally, the results and the performance have been compared with SVM and NB. It is found that 2LSI-SF yields better performance. [8]

SarwatNizamani ,NasrullahMemon , MathiesGlasdam , Dong Duong Nguyen (2014) have evaluated the email detection method based on various feature sets. This paper measures the performance of the spam email detection method based on the various set of features extracted from the email. Cluster based classification model has been used to detect the spam email. Feature construction engine has been used to extract the features from the email dataset and then the required features are selected for further classification process. Different types of classification algorithms have been applied on the selected features using WEKA tool. This paper emphasizes the importance of feature set in the classification of emails. Spam emails could be detected using advanced features with an accuracy of 96% [9].

Son Dinh, TaherAzeb, Francis Fortin, DjedjigaMouheb, MouradDebbai have introduced Spam campaign detection, analysis and investigations for grouping the similar spam emails. Raw emails are sent through parsing engine and the set of features produced as output. The extracted features are content type, character set, subject, Email layout, URL tokens and attachment name. Spam campaign detection consists of two methods namely DFS and Incremental FP-Tree. In DFS traversing, various conditions are checked for considering the set of nodes that belong to the same campaign. FP-Tree is constructed dynamically using IFP-Tree technique. Once spam campaigns are identified, they are labelled using frequent words. [10].

AL-Rawashdeh, Ghada & Mamat, Rabiei & Abd Rahim, Noor Hafhizah. (2019) proposed Hybrid Water Cycle Optimization Algorithm with simulated Annealing for spam detection. The focus of this system is on reducing feature set for spam detection and improving the accuracy of feature selection in order to optimize the outcome. Cross-Validation is used for both training dataset and validation. The methodologies are tested on seven different dataset. Support Vector Machine classifier is applied for classification. The conclusion is that the Hybridization algorithm is better than Harmony search, Genetic algorithm and Particle swam algorithm. An accuracy of 96.3% is obtained. More than 50% of the features are reduced in this work [11].

M. Singh, R. Pamula and S. k. Shekhar have been proposed Email Spam Classification by Support Vector Machine. Non linear SVM classifier is used to classsiy the email dataset. This methodology considered Linear and Gaussian kernal funcitons on the SpamAssasin public dataset and emails from Gmail inbox. It is found that Linear kernel gives the highest level of accuracy for test dataset compared to Gaussian kernal function. Time needed for classifying test dataset using these two kernel function is same. [15]

Table 1: Review of Spam email detection approaches

Author	**Title**	**Proposed work**	**Algorithm**	**Features**	**Dataset**
Aakash Atul Alurkar, Sourabh Bharat Ranade et., all [16]	A Proposed Data Science Approach for Email Spam Classification using Machine Learning Techniques	ML used to detect Repetitive words for classification	Data preparation, Data analysis, Assessment and deployment	Cc/Bcc, domain and header	Enron and UCI (0.5M)

Amany A. Naem, Neveen I, Ghali, Afaf A. Saleh [20]	Antlion optimization and boosting classifier for spam email detection	Antlion optimization and Boosting	Support Vector Machine, K-nearest neighbor and Bagging	Vector created using tokens from email body	CSDMC2010 dataset
XiangHui Zhao, Yangping Zhang, Junkai Yi, 2016 IEEE.[17]	Statistical-based Bayesian Algorithm for Effective Email Classification	Using hash-spam, hash-legal and hash-probability tables.	Improved Bayesian Algorithm	Tokens from the body of email.	Chinese email
Aviad Cohen, Nir Nissim , Yuval Elovici 2018 Elsevier Expert Systems With Applications [18]	Novel set of general descriptive features for enhanced detection of malicious emails using machine learning methods	WEKA datamining tool is used. Feature selection: filter methods, wrapper methods, and embedded methods. But filter method is used.	J48 ,Random Forest (RF), Naïve Bayes ,Bayesian, Logistic Regression, LogitBoost, Sequential Minimal Optimization, Bagging and AdaBoost	Header, body and Attachments	Received email dataset from Virustotal
Muhammad Ali Hassan, Nhamo Mtetwa, 2018 IEEE [19]	Feature Extraction and Classification of Spam Emails	Bag-of-words and TF-IDF vectorization is used for feature extraction	SVM and Naïve Bayes	Email Body	Ling-spam dataset and Enron dataset

3. Proposed Technique

Our approach classifies the emails in an Enron dataset into spam or ham using subject and content features of the emails. Subject field describes the content of email. Spammers tempt the user to open their spam emails by including attractive words in the subject field. Subject is analyzed using RBSA for classifying emails. Rules are framed in the proposed technique RBSA by analysing subject field of the spam emails. Spam_term_weights are computed by applying these rules on subject. The words in subject are compared with the list of spam words. Content of the emails are also considered for classification. Entire content cannot be used for classification because it will contain lot of words. SBFS method is implemented to reduce the number of features. In this technique, number of words is reduced by eliminating meaningless words from the email content. CountVectorize( ) method is applied on the content of email to convert the text into numeric values. Classification algorithms such as Support Vector Machine [15], Multinomial Naive Bayes[12], Gaussian Naive Bayes and Bernoulli Naive Bayes have been applied on the output of CountVectorizer( ) method. The classification accuracy of these four algorithms is compared. This proposed approach includes five functional units such as Data collection, Data pre-processing, Rule Based Subject Analysis, Semantic Based Feature Selection and Classification.

Table 2: List of Subjects that include spam terms.

Subject	Spam term
Do you want to chat?	Includes question mark
CONGRATULATIONS!	Capitalized and also a spam word
Get Free COUPOUNS for Rs.500000	More number of spam words
Comple B.Tech, M.Tech, Diploma, B.Com, BBA, MCA any degree from recognized University	Lengthy subject
U aRe A lUcKy WiNnEr	Combination

3.1 Data collection and Pre-processing

3.1.1 Data collection

In this proposed scheme we used Enron email dataset. The dataset D consists of totally 5132 emails. Unnecessary columns have been removed from the dataset since we focus only on subject and content which in turn reduce the features. Label, Subjects, Text and Weight are stored as columns in a data frame for further processing. Data frame is denoted by DF (L, S, X, W) where L, S, X and W denote Label, Subject, Text and Weight respectively.

3.1.2 Pre-processing

Pre-processing is an important task in Natural Language Processing and in Text processing. Pre-processing is applied on contents of the emails to reduce the number of features. Generally text contains unwanted words, URL, HTML tags, multiple words with same meaning, meaningless words, special characters, numbers etc.,. These are removed from the given input text in order to improve accuracy of Machine Learning algorithms.

Spam is encoded as 1 whereas ham as 0. These encodings can be used to normalize the class labels. Since text labels are not comparable they should be converted into numeric labels.

Next, tokenization is applied on both Subject and Text columns. The content in the Subject and Text column is tokenized into words. Stop words are words such as “in”,” the”, “this”, “that”, “is” etc. These words should be ignored before applying Machine Learning techniques since these words appear more in quantity and also these words are not beneficiary in classification process. The removal of stop words helps in improving classification accuracy as ML algorithms work on less number of tokens.

3.2 Rule Based Subject Analysis

The main objective of RBSA is analyzing the subject field of the emails to find the spam terms. The most common spam terms are collected from Internet sources. This proposed technique uses several rules. These rules are formulated by referring to various spam emails. Spam_term_weight is computed for each and every email in the dataset based on these rules. The proposed Rule Based Subject Analysis (RBSA) algorithm helps in identifying spam emails efficiently.

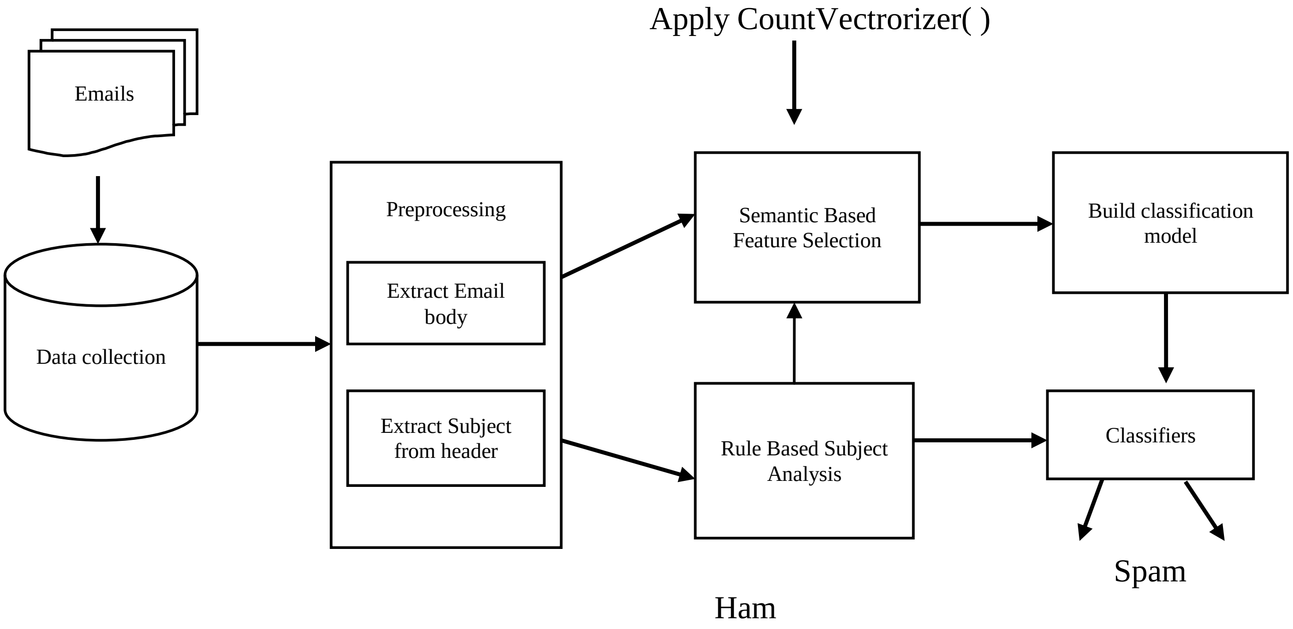
Apply CountVectrorizer( )
Emails
Data collection
Preprocessing
Extract Email body
Extract Subject from header
Semantic Based Feature Selection
Rule Based Subject Analysis
Build classification model
Classifiers
Spam
Ham

Fig.1. Proposed Architecture diagram

```
Algorithm 1:  Subject_Analysis
Input : DF(S); Subject field S in the data frame DF.
Output: Spam_term_weight of each email
For each row I in DF
    For each subject S in I
        For each token T in S
            If ((T is a Question)  and (T is spam word)) or  ((T is Capitalized)  and (T
        is spam word)) or ((T is AlphaNumeri) and (T is spam word)) or (length(T) > 25 )
        or (T is spam word)  then
                        Spam_term++
        End For
        //compute Spam_term_weight
        Spam_term_weight=Spam_term
    End For
End For
```

3.3 Semantic Based Feature Selection

This module plays an important role in our approach. In classification procedure the results can be enhanced by extracting relevant features. The required features for classification process are also reduced. This is achieved by removing meaningless words from the email contents. Reducing number of features improves classification accuracy.

Not all the words in the content are meaningful words. Meaningless words are eliminated from the content using Find_Meaning algorithm in order to achieve feature reduction. Since Machine Learning techniques are not properly executing on the non-numerical data, the text content must be transformed into numerical values. This module is used to prepare an input for the classifiers. Input is a matrix that contains token counts. CountVectorizer( ) method is used to transform the given textual content into matrix of token counts. This CountVectorizer( ) performs tokenization of the given text. It also constructs a vocabulary of familiar words and also encodes the new text documents by utilizing the constructed vocabulary.

```
Algorithm 2 :Find_Meaning
Input:  DF(X); Text field X in data frame DF.
Output: Text with meaningful tokens.
For each row I in DF
    For each Text X in I
        For each Token T in X
            Synonym_set=Syn_set (T)
                If (Synonym_set is null) then
                    Remove T
        End For
    End For
End For
```

3.4. Classification

The most important module in our work is Classification module. Four classifiers namely Support Vector Machine[22], Multinomial Naive Bayes[12], Gaussian Naive Bayes and Bernoulli Naive Bayes have been implemented to categorize the emails into spam or ham. This dataset has been split into 80 percent training dataset and 20 percent testing dataset.

Classification models using Support Vector Machine, Multinomial Naive Bayes, Gaussian Naive Bayes and Bernoulli Naive Bayes have been built for the training dataset. SVM classifier is first used to build a model for the training dataset. This classifier handles huge volume of dataset effectively. SVM is a supervised machine learning algorithm suitable for classification and linear regression. This algorithm works well both on the low and high dimensional spaces. It requires a line or plane to be drawn in order to separate the data points. This algorithm is suitable for pattern recognition, text classification and prediction. It is extensively used in solving many machine learning applications. Because of its fast working nature, it is used in our proposed approach for classifying the given set of emails into two classes namely spam or ham. This separation is done by drawing suitable hyper plane that also results in maximal margin.

Fig. 2 reveals binary classification since our dataset contains two classes. Emails are represented as data points (p) that are categorized into spam or ham based on the features extracted in the section 3.3. Spam emails are categorized above the hyper plane when Q=1 whereas ham emails are grouped below the hyper plane when Q=-1. We assume that our dataset contains N emails (data points). Emails are categorized into two groups as depicted in Fig. 2 using the linear function,

$$Q = Sp_n + B \quad where\ n = 1, \dots, N \tag{1}$$

In linear function, p_n denotes email from the dataset, Q is a prediction value, B is a constant and S is a vector. Prediction value can be 1, -1 or 0. The point (email) can be either 1 or -1(spam or ham) that is expressed as the following equations (2) and (3).

$$Sp_n + B \geq 1 \quad when\ Q = 1 \tag{2}$$

$$Sp_n + B \leq -1 \quad when\ Q = -1 \tag{3}$$

Finding largest margin of the hyper plane is an important process in SVM classification. The value of ‘S’ will decide the size of margin. Margin = 2/||S||. In order to maximize the margin size, the smallest ||S|| must be found. Once it is found, the emails are categorized as spam or ham based on the input features.

Received 20 January 2021; Accepted 08 February 2021.

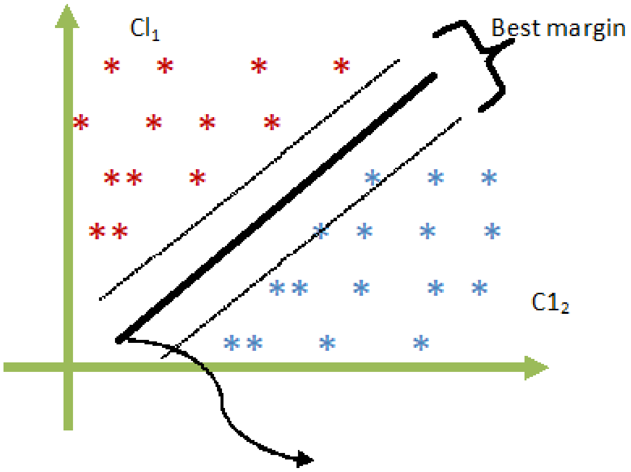

Fig. 2 Hyper plane that categorizes dataset into ham and spam

Naïve Bayes classifier[12] for multinomial model is built for the training dataset and email dataset is classified into spam or ham. Text classification can efficiently be performed by MNB classifier. This MNB classifier requires the word counts or frequency of the words as its input. Let M be a sparse matrix containing word count of the tokens in the email body and D be a training dataset, D =$\{e_1, e_2, ..., e_m\}$ which contains 'N' emails and 'm' attributes including class labels. Note that the emails can be classified into two classes namely spam and ham.
Cl ∈ {spam,ham}. Classifier will predict that the given email belong to the class Cl_i provided prob(Cl_i |N) is the maximum. In other words, prob(Cl_i |N) > prob(Cl_j|N), (j≠i).
Using Bayes Theorem, we have

$$\text{prob}(Cl_i|N) = \frac{\text{prob }(N|Cl_i)\ \text{x}\ \text{prob }(Cl_i)}{\text{prob }(N)} \tag{4}$$

The classifier predicts the class label C_i (i=1,2) by calculating prob(N|Cl_i) x prob(Cl_i). If prob(N|Cl_1) X prob(Cl_1) is maximum then the given email is classified as spam. Otherwise the email is classified as ham. The Emails in the training data set are described by the attributes (Labels, sparse matrix of tokens and Weight). Sparse matrix has word counts for the tokens. The number of tokens for every email in the data set is varied and can be represented as,
$e_i = \{t_{i1}, t_{i2}, \ldots, t_{ik}\}, \ldots, e_j = \{t_{j1}, t_{j2}, \ldots, t_{jl}\}$, where e_i and e_j are emails and $t_{i1}, t_{i2}, \ldots, t_{ik}, t_{j1}, t_{j2}, \ldots, t_{jl}$ are tokens in the emails. Bayesian classifier predicts the class label for N as either spam or ham depending on the priori probabilities.

Next, Gaussian Naïve Bayes model for the training dataset is built. The given set of emails is again classified into ham or spam. Following Gaussian distribution, the continuous values of every class are distributed. Note that this classifier works on continuous data. Test data are again classified into spam or ham by using Gaussian Naïve Bayes model which is already built.

Next classifier is Bernoulli Naive Bayes algorithm which performs similar to multinomial Naïve Bayes. Only difference is that Bernoulli algorithm uses the predictors as Boolean variables. In other words, the parameter values are either yes or no and we proceed for the prediction of class variables.

The classifier model is built for the training dataset using Bernoulli Naïve Bayes algorithm. The same model is also applied on the test dataset for classifying the Emails.

Four confusion matrices for four classifiers are constructed and listed in Table-9. The accuracy levels of the four classifiers are also compared. It is observed that Support Vector Machine gives the highest level of accuracy.

4. Experimental study and implementation

The performance of each of the four classifiers is computed using various metrics such as precision, recall, F-score and support. Confusion matrix is used to compute these measures. Four terms such as True Positive, True negative, False Negative and False Positive are used to generate Confusion matrix.

TN specifies number of ham emails correctly classified as ham.
TP specifies number of spam emails correctly classified as spam.
FN specifies number of spam emails classified as ham.
FP specifies number of ham emails classified as spam.

Precision is the relationship between true positive and predicted positive whereas recall is the ratio of true positives over all positives. Precision determines the percentage of spam emails actually predicted as spam among all predicted positives. Recall determines the percentage of spam emails actually predicted as spam among total number of spam emails predicted as ham and spam.

Precision = number of spam emails classified as spam divided by total number of positive predictions.

Recall = number of spam emails classified as spam divided by sum of True Positive and False Negative.

Precision and Recall are an important measure for our dataset. But Accuracy is not always a suitable measure for email dataset especially when the dataset is a skewed one.

F1-score is the mean value of precision and recall. Precision, recall and F1-score of training and testing dataset using four algorithms are listed in the Tables 5, 6, 7 and 8 respectively. False positive rate is an important measure which is calculated by using FP and TN. False Positive rate of the algorithms are visualized in Fig. 3.

$$FP_{rate} = FP/(FP + TN) \quad (5)$$

The percentage of misclassification is also computed using Error_rate. Generally 0.0 is a best and 1.0 is a worst rate of machine learning algorithms. Error rate of the algorithms are visualized in Fig. 4.

$$Error_{rate} = (FP + FN)/N$$
(6)

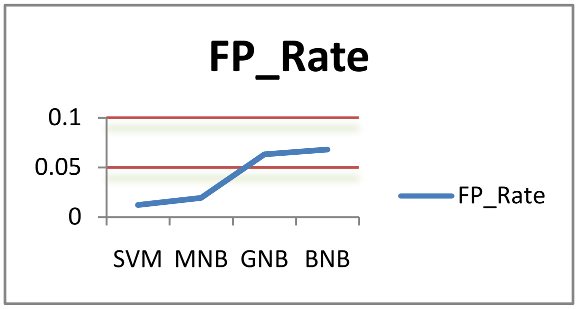

Fig. 3. False positive rate of SVM, MNB, GNB and BNB algorithms

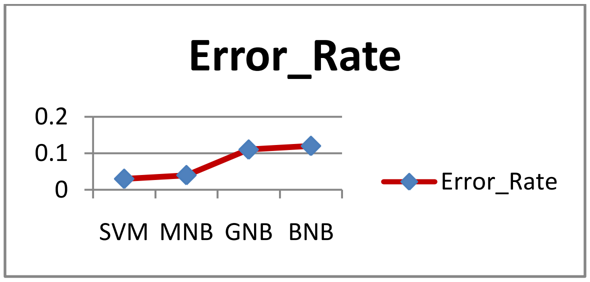

Figure - 4 Error_rate of SVM, MNB, GNB and BNB algorithms

The dataset is taken from Enron which consists of nearly 5132 emails. Unnecessary columns are removed. The labels are "spam" and "ham" which are converted into 1 and 0 respectively. Preprocessing techniques are then applied for removing special characters, stop words and numerals. Tokenization is applied to subject and content of the emails for splitting the sentence into words.

Table 3: Tokenization of Email content

Label	**Text**
0	[Go, join, point, crazy, Available, product]
0	[come, cool, Joking, wife, online]
1	[Free, entry, link, computer, win, Cup, fine]
0	[done, thank, say, early, hot, already]
0	[account, comment, think, goes, lives, around]

Subject_Analysis algorithm is executed for computing spam_term_weight for each of the emails by examining subject terms. Once calculated, it is stored under Weight column of data frame DF. Semantic Based Feature Selection is implemented in email contents to determine the meaningless word. Meaningless words are not important for classification process since they have low values. These words are removed from the content of emails to enforce feature reduction. To identify the spam words in the content of the email body, words with highest count are to be identified. Using CountVectorizer( ) method, the sparse matrix is generated. It is a numerical representation of the Text in the email dataset. This method not only determines word count but also preprocesses the text in order to extract the additional features. Hence the word count for all the words in the email dataset is calculated. Words with zero and minimum count in the documents are discarded. For this, min_df is taken as 0.25.

Table 4: Matrix representation of Text in the email dataset after applying CountVectorizer( )

	Go	join	come	Point	Crazy	computer	Product	account	free
e0	1	1	1	1	1	1	1	1	1
e1	0	0	0	0	2	2	1	2	1
e3	1	1	0	2	1	1	0	1	0
e4	0	0	0	0	0	0	0	1	1
e5	0	0	1	1	1	1	0	0	1

After obtaining the sparse matrix for the email dataset, the resulting dataset is divided into training dataset and test dataset in the ratio of 8:2. The classification model is next built for the training dataset using SVM and MultinomialNB[12] classification procedure. These algorithms yield output with 99% and 98% accuracy respectively. The test dataset is then applied on the same model which gives 97% and 95% accuracy, 0.03 and 0.04 error rate. These algorithms also produced lowest false positive rates 0.2 and 0.3. It is observed that SVM yields FP rate 0 for training dataset. None of the ham messages are classified as spam using SVM in training dataset. SVM produced better Precision and recall values for both training and test dataset.

Table 5: Classification report for training and test dataset using Support Vector Machine

Test dataset / Training dataset	Email	Precision	Recall	F1-score	Support
Training dataset	Ham	1	1	1	3926
Training dataset	Spam	1	1	1	409
Test dataset	Ham	0.98	0.99	0.98	987
Test dataset	Spam	0.86	0.76	0.81	97

Table 6: Classification report for training and test dataset using Multinomial Naïve Bayes.

Test dataset / Training dataset	Email	Precision	Recall	F1-score	Support
Training dataset	Ham	0.98	0.99	0.99	3926
Training dataset	Spam	0.93	0.83	0.88	409
Test dataset	Ham	0.97	0.98	0.98	987
Test dataset	Spam	0.78	0.69	0.73	97

Next model is built using Gaussian Naïve Bayes Model for the 80% of training dataset and then applied for the 20% test dataset. This algorithm gives 89% accuracy for training dataset and 88% accuracy for test data. Finally, the model is built using Bernoulli Naïve Bayes algorithm for training dataset. The same model is then tested for the test dataset. Accuracy of this algorithm for training and test dataset is 91% and 87%. False positive and error rate of this algorithm is 0.7 and 0.12 respectively.

Table 7: Classification report for training and test dataset using Gaussian Naïve Bayes

Test dataset / Training dataset	Email	Precision	Recall	F1-score	Support
Training dataset	Ham	0.94	0.97	0.95	3926
Training dataset	Spam	0.54	0.38	0.45	409
Test dataset	Ham	0.93	0.93	0.94	987
Test dataset	Spam	0.29	0.29	0.29	97

Table 8 Classification report for training and test dataset using Bernoulli Naïve Bayes

Test dataset / Training dataset	Email	Precision	Recall	F1-score	Support
Training dataset	Ham	0.93	0.95	0.94	3926
Training dataset	Spam	0.43	0.31	0.36	409
Test dataset	Ham	0.92	0.92	0.92	987

Test dataset	Spam	0.34	0.34	0.34	97

The results obtained from the four algorithms are analyzed and visualized in Fig. 5. It is observed that the Support Vector Machine algorithm gives the highest level of accuracy 97% compared to others.

Table 9: Confusion matrices of SVM, MNB, GNB and BNB

Classification Algorithm		Predicted as ham		Predicted as spam	
		Training dataset	Test dataset	Training dataset	Test dataset
Support Vector Machine	Actual ham	3926	975	0	12
	Actual spam	1	23	408	74
Multinomial Naïve Bayes	Actual ham	3899	968	27	19
	Actual spam	69	30	340	67
Gaussian Naïve Bayes	Actual ham	3761	924	165	63
	Actual spam	282	64	127	33
Bernoulli Naïve Bayes	Actual ham	3791	919	135	68
	Actual spam	252	69	157	28

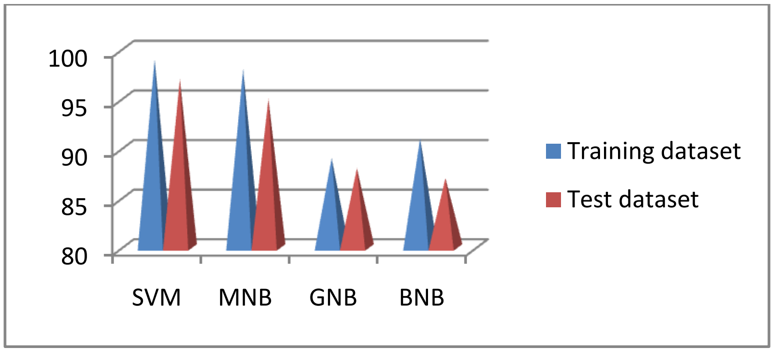

Fig. 5. Classification accuracy of SVM, MultinomialNB, GaussianNB and BernoulliNB algorithms

5. Comparative analysis

A semantic-based classification approach for an enhanced spam detection [13] have been proposed by Nadjate Saidani , Kamel Adi , Mohand Said Allili. In this paper, text based semantic analysis have been explored to improve the accuracy of spam detection in two levels. In the first level, emails are categorized by various domains such as education, health, finance etc., This methodology categorizes the emails based on the multiple domain. In the second level, manually collected features are combined with automatically extracted features to detect the spam emails in each domain. Spam emails are detected using KNN, Naïve Bayes, Decision Tree, Random Forest and Support Vector Machine algorithms. Naïve Bayes and Support Vector Machine algorithms yield an accuracy of 96% which is lesser than our methodology.

Content Based Spam Detection in Email using Bayesian Classifier [12] has been developed by Sunil B. Rathod, Tareek M. Pattewar. In this work, content of the email is used for classifying emails into spam or ham with accuracy 96%.

In our approach, subject and content of emails are used for classification since our methodology aims at increasing classification accuracy and reducing false identification.

6. Conclusions and Future work

Internet spam can be filtered in many ways. Considering the everyday escalation of spam and spammers, it is very important to offer effective mechanisms and implement competent software packages to handle spam. Various machine learning based classification algorithms are available for email classification. In this paper, classification model is built based on subject and content of the emails using four classification algorithms namely Support Vector Machine, Multinomial Naive Bayes, Gaussian Naive Bayes and Bernoulli Naive Bayes to classify the emails into ham or spam. Their performances are compared. Rule Based Subject Analysis is proposed for checking whether spam terms exist in the subject and also for computing spam_term_weight. Semantic based Feature Selection is used to reduce features. This proposed methodology is evaluated using four Machine Learning algorithms. The study indicates that SVM gives the highest level of accuracy. False identification is also reduced. The proposed algorithm can be refined using fuzzy logic for unlabeled records and employed on a Mail Server and Mail Client in future. We have also planned to investigate sources of spam emails and group them accordingly.

References

1. Venkatraman, s & B, Surendiran & Kumar, P.. (2019). Spam e-mail classification for the Internet of Things environment using semantic similarity approach. The Journal of Supercomputing. 76. 1-21. 10.1007/s11227-019-02913-7.
2. Yong Fang , Cuirong Zhao , Cheng Huang , Liang Liu. (2020). SankeyVis: Visualizing active relationship from emails based on multiple dimensions and topic classification methods. Forensic Science International: Digital Investigation. 35 (2020) 300981
3. Li, Wenjuan & Meng, Weizhi & Tan, Zhiyuan & Xiang, Yang. (2019). Design of multi-view based email classification for IoT systems via semi-supervised learning. Journal of Network and Computer Applications. 128. 56-63. 10.1016/j.jnca.2018.12.002.

4. Shams, Rushdi & Mercer, Robert. (2013). Classifying Spam Emails Using Text and Readability Features. Proceedings - IEEE International Conference on Data Mining, ICDM. 10.1109/ICDM.2013.131.
5. Olatunji, Sunday. (2017). Improved email spam detection model based on support vector machines. Neural Computing and Applications. 10.1007/s00521-017-3100-y.
6. N. Arulanand, K. Premalath (2014) Bin Bloom Filter Using Heuristic Optimization Techniques for Spam Detection International Scholarly and Scientific Research & Innovation 8(8).
7. Shams, Rushdi & Mercer, Robert. (2015). Supervised classification of spam emails with natural language stylometry. Neural Computing and Applications. 27. 10.1007/s00521-015-2069-7.
8. Jiang, Eric. (2006). Learning to Semantically Classify Email Messages. 10.1007/978-3-540-37256-1_86.
9. Nizamani, Dr. Sarwat & Memon, Nasrullah & Glasdam, Mathies & Nguyen, Dong. (2014). Detection of fraudulent emails by employing advanced feature abundance. Egyptian Informatics Journal. 15. 10.1016/j.eij.2014.07.002.
10. Dinh, Son & Azeb, Taher & Fortin, Francis & Mouheb, Djedjiga & Debbabi, Mourad. (2015). Spam campaign detection, analysis, and investigation. Digital Investigation. 12. 10.1016/j.diin.2015.01.006.
11. AL-Rawashdeh, Ghada & Mamat, Rabiei & Abd Rahim, Noor Hafhizah. (2019). Hybrid Water Cycle Optimization Algorithm with simulated annealing for Spam Email Detection. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2944089.
12. Rathod, Sunil & Pattewar, Tareek. (2015). Content based spam detection in email using Bayesian classifier. 1257-1261. 10.1109/ICCSP.2015.7322709.
13. Nadjate, Saidani & Adi, Kamel & Allili, Mohand. (2020). A Semantic-Based Classification Approach for an Enhanced Spam Detection. Computers & Security. 94. 101716. 10.1016/j.cose.2020.101716.
14. Eman M.BahgatSherineRadyWalaaGadIbrahim F.Moawad “Efficient email classification approach based on semantic methods” 2018 Ain Shans Engineering Journal.
15. M. Singh, R. Pamula and S. k. shekhar, "Email Spam Classification by Support Vector Machine," 2018 International Conference on Computing, Power and Communication

Technologies (GUCON), Greater Noida, Uttar Pradesh, India, 2018, pp. 878-882, doi: 10.1109/GUCON.2018.8674973.

16. Alurkar, Aakash & Ranade, Sourabh & Joshi, Shreeya & Ranade, Siddhesh & Sonewar, Piyush & Mahalle, Parikshit & Deshpande, Arvind. (2017). A proposed data science approach for email spam classification using machine learning techniques. 1-5. 10.1109/CTTE.2017.8260935.
17. Zhao, XiangHui & Zhang, Yangping & Yi, Junkai. (2016). Statistical-Based Bayesian Algorithm for Effective Email Classification. 636-639. 10.1109/ICISCE.2016.141.
18. Cohen, Aviad & Nissim, Nir & Elovici, Yuval. (2018). Novel Set of General Descriptive Features For Enhanced Detection of Malicious Emails Using Machine Learning Methods. Expert Systems with Applications. 110. 10.1016/j.eswa.2018.05.031.
19. Hassan, Muhammad & Mtetwa, Nhamoinesu. (2018). Feature Extraction and Classification of Spam Emails. 93-98. 10.1109/ISCMI.2018.8703222.
20. Naem, Amany & Ghali, Neveen & Saleh, Afaf. (2018). Antlion Optimization and Boosting Classifier for Spam Email Detection. Future Computing and Informatics Journal. 3. 10.1016/j.fcij.2018.11.006.