Email Spam Detection using Naive Bayes Classifier

Megha Tope

ME Student,
Computer Science and Engineering,
CSMSS College of Engineering, Aurangabad, India

***Abstract*: When the world wide web is expanding and spreading every day, email seems to be a reliable form of communication and is the fastest way to send information from one place to another. Now is the day where most transactions, whether they are general or business, are happening using email as a mode of communication. Email is an effective solution for communication because it helps in real time communication, which saves time and money. In addition to their advantages, e-mail has also been affected by spam attacks. Spam emails are often used to send bulk emails to the sender. Spam will flood the Internet with many in-depth copies of the messages. These messages will be sent to recipients who do not wish to receive anything else. We will analyze many data mining methods for spam data to find the best classification for email sorting.**

***Keywords*: E-mail spam, Classification, Feature Extraction, Naïve Bayesian Classifier**

I. INTRODUCTION

E-Mail is a powerful online communication mode because it saves money and helps reduce communication time, which makes it a favorite communication medium for personal communication and professional communications. Business E-mails provide easy data transfer, as well as original and other files that can be sent worldwide. There are many other cases where emails we send are affected by multiple attacks, which may be active or interactive. have We occasionally receive emails from unknown sources and have a few emails containing irrelevant content that is not important to the user. Spam Mails is a well known practice for sending unwanted or large data to a set of unique or random e-mail accounts. Spam Mail A subset of online spam related to the same or identical messages all sent to recipients by email. Spam includes some malware in scripts or other files that are executable and may harm the user's system. Most e-mail and spam lists are created by scanning the Usenet ad thoroughly by stealing the Internet email list. E-mail is an email that meets the following three criteria:

1) Anonymity: The address and identity of the sender are hidden.

2) Mass Mailing: Mail messages that are sent to a huge group of people.

3) Unsolicited: Email is not requested by recipients.

Spam has become a growing problem over the years. About 70% of all email is spam. As with web extensions, the problem of email spam is also growing as well. According to [1], it was found that an average of 10 days a year was compromised in spam processing. Spam is a costly issue which can cost a lot in the following years to lower bandwidth providers. Spam is an important issue that still attacks the presence of email. Therefore, it is very important to distinguish junk e-mail from various methods that are recommended to identify and classify e-mail messages as spam or non-spam or e-mail, and to find out the success rate of the algorithm. The speed of learning with the machine is very high. Many algorithms are processed for the classification of unsolicited emails that are widely used and analyzed among them as vector machines. Naïve Bayes Tree Deciding the classification of neural networks is a categorization. Well known In this article we have tried our algorithms: Naive Bayes, Bayes Net, Vector Machine Support (SVM), Tree Function (FT), J48, Random Forest and Random Tree.

In order to process and analyze the results of the categorization function, we need to use the attribute selection algorithm on the same data set. (The algorithm we use here is the first matching algorithm.) And we can do it. Use the selected category and entity. With the help of this study, we were able to find the identity of each category better when we chose our Best-First algorithm based feature compared to all the classifiers we got in. This data set makes Naive Bayes better. Result in the context of precision.

2.1 *Naïve Bayes*: A naive Bayes classifier is a simple probabilistic classifier with strong assumptions of independence. Simply put, a naive bayes classifier assumes that the presence / absence of a particular property of a class is not related to the presence / absence of any other feature, considering the class variable as a function of the Class Probability Model, Trained in a supervised learning environment. An advantage of the naive bayes classification is that it requires only a small amount of training data to estimate the parameters required for classification. The Bayesian classification assumes that the data belongs to a particular class. We then calculate the probability that the assumption is true. Bayesian classmates are basically statistical classifiers, that is, they can predict probabilities of class membership, such as the probability that a given test belongs to a particular class.

The naïve technique of Bayes is based on a Bayesian approach, it is therefore a simple, clear and fast classifier [4]. Before reaching the main term of the Bayes theorem, we will first analyze certain terms used in the theorem. P (A) is the probability that event A occurs. P (A / B) is the probability that event A occurs because event B has already occurred or we can define it as the conditional probability of A as a function of the condition that B has already occurred. The Bayes theorem is defined in Equation 1.

$$P(A/B) = P(B/A)\ P(A)\ P(B) \qquad (1)$$

If we consider OBJX as an object to be classified with the likelihood of belonging to one of classes CLS1, CLS2, CLS3, etc. By calculating Prob (CLSi / OBJX). Once these probabilities have been calculated for all classes, we simply assign OBJX to the class with the highest probability.

$$\text{Prob}(CLSi / OBJX) = [\text{Prob}(OBJX / CLSi)\ \text{Prob}(CLSi)] / \text{Prob}(OBJX) \qquad (2)$$

Where Prob (CLSi / OBJX) is the probability that the object OBJX belongs to a class CLSi, Prob (OBJX / CLSi) is the probability of obtaining attribute values OBJX if we know that it belongs to class CLSi Prob (CLSi) is the probability of an object belonging to a class CLSi without any other information, and Prob (OBJX) is the probability of obtaining OBJX attribute values regardless of the class to which the object belongs.

C. Classification and Prediction

Classification is the separation of objects into classes. If classes are created without looking at the data, the classification is known as a priori classification. If classes are created by looking at the data, the classification method is known as the later classification. When classifying, it is assumed that the classes have been considered a priori and the classification then consists in forming the system so that when a new object is introduced into the formed system it can affect the object to one of the existing classes. This approach is popularly known as the supervised learning process where in meaningful inputs are given to the system to learn for iterative process. Data classification can be divided into a two-step process as depicted the figure given below (refer Figure 2). In the first step, the model is constructed describing a predetermined set of data classes. The model is constructed by analyzing the database tuples described by the attributes. It is assumed that each tuple belongs to one of the existing classes, as determined by the class label attribute. The data tuples analyzed to construct the model collectively form the training set.

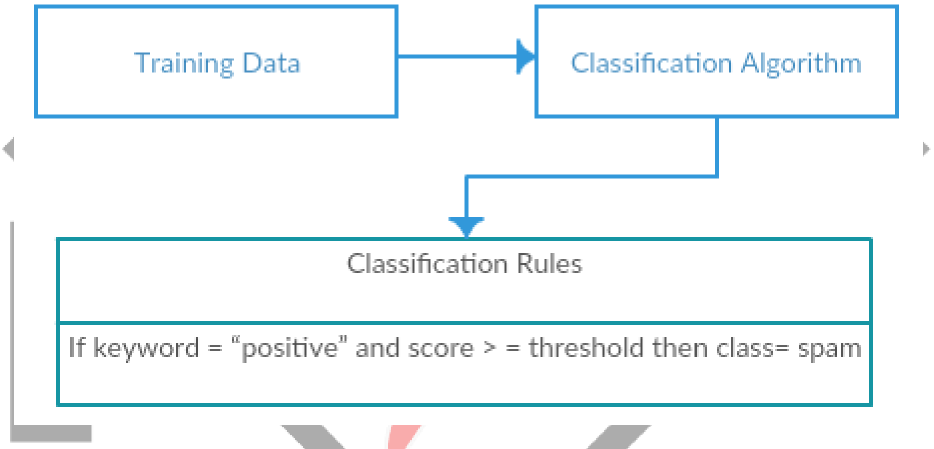

Training Data
Classification Algorithm
Classification Rules
If keyword = "positive" and score > = threshold then class= spam

Figure 2: Learning and Training of Classifier

Email ID	Department	Keywords	Class Label
abc@msn.com	Testing	Credit, Dollars	Spam
xyz@msn.com	HR	Weight loss, Herbal	Spam
ert@msn.com	Development	Lottery, Polling Results	Spam

In the second stage, as shown in Figure 3, the model is used for classification. The predictive accuracy of the model is evaluated first. The accuracy of a model in a given test data set is the percentage of samples of test sets correctly classified by the model. For each test sample, the known class tag is compared to the class prediction of the model learned for that sample.

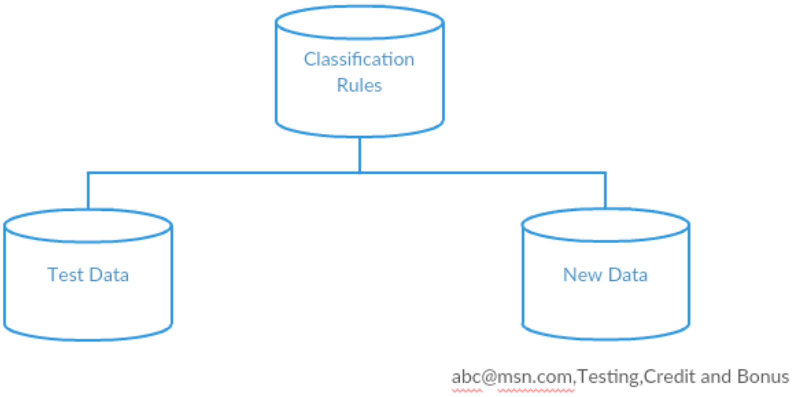


Figure 3: Classification Model

The prediction can be considered as the construction and use of a model to evaluate the class of an unmarked sample or to evaluate the ranges of values of an attribute that a given sample is likely to have. In this context, classification and regression are the two main types of prediction problems, where classification is used to predict discrete or nominal values, while regression is used to predict continuous or ordered values.

IV. PROPOSED WORK

Description: In this paper we are describing the method that is used to perform spam email classification. The first step is to select the data set file and apply the feature extraction technique for the extracted feature. For which we are using the word-count algorithm. The next step is to form the set of data that is extracted using the characteristic extraction technique. For the formation of the data we can calculate the probability of spam and not spam words in the document. The next step is to test the data with the help of Naïve Bayesian Classifier for which it calculates the probability of spam and non-spam mail and make a prediction whose value is greater. If spam words are larger than words that are not spam in an email, the mail is unwanted emails.

In the next step we are calculating the words that are misclassified by the classifier and we calculate the accuracy of the classifier and also calculate the classifier error rate by calculating the fraction of the word that is misclassified and the total number of words in the document.

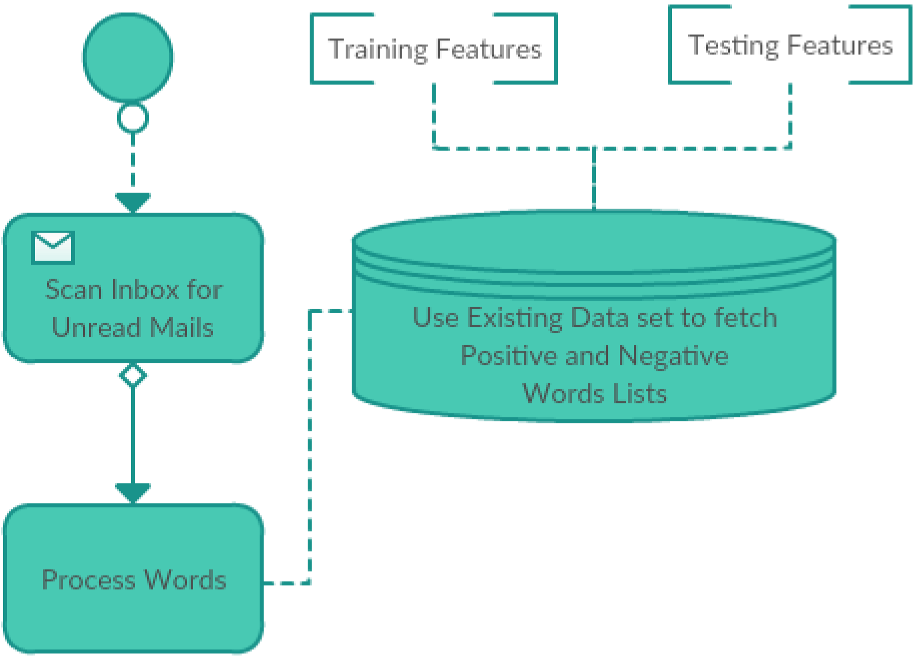

Training Features
Testing Features
Scan Inbox for Unread Mails
Use Existing Data set to fetch Positive and Negative Words Lists
Process Words

Figure 4: Word processing and classification for training using existing dataset

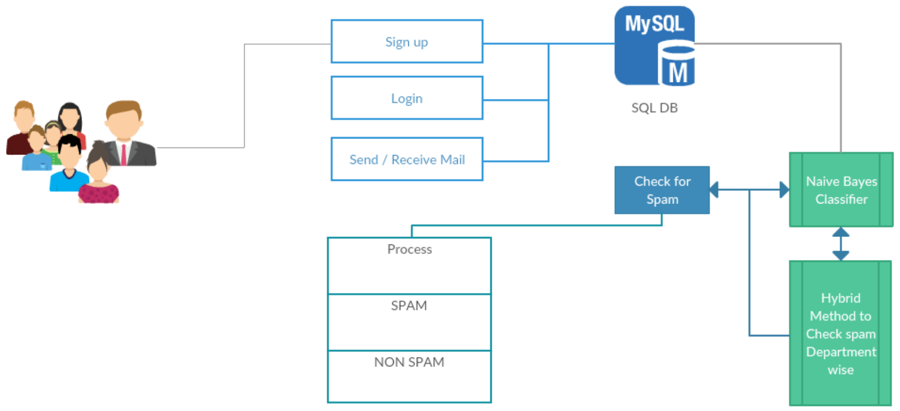


Figure 5: Proposed Architecture

PROPOSED ALGORITHM

Step 1: Select the email

Step 2: Extract features with help of tokenization and word count algorithm.

Step 3: Training the dataset with the help of Naive Bayesian Classifier.

Step4: Find the probability of spam and non-spam mails.
Prob_spam = (sum(train_matrix (spam_indices,)) + 1) ./ (spam_wc + numtokens)

Prob_nonspam = (sum(train_matrix(nonspam_indices,)) + 1) ./ (nonspam_wc + numtokens)

Step 5: Testing the dataset

log_a = test_matrix*(log(prob_tokens_spam))' + log(prob_spam)

log_b = test_matrix*(log(prob_tokens_nonspam))'+ log(1 - prob_spam)

if

output = log_a > log_b then document are spam

else the document are non-spam

Step 6: Classify the spam and non-spam mails.

Step 7: compute the error of the text data and calculate the word which is wrongly classified

Numdocs_wrong = sum(xor(output, text_lables))

Step 8: display the error rate of text data and calculate the fraction of wrongly classified word

Fraction_wrong = numdocs_wrong/numtest_docs

V. RESULT & DISCUSSION

As part of this project, we explain the classification of unwanted emails to identify spam and not spam. For this we use the Naïve Bayesian Classifier. In this project, we created an e-mail classification system to classify spam and not spam. To do this, we created a custom dataset to run this experiment. In our data set, we took a total of 960 emails in which 700 train data and 260 test data. Of the 700 data streams, 350 are spam and 350 are non-spam. Similarly, the test data set 260 contains 130 spam e-mails and 130 non-spam e-mails.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$FMeasure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Here we are present reading different for the four sets of data formed that are tested by the classifier, i.e. naive Bayesian Classifier and Support Vector Machine. Hence the different readings and calculation of the result:

Train Dataset	Support Vector Machine	Naive Bayes Classifier
Dataset 100	68	78.66
Dataset 200	66.23	79.12
Dataset 700	65.11	81.34

Table: Reading of different Classifiers

This reading contains the text data classified by the classifier and provides the word that is misclassified and the classifier error rate. Therefore, we can show the overall result provided by the classifier. And to say that the naive Bayesian classifier classifies most of the word accurately. When the number of data sets is increased, the Bayesian naive classifier produces a better result compared to the vector machine.

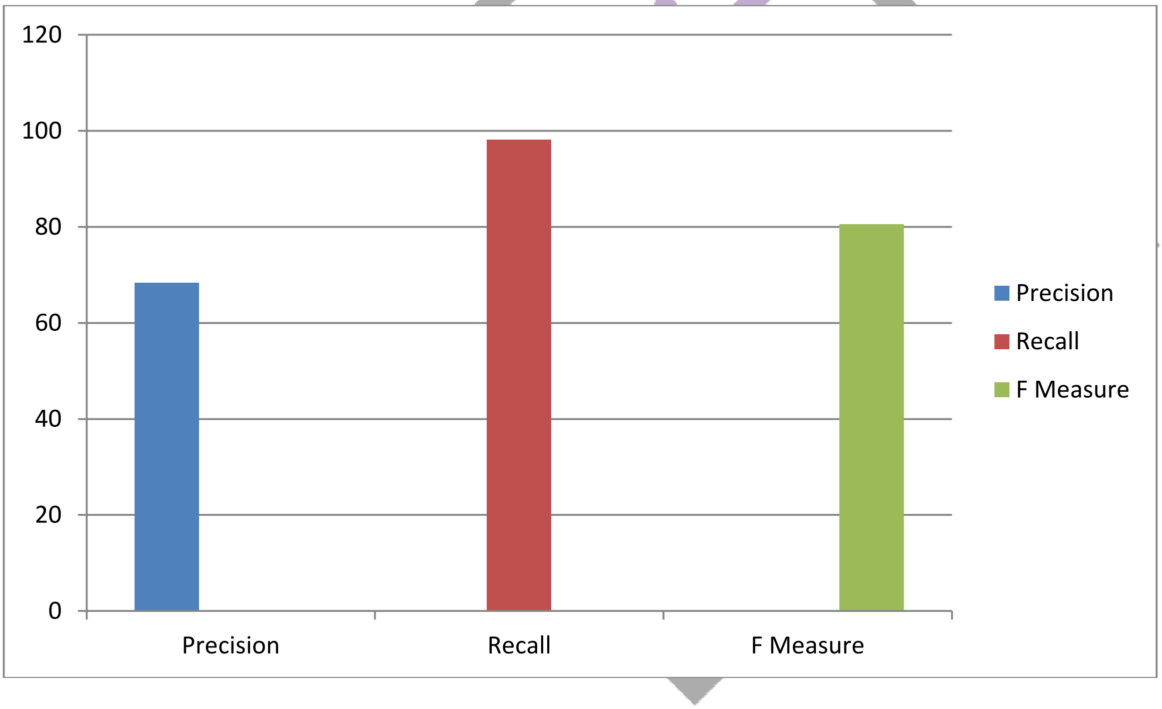
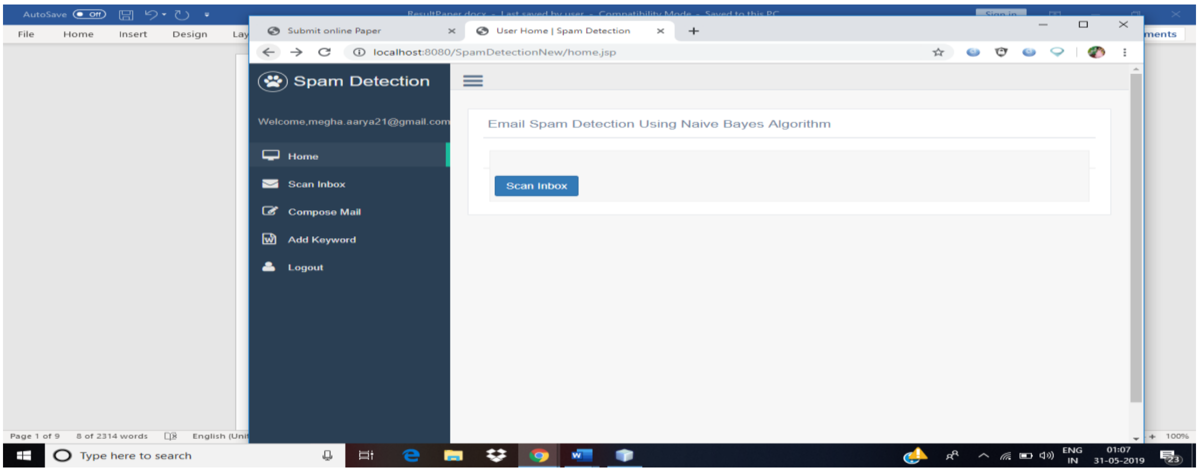


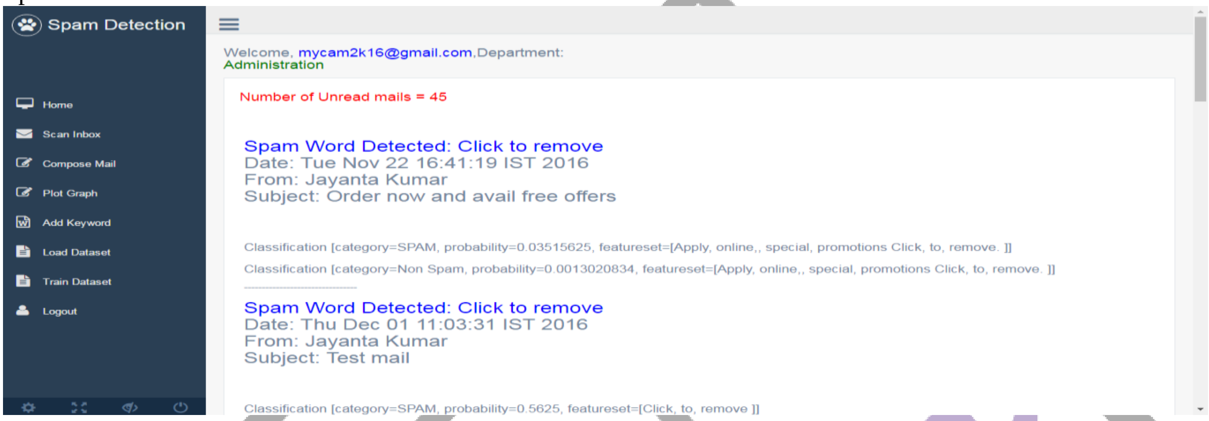
Figure 6: Graphical Representation of Performance Metrics

Therefore, we can show the graph to display the accuracy and error rate of both the Naive Bayesian Classifier and Vector Machine.
Result:-Login



Email Spam Detection Using Naive Bayes Algorithm

Spam Detected:-



Welcome, mycam2k16@gmail.com,Department: Administration

Number of Unread mails = 45

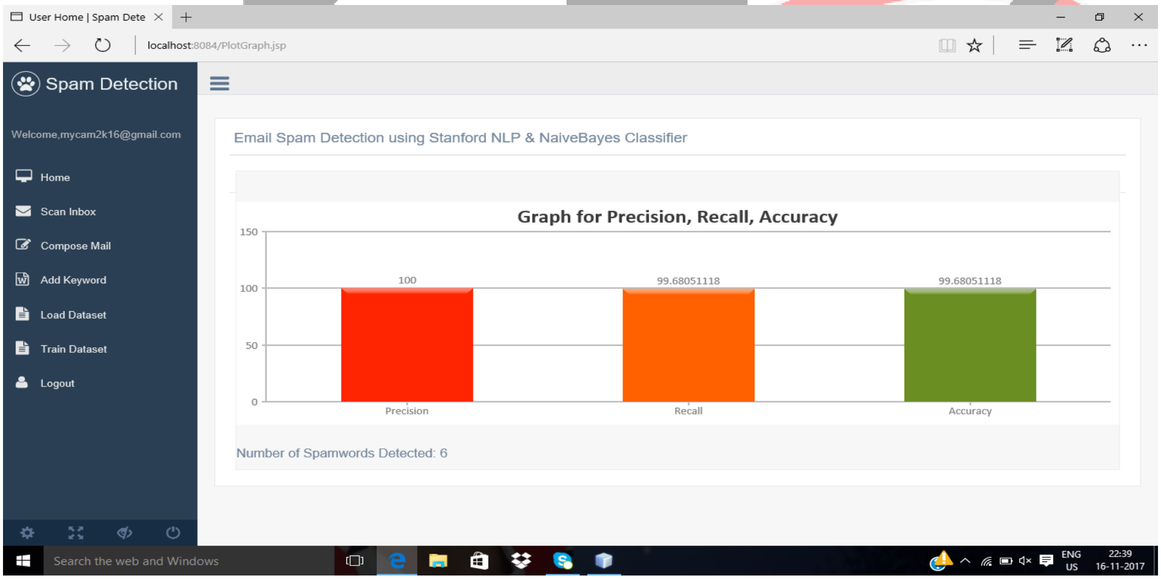
Spam Word Detected: Click to remove
Date: Tue Nov 22 16:41:19 IST 2016
From: Jayanta Kumar
Subject: Order now and avail free offers

Classification [category=SPAM, probability=0.03515625, featureset=[Apply, online,, special, promotions Click, to, remove.]]
Classification [category=Non Spam, probability=0.0013020834, featureset=[Apply, online,, special, promotions Click, to, remove.]]

Spam Word Detected: Click to remove
Date: Thu Dec 01 11:03:31 IST 2016
From: Jayanta Kumar
Subject: Test mail

Classification [category=SPAM, probability=0.5625, featureset=[Click, to, remove]]

Result Graph:-



Email Spam Detection using Stanford NLP & NaiveBayes Classifier

Number of Spamwords Detected: 6

VI Conclusion:

In the classification of spam, the main source of concern is the classification of e-mail and unwanted threats. So today, most researchers are working in this area to find the best classifier to detect spam. Therefore, you need a filter with great precision to filter out spam mails or spam mails. In this article, we have focused on finding the best classifier for spam classification using data mining techniques. Therefore, we apply different classification algorithms in the given input data set and verify the results. In this study, we analyzed that classifiers work well when we incorporate the feature selection approach into the classification process,

that is, accuracy is dramatically improved when classifiers are applied to the classifier, Dataset rather than the data set.
We use the Naïve Bayes Classifier here and extract the word using the word count algorithm. The error rate is very low when we use the Bayesian Naïve Classifier. It can be said that Naïve Bayesian Classifier produces a better result than Support Vector Machine.

REFERENCES

[1] Masurah Mohammad, Ali Selman “An Evaluation on the Efficiency of Hybrid Feature Selection in Spam Email Classification”, IEEE,2015.
[2] C. Bala Kumar, D. Ganesh Kumar “A Data Mining Approach on Various Classifiers in Email Spam Filtering”, IJRASET, May 2015
[3] Vinod Patidar, Divakar Singh, Anju Singh "A Novel Technique of Email Classification for Spam Detection ", International Journal of Applied Information Systems (IJAIS), Volume 5 – No. 10, August 2013.
[4] Cormack, Gordon. Smucker, Mark. Clarke, Charles "Efficient and effective spam filtering and re-ranking for large web datasets" Information Retrieval, Springer Netherlands. January 2011
[5] Archit Mehta ,Raunakraj Patel “Email Classification using data Mining”, IJARCCE, 2011.