ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijareeie.com

Vol. 6, Issue 10, October 2017

An Evaluation of Naïve Bayesian Classifier for Anti-Spam Filtering Techniques

Deepika Mallampati

Assistant Professor, Dept. of Computer Science and Engineering, Sreyas Institute of Engineering & Technology, Hyderabad, Telangana, India

ABSTRACT: An efficient anti-spam filter that would block all spam, without blocking any legitimate messages is a growing need. To address this problem, we examine the effectiveness of statistically-based approaches Naïve Bayesian anti-spam filters, as it is content-based and self-learning (adaptive) in nature.To solve thisproblem the different spam filtering technique is used. Thespam filtering techniques are used to protect our mailbox forspam mails. In this project, we are using the Naïve BayesianClassifier for spam classification. The Naïve BayesianClassifier is very simple and efficient method for spamclassification. Here we are using the Lingspam dataset forclassification of spam and non-spam mails. The featureextraction technique is used to extract the feature.

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

I. INTRODUCTION

The problem of unsolicited bulk e-mail, orspam, gets worse with every year. The vastamount of spam being sent wastes resourceson the Internet, wastes time for users, and mayexpose children to unsuitable contents (e.g.pornography). This development has stressedthe need for automatic spam filters.Early spam filters were instances of knowledge engineering and used hand-crafted rules(e.g. the presence of the string “buy now” indicates spam). The process of creating the rulebase requires both knowledge and time, andthe rules were thus often supplied by the developers of the filter. Having common and, moreor less, publicly available rules made it easyfor spammers to construct their e-mails to getthrough the filters. The difficulty in eliminating spam lies in differentiating itfrom a legitimate message. However, the message content ofspam typically forms a distinct category rarely observed inlegitimate messages, making it possible for text classifiers tobe used for anti-spam filtering. The goal of this research is toexamine the effectiveness of Naïve Bayesian anti-spam filtersand the effect of parameter settings on the effectiveness ofspam filtering. Additionally, we look at a novel modificationto existing filters and incorporate it into the evaluation.
Mail filters have differing degrees of configurability. Once ina while they settle on choices taking into accountcoordinating a consistent expression. Different times,essential words in the message body are utilized, or maybethe email location of the sender of the message. Some morepropelled channels, especially hostile to spam channels, usemeasurable archive order methods, for example, the guilelessBayes classifier. Picture sifting can likewise be utilized thatutilization complex picture examination calculations toidentify skin-tones and particular body shapes typicallyconnected with obscene pictures.Mail filters can be introduced by the client, either asindependent projects (see interfaces underneath), or as amajor aspect of their email project (email customer).In email programs, clients can make individual, "manual"channels that then naturally channel mail as indicated by thepicked criteria. Most email projects now likewise have aprogrammed spam separating capacity. Network accesssuppliers can likewise introduce mail channels in their mailexchange operators as a support of the greater part of theirclients. Because of the developing danger of fake sites,Internet administration suppliers channel URLs in emailmessages to uproot the risk before clients click.Normal uses for mail filters incorporate arranging incomingemail and evacuation of spam and PC infections. A less basicutilization is to investigate active email at a feworganizations to guarantee that workers consent to properlaws.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijareeie.com

Vol. 6, Issue 10, October 2017

Clients may additionally utilize a mail filter to organizemessages, and to sort them into organizers in light of topic orother criteria.

II. BACKGROUND STUDY AND RELATED WORK

There has been numerous numbers of studies on activelearning for text classification using machine learningtechniques [9]-[11], probabilistic models [12], [13]. Thequery by committee algorithm (Seung et al. 1992, Freund etal., 1997) used priori distribution than hypothesis. Thepopular techniques for text classifications are decision trees[14], [15], Naïve Bayes [14]-[16], rule induction, neuralnetworks [14]-[16], nearest neighbors and later on SupportVector Machine [17]. Though there is lot of techniques andalgorithms which have been proposed so far, the textclassification is not yet accurate and faultless and still indemand of improvement.

Web spam which is a major issue throughout today's websearch tool; consequently it is important for web crawlers tohave the capacity to detect web spam amid creeping. TheClassification Models are designed by machine learningorder algorithm. [2] The one machine learning algorithm isNaïve Bayesian Classifier which is also used in [1] toseparate the spam and non-spam mails. Big Data analyzingframework which is also outline for spam detection.Extricating the feeling from a message is a method for get thevaluable data. In Machine learning innovations can gainfrom the preparation datasets furthermore anticipate thechoice making framework hence they are broadly utilized asa part of feeling order with the exceptionally precision offramework. [3]

Email Spam is most crucial matter in a social network. Thereare many problem created through spam. The spam isnothing this is unwanted message or mail which the end userdoesn’t want in our mail box. Because of these spam theperformance of the system can be degraded and also affectedthe accuracy of the system. To send the unsolicited orunwanted messages which are also called spam is used inElectronic spamming. In this project explain about the emailspam, where how spam can spoil the performance of mailingsystem. In the previous study there are many types of spamclassifier are present too detect the spam and non-spammails.

There are different email filtering techniques are also used inspam detection. Mostly popular filters or classifier are:Decision tree classifier, Negative Selection Algorithm,Genetic Algorithm Support Vector Machine Classifier,Bayesian Classifier etc. From the previous study we identifythat Support Vector Machine (SVM Classifier) are used foremail spam classification. But it takes very much time fordetecting spam. The SVM Classifier has also wronglyclassified the messages. So the system can be on a risk. Theerror rate of SVM Classifier is very high. In this project thereis also discussion in the Feature Selection process. There aredifferent feature extractions techniques are present which areused in extracting the messages.

Solution of the Problem: To solve the problem of previous study in this project we areuses the Naive Bayesian Classifier for classify the spam andnon-spam mails. The naive Bayesian Classifier is one of themost popular and simplest methods for classification. Naïve Bayesian Classifiers are highly scalable, learning problemthe number of features are required for the number of linearparameter. Training of the large data simple can be easilydone with Naive Bayesian Classifier, which takes a very lesstime as compared to other classifier. The accuracy of systemis increase using Naïve Bayesian Classifier.

III. METHODOLOGY

E-mail spam classification has major issue in today’selectronic world. To solve this problem the different spamclassification methods are used. Using this spam detectiontechnique we can identifies the spam and non-spam mails inour mailbox. In this work we are using the Naïve BayesianClassifier for email spam classification.

In this work also use feature extraction techniques forproviding efficient dataset. The feature extraction techniquesare used when the input data is too large and it is redundantin nature so feature is extracted to obtain an accurate result.In this work we are using the word-count algorithm forextracting feature from the dataset.

Here we use theLingspam data set which contains total 960 mails in which700 are train dataset and 260 are test dataset. The train andtest data are further divided in two parts spam mails andnon-spam i.e. 50% of train dataset are spam dataset and 50%are non-spam dataset as same for the test dataset.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijareeie.com

Vol. 6, Issue 10, October 2017

Step 1: Select the file from the dataset.
Step 2: Pre-process the file and removing the stop-word.
Step3: Count the total word of the file and find the uniqueword of that file.
Step 4: Calculate the frequency of words.
Step 5: Make a dictionary and store the file path.
Step 6: Extracted Feature.

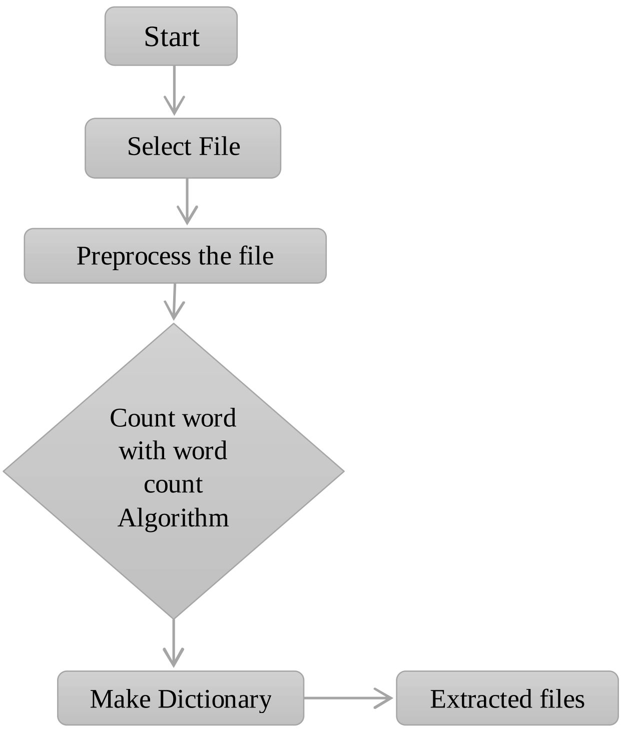

Figure.1 Feature Extraction Method

Description: In this feature extraction word-count algorithmin which three steps are present they are pre-processing,count the word, make dictionary. The first is to select the filefrom the dataset. Then second pre-process the data in whichfirst remove the stop word and non-words from thedocument. The third step for feature extraction is to count theunique word from total number of words. So we can calculatethe frequency of that word in a document. The forth step is tomake a dictionary and store the path of document this cansolve the redundancy problem. The extracted data arereceived after all steps are complete.
The proposed methodology:

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijareeie.com

Vol. 6, Issue 10, October 2017

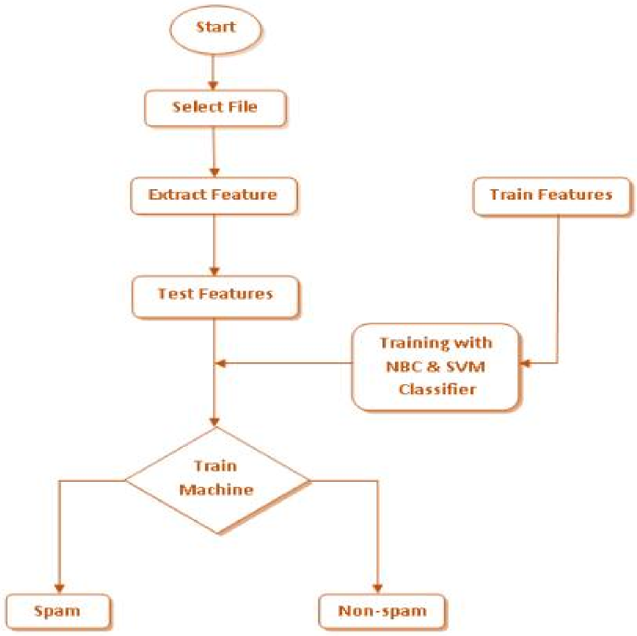

Figure.2 The Proposed MethodologyAlgorithm:

Step 1: Select the file
Step2: Extracting the feature with help of word countalgorithm.
Step 3: Training the dataset with the help of Naive BayesianClassifier.
Step 4: 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 datasetlog_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_bthen 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 theword which is wrongly classifiedNumdocs_wrong = sum (xor (output, text_lables))
Step 8: display the error rate of text data and calculate thefraction of wrongly classified wordFraction_wrong = numdocs_wrong/numtest_docs
Description: In this work we are describing the methodwhich is used to perform e-mail spam classification. The firststep is to select the file from the dataset and apply the featureextraction technique for extracted feature. For which we areusing the word-count algorithm. The next step is training thedataset which are extracted by the feature extractiontechnique. For training the data we can calculate theprobability of spam and non-spam words in the document.The next step is to test the data with the help of NaïveBayesian Classifier for which calculation the probability ofspam and non-spam mails and make a prediction whichvalue is higher. If spam words are greater than non-spamwords in a mail then the mail is spam mails otherwisenon-spam mails.

In the next step we are calculating the words which arewrongly classified by the classifier and calculate accuracy ofthe classifier and also calculate the error rate of classifier bycalculating the fraction of word which is wrongly classifiedand total number of words in document.

V. CONCLUSION& FUTURE SCOPE

Automatic text categorization is the task of assigning level of different categorization. In our paper it's between spam and ham and to make this procedure in reality we have incorporated. To solve this problem create an email spam classificationsystem and identifies the spam and non-spam mails.Here we are using the Naïve Bayesian Classifier andextracting the word using word-count algorithm. Aftercalculation we find that naïve Bayesian classifier has moreaccurate the support vector machine. The error rate is verylow when we are using the Naïve Bayesian Classifier.

REFERENCES

[1] Sharma K. and Jatana N. (2014) "Bayesian Spam Classification: TimeEfficient Radix Encoded Fragmented Database Approach" IEEE 2014pp. 939-942.
[2] Sharma A. and Anchal (2014), "SMS Spam Detection Using NeuralNetwork Classifier", ISSN: 2277 128X Volume 4, Issue 6, June 2014, pp.240-244.
[3] Ali M. et al (2014), , "Multiple Classifications for Detecting Spam emailby Novel Consultation Algorithm", CCECE 2014, IEEE 2014, pp. 1-5.
[4] Liu B. et al (2013) "Scalable Sentiment Classification for Big DataAnalysis Using Naive Bayes Classifier" IEEE 2013 pp.99-104.
[5] Belkebir R. and Guessoum A. (2013), "A Hybrid BSO-Chi2-SVMApproach to Arabic Text Categorization", IEEE 2013, pp. 978-984.
[6] Blasch E. et al (2013), Kohler, "Information fusion in a cloud-enabledenvironment," High Performance Semantic Cloud Auditing, SpringerPublishing.
[7] Allias N. (2013) "A Hybrid Gini PSO-SVM Feature Selection: AnEmpirical Study of Population Sizes on Different Classifier" pp 107-110.
[8] Prasad N. et al (2013) "Comparison of Back Propagation and ResilientPropagation Algorithm for Spam Classification", Fifth InternationalConference on Computational Intelligence, Modeling and Simulation,IEEE 2013, pp. 29-34.
[9] W.-W. Deng and H. Peng, "Research On A Naïve Bayesian BasedShort messaging Filtering System," in Proc. Fifth InternationalConference on Machine Learning and Cybernetics, Dalian, August13-16, 2006.
[10] J. M. G. Hidalgo et al., "Content based SMS spam filtering," in Proc.the 2006 ACM Symposium on Document Engineering, Amsterdam,The Netherlands, October 10-13, 2006.
[11] G. V. Cormack et al., "Feature engineering for mobile (SMS) Spamfiltering," in Proc. the 30th Annual international ACM SIGIRConference on Research and Development in information Retrieval,Amsterdam, The Netherlands, July 23-27, 2007.
[12] M. T. Nuruzzaman and C. Lee "Independent and Personal SMS SpamFiltering," presented at 11th IEEE International Conference onComputer and Information Technology, 2011.
[13] Cormack et al., "Spam filtering for short messages," in Proc. theSixteenth ACM Conference on Conference on Information AndKnowledge Management, November 06-10, 2007, Lisbon, Portugal.
[14] T. M. Mitchell, Machine Learning, McGraw-Hill.
[15] C. M. Bishop, Pattern Recognition and Machine Learning, Springer,2006.
[16] K. P. Murphy, Machine Learning: A Probabilistic Perspective.
[17] S. Tong and D. Koller, "Support vector machine active learning withapplications to text classification," Journal of Machine LearningResearch, pp. 45-66, 2001.

BIOGRAPHY

A. Deepika Mallampati, completed her M.Tech (SE), she is having teaching experience of 9years. Presently working as Assistant Professor, in the Dept. of Computer Science and Engineering, Sreyas institute of Engineering &Technology, Nagole, Hyderabad, Telangana India.