Email spam classification using neighbor probability based Naïve Bayes algorithm

Conference Paper · November 2017

DOI: 10.1109/CSNT.2017.8418565

3 authors, including:

Chakunta Venkata Guru Rao
SR Engineering College, Warangal, India
177 PUBLICATIONS **750** CITATIONS

SEE PROFILE

Email Spam Classification using Neighbor Probability based Naïve Bayes Algorithm

P. U. Anitha
Dept. of CSE
Jawaharlal Nehru Technological University, Hyderabad, India
anitha_podishetty@yahoo.co.in

C. V. Guru Rao
Dept. of CSE
S.R. Engg. College
Warangal, India
guru_cv_rao@hotmail.com

Suresh Babu
Dept. of CSE
Kakatiya Government College,
Hanamkonda, India
sureshd123@gmail.com

Abstract—Email spam is a kind of electronic spam, which tends to be a more difficult problem nowadays among all internet challenges. Spam mails are mostly sent in commercial purpose, some of them may contain malware links that lead to phishing websites. The aim of this study is to classify into ham and spam emails with an optimized and well efficient classification technique. Ham holds emails that are legitimate or legally valid message can get accepted by users. Spam emails are unwanted emails that a user doesn't want and to get rid of it. This study emphasizes on the improvement in classifying all mails into these two groups with minimal requirement of training and with an accuracy of hundred percent. Here in this study, Modified Naïve Bayes (MNB) classifier ensured the requirements with very low percentage of training and produces accurate results than existing Naïve Bayes (NB) or Supporting Vector Machine (SVM) classifier.

Keywords— Ham; Modified Naïve Bayes; Naïve Bayes; Spam; Supporting Vector Machine;

I. INTRODUCTION

In recent years, e-mail is one of the fastest and most economical forms of intercommunication medium over the internet. An email communication provides many facilities such as, people in order to communicate to parents and friends, share files, data or any other information. Emails are classified into two groups like ham or spam [1]. Ham emails are legal emails or a valid emails and spam emails are unsolicited and unnecessary emails. Hence, spam emails are producing a lot of issues and extremely rooted in internet [2]. Spam emails will misuse storage space, cause waste of time, produce harmful malware and significantly affects phishing links of users. Apart from those issues, spam email bandwidth costs are billions of dollars with dial-up connections to users [3], [4]. To overcome those problems, spam email filtering mechanism is significantly required. Various methods like Supporting Vector Machine (SVM), Genetic Algorithm, Bat Algorithm, Artificial Bee Colony (ABC) and Naïve Bayes have been employed to classify spam emails in real time which are all tried to improve accuracy but cannot able to do classification perfectly and also requires high training time.

The main function of an email spam classifier is to identify the mails which can be unwanted or harmful mail to a user and mark it as spam. Finally, it should prevent the spam mail from not going into the mailbox of the recipient. The classifier should be able to have a great influence over the spam vocabulary in order to predict it as a spam and should produce reliable classification [5]. Among all the existing methodologies of filters that it should be able to identify certain features and must rely on those values for classification. These kinds of features are unreliable and also have a risk of misclassification of ham email as a spam email or removal of legitimate mails [6]. Many spam filtering methods are not perfect or precise in classification. However, classification is so important for an email recipient who supposes to go through a large burden of emails [7], [8]. Effects of a misclassification of email cause not only time wastage but also many valuable information. Classification based on just the subject of mail or just glancing for every word in the concept is not at all efficient [9], [10]. The classifier should be always updated with the likelihood of the user in order to avoid error in classification.

In existing methodologies of email classification, it is summed up the probability of each word into priority value of mail to be spam. But in the real scenario each word's probability of spam is independent of other and also combination of two words probability of spam is independent of the probability of the same words in individual. For example, consider "Bumper" is a ham word and "Prize" is a ham word but the combination of this "Bumper Prize" will create spam which is not evaluated in existing methodology. In proposed scheme, this combination of words is evaluated in same sequence as they appear in sentence to evaluate the spam probability. The reason why it is taken as in the same sequence as appear in sentence is if a classifier takes all combinations of words it may even combine first and last word of email or sentence as a combination word which has no necessity to compute and may introduce unwanted values to computation. So, in order to avoid it same sequence as appear in actual sentence in email is taken as combination words.

This paper is organized in a manner that Section II briefs out the various mail classification strategies and its causes of performance degradation. Section III illustrates about the fundamental functioning of NB and factors that influence efficient performance of MNB with its working nature. Experimental arrangements and its corresponding results are described and its comparison towards existing

978-1-5386-1860-8/17/$31.00 ©2017 IEEE
DOI 10.1109/CSNT.2017.67

methodologies was said at section IV. Section V illustrates conclusions of the work.

II. Literature Review

Doaa Hassan [11] proposed a methodology of combining text clustering using K-means algorithm with various classification mechanisms to improve accuracy of classification of emails into spam or non-spam. The conjunction of clustering and classification mechanisms was carried out by adding extra features classification and also the classifier's performance was improved by clustering, results of this work show that combining K-means clustering with supervised classification in this methodology does not improve the classification performance for all mails. Further, the situations where the classifiers performance is improved by clustering, is found to be only slight increase in the the performance of classifiers in terms of accuracy with a very small amount which is not enough to meet requirements.

Gillani, et al. [12] presented an economic metric, based on the spam economic system by associating the detection accuracy of the detectors with the spammers cost. Hence, the sensitivity of a detector does not need to be tuned all the way up to maximize detection, but enough to make spamming cost intolerable to the spammer. So, spam detector will employ statistical features, in order to easily differentiate the spam emails. The advanced method estimations have presented the effectiveness and significantly decreased the false positives in spam detector. But, the pitfall associated with this method is to fix the spamming cost to a level that all average spam mail possess without knowing any value regarding them and also not efficient in initial conditions of mail box.

Sunday Olusanya Olatunji [13] presented a method on email spam detection based on Support Vector Machines (SVM) for spam detection while paying attention to appropriately search for the optimal parameters to achieve better performance. SVM has certain drawbacks like not concentrating towards priority of a word to be a spam and ham. And also requires large amount of mails in order to perfectly classify the mails.

Rushdi Shams et. al [14] implemented a work on supervised classification of spam emails with natural language stylometry attributes. This method will extract all attributes from a mail related to writer stylometry in order to classify mails. The major limitation of this methodology is it is only suitable for personalized mails and not suitable for commercial or official mails. Moreover, this method is not suitable for new writer's mails and it has higher probability of rejecting ham mails.

Xiuyi Jia et al. [15] have presented a three-way decision solution for filtering of spam, which can reduce the error rate of classifying a legitimate email into spam with minimum misclassification cost. Also the solution can provision a more efficient decision procedure for users because it is not restricted to a specific classifier.

III. Proposed Method

In the handling of electronic spam, it is a tougher job to segregate a huge burden of emails in a recipient's inbox and preventing from the attack of spam emails. It depends on the taste acceptance and the approach towards utilizing email conversations by an individual recipient. A spam for an ordinary person could be a ham for an authority or official who used to take actions against it. Some mails also may be sent by the control authorities or in a noble cause to aware people from spam could be classified as spam because the only reason it uses such spam words often.

In order to avoid these kinds of misclassification and also strictly prevent from attack of spam with less requirement of training the proposed methodology is derived. This methodology will utilize the probability of occurrence of several independent words in an email and their probability of spam and make conclusions out of it like whether the mail is spam or ham. Proposed methodology uses MNB classifier for classification purpose to make accurate decisions on a mail to be spam or ham. MNB works mainly to accomplish two purposes; one is to classify mails precisely into ham and spam emails; second is to classify a mail according to the relative occurrence of words to specify ham or spam with the approach to make sure that none of the healthy mails for recipient should not specify as spam.

In general NB classifier classifies set of objects based on training to identify what kind of data belongs to a certain category. If it finds similar while testing phase, then it will mark it up to that corresponding category. The basic work function of such NB classifier is described as follows in order to understand the fundamental classification mechanism.

A. Naïve Bayes Classifier

In the field of machine learning NB is a probabilistic approach to classification it applies a traditional Bayes theorem to calculate probabilities of a particular category with a powerful assuming of the difference among features. It still retains its state as the best scheme for classification of text or for problems that corresponds to make decisions over documents, whether it belongs to a certain category or not based on frequency of occurrence of words.

The spam email feature classification purpose NB classifier is utilized and the best features are recognized. By using this classifier, for each email in testing spam dataset with preferred essential attributes, algorithm calculates the spam and non-spam emails.

According to NB algorithm, Emails are classified into individual words $w_1, w_2, ..., w_n$ and selected features are denoted as F. The probabilities of receiving emails are equal to the probability of receiving the list of words.

$$P(F) = P(w_1, w_2, ..., w_n) \quad (1)$$

By the above equation naive Bayes assumption becomes,

$$P(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} P(w_i) \quad (2)$$

Next, two classes of emails are indicated as $spam(S)$ and $ham(H)$. After that the probabilities of $P(F|S)$ (probability of given feature from email class S) and $P(F|H)$ (probability of given feature from email class H). Thus, $P(F|S)$ and $P(F|H)$ can be expressed as,

$$P(F|S) = P(w_1, w_{2,...,} w_n|S) \quad (3)$$

$$\prod_{i=1}^{n} P(w_i|S) \quad (4)$$

And,

$$P(F|H) = P(w_1, w_{2,...,} w_n|H) \quad (5)$$

$$\prod_{i=1}^{n} P(w_i|H) \quad (6)$$

Then, a training dataset is to estimate how spammy each word is, where probabilities $P(w_i|S)$ (probability of given an email from email class S which it contains the word w_i) and $P(w_i|H)$ (probability of given a email from email class H which it contains the word w_i) are needed. In the following formula, $P(w_i \cap S)$ is the probability that a given email is a spam email and contains the word w_i. Thus, by Bayes theorem:

$$P(w_i|S) = \frac{P(w_i \cap S)}{P(S)} \quad (7)$$

$$P(w_i|H) = \frac{P(w_i \cap H)}{P(H)} \quad (8)$$

The following step is to compute the posterior probability of spam email given the overall probability of the sampling email by Bayes' rule; this is the crucial part of the entire classification.

$$P(S|E) = \frac{P(E|S)P(S)}{P(E)} \quad (9)$$

$$P(S|E) = \frac{P(S)\prod_{i=1}^{n} P(w_i|S)}{P(E)} \quad (10)$$

And similarly,

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \quad (11)$$

$$P(H|E) = \frac{P(H)\prod_{i=1}^{n} P(w_i|H)}{P(E)} \quad (12)$$

Therefore we can classify the email by comparing the probabilities of $P(S|E)$ (probability of a given email is classified as spam which belongs to the email class S) and $P(H|E)$ (probability of a given email is classified as ham which belongs to the email class (H). Initially find the ratio of the two probabilities.

$$\frac{P(S|E)}{P(H|E)} = \frac{P(E|S)P(S)}{P(E|H)P(H)} \quad (13)$$

$$\frac{P(S|E)}{P(H|E)} = \frac{P(S)\prod_{i=1}^{n} P(w_i|S)}{P(H)\prod_{i=1}^{n} P(w_i|H)} \quad (14)$$

The above equations 13 and 14 denotes the amount of probability ratio of an email as spam to ham value based on the combination of independent probabilities of n words in that mail and expected probability of ham and spam.

$$\frac{P(S|E)}{P(H|E)} = \frac{P(S)}{P(H)}\prod_{i=1}^{n} \frac{P(w_i|S)}{P(w_i|H)} \quad (15)$$

The products in the above equation can be extremely small values if we have a big amount of words, w_i. To overcome this issue, we apply log to probability ratio.

$$\log\frac{P(S|E)}{P(H|E)} = \log\frac{P(S)}{P(H)}\prod_{i=1}^{n} \frac{P(w_i|S)}{P(w_i|H)} \quad (16)$$

$$\log\frac{P(S|E)}{P(H|E)} = \log\frac{P(S)}{P(H)} + \sum_{i=1}^{n} \log\frac{P(w_i|S)}{P(w_i|H)} \quad (17)$$

Here, using above equation, to calculate the log posterior probability when receive a new email. If the result is greater than zero (which means $P(S|E) > P(H|E)$), we classify email E as spam. Similarly, we classify the email as ham if it is less than zero (which means $P(S|E) < P(H|E)$). Finally, NB classifier decides whether the email as ham or spam depending on individual word's probabilistic performance. But, when it comes to an email as a whole which has lots of words, each word has certain spam level independently. Then one cannot decide the level of spam of

an email by simply summing all those values because independent event's probability cannot be supposed to sum together. This consequence has put forth a need for a new approach in NB classification of documents.

B. Modified Naïve Bayes Classification

All words in a mail have independent nature of spam level according to laws of probability the probabilities of independent event should not be added to sum of probabilities, which results more than one. For example, consider the word "Bumper" is a hammy word and "Prize" is also a hammy word but when these words combine together "Bumper Prize" which is a spammy word. This example shows that the combination of words can also create a spam which cannot be calculated by ordinary classification methods.

Proposed scheme introduces a method to combine the probabilities of many independent events and take it as a single probability of an email and utilize it to evaluate whether the given mail is spam or not. This scheme is implemented via a slightly different approach in NB classifier. Its training Enron dataset also contains difference in ratio of ham and spam mail in order to show that a recipient will receive ham mails more than the spam emails. It starts by counting the number of appearance of ham words in a test document (AH) and number of appearance of spam word in a test document (AS).

The probability of whether a given word is spam can be calculated as follows

$$P(w_j / S) = \frac{AS / TotalSpam}{((AS / TotalSpam) + (AH / TotalHam))} \quad (18)$$

Where,

$P(w_j / S) \rightarrow$ Probability of a word to be a ham

$AS \rightarrow$ Appearance of spam words in test document

$AH \rightarrow$ Appearance of ham words in test document

$TotalSpam \rightarrow$ Total number of spam words in training set

$TotalHam \rightarrow$ Total number of spam words in training set

Probability of finding a word to be ham or spam in testing email can be given as

$$P(S_n / E) = \frac{P(S)\prod_{j=0}^{AS} P(w_j / S)}{P(H)\prod_{i=0}^{AH} P(w_i / H) + P(S)\prod_{j=0}^{AS} P(w_j / S)} \quad (19)$$

$$P(H_n / E) = \frac{P(H)\prod_{i=0}^{AH} P(w_i / H)}{P(H)\prod_{i=0}^{AH} P(w_i / H) + P(S)\prod_{j=0}^{AS} P(w_j / S)} \quad (20)$$

Where,

$P(H_n / E) \rightarrow$ Probability of an email to be a ham

$P(S_n / E) \rightarrow$ Probability of an email to be a spam

$P(w_j / S) \rightarrow$ Probability of a word to be a spam

$P(w_i / H) \rightarrow$ Probability of a word to be a ham

Probability calculated from the above equation can be used to evaluate whether the given email to be spam with the relation among combination of adjacent words in a line. That we can able to compute spamliness of an email without any complex computations. Thus, spam combinations can be easily identified which helps us to avoid spam emails more precisely and prevent it from going to recipient's mailbox. Proposed system does not utilize the combinational computation among all words in the document because it is so complex. Moreover, the probability computations may be wrongly calculated because of unnecessary combinations of words in different lines creating spam.

Proposed methodology only uses the combinations of adjacent words to be spam and combines its probability of spamliness by traditional way to combine independent probabilities as a product or Geometric Progression (G.P.). Thus, this scheme requires only less amount of training data set as a minimum of about 60% for testing and 40% testing with that ratio MNB can able to produce precise results. Moreover, MNB produces with more efficient approach of classification along with minimum number of training data and lesser training time.

MNB produces an accurate classification of ham and spam mails by an adaptable ratio of priority that can be fixed according to how often ham and spam mails are received by recipient. This scheme also provides reliability to recipient that an email can be categorized with a relative likelihood of ham or spam by the combination of consecutive words in the content. MNB also provides a score of ham and spam for every word in training dataset in addition words itself will improve its priority of ham or spam according to its consecutive word. This priority will improve itself by equations that combine independent probabilities of word in content of an email. Thus, MNB classifier can be able to adapt its classification to any kind of mail recipients with very minimal ratio of training data.

IV. EXPERIMENTAL SETUP

Implementation of proposed scheme in an experimental setup rises up for a need of standard dataset that acts a model recipient mailbox with both the collections of ham

and spam email messages. In this study we are using Enron dataset which contains about 2736 ham mails and 972 spam mails in total as a sample to process. Implementation of this work is done using tool java NetBeans version 8.2 with minimum requirements of operating system as windows 7 XP and 2GB RAM for smoother working of scheme.

From the above discussions it is clear that MNB works on combination of consecutive words which leads to a good ethical approach of classification of ham and spam emails in a recipient mail box. Doaa Hassan [11] study states the accuracy of various existing methodologies on enron dataset not to be perfect and does not utilizes combination of consecutive words. Following chart (Fig 1) illustrates the resultant accuracy comparison of proposed scheme with various existing methodologies:

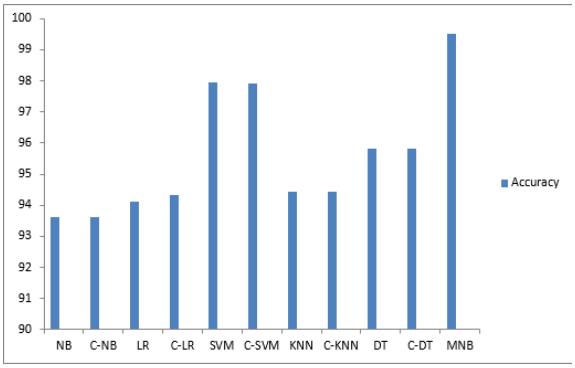

Fig. 1. Accuracy of various mail classification systems

Accuracy chart illustrates that MNB is more precise in classification of mails of ham and spam when compared to all other previous strategies. It also shows that all the systems aim only to reach the accuracy of calculation not concerned about the content classified is efficiently classified or not. From the data inferred from Doaa Hassan's study following table (Table 1) [11] illustrates the accuracy of classification of mails by various methodologies in precise percentages.

Accuracy values of various existing methodologies

Methodology	Accuracy
NB [11]	93.6
C-NB [11]	93.6
LR [11]	94.1
C-LR [11]	94.34
SVM [11]	97.94
C-SVM [11]	97.9
KNN [11]	94.45
C-KNN [11]	94.45
DT [11]	95.83
C-DT [11]	95.83
MNB	99.5

MNB also takes fewer amounts of data for training and also will operate with less complexity. MNB is able to work efficiently even in lesser ratio of training and testing messages. Following table (Table 2) will illustrate the different ratio of training and testing of MNB and its corresponding accuracy of classification.

TABLE I. DIFFERENT TRAINING TESTING RATIO OF MNB AND ITS ACCURACY VALUES

Training %	Testing %	Ham Accuracy	Spam Accuracy
60	40	99.03	98.5
70	30	99.5	99.5
80	20	99.5	99.5

MNB has clearly stated from the above table 2 that MNB requires only requires 60% of data to train in order to produce the greater than or equal to accuracy provided by the existing algorithm. This helps to reduce the time span of the whole process while other methodologies are so firm theory that for improving accuracy one need to improve the amount of training data. Getting more data and sorting it to relevant ham and spam for training the classifier is tougher job and doing it in large numbers is more time consuming process.

MNB undergoing training with minimal amount of data get trained under very less span of time. For standardization consider taking 70% of training and 30% of testing as standard ratio for training and testing. Time span to build training dataset for several methodologies shown in below table (Table 3) illustrates betterment of MNB training performance with them from Doaa Hassan work [11].

TABLE II. TRAINING TIME REQUIRED FOR VARIOUS CLASSIFICATION TECHNIQUES

Methodologies	Training time in s
NB [11]	6.96
C-NB [11]	61.848
LR [11]	41.22
C-LR [11]	613.01
SVM [11]	4.96
C-SVM [11]	63.59
KNN [11]	4.23
C-KNN [11]	55.33
DT [11]	41.55
C-DT [11]	106.66
MNB	3.5

From the above table it is notable that MNB has very less time span for training data to classifier than all other existing schemes of classification.

V. CONCLUSION

MNB is an email spam classifier which can capable of classifying with an average of 99.5% accuracy. Moreover, it requires a lesser amount of data for training and to give its standard performance with a very low training time of 3.5 seconds. So far from this study, it is inferred that MNB is a fast and reliable classifier because of its nature of relating

independent probabilities of words in the content of an email. MNB gives out a new ethical approach of email classification with combining independent probabilities of consecutive words. In future, by improving the method for classifying unidentified or new words from a test email efficiently MNB can be more accurate in classification of emails. And also by decreasing the total number of mails in dataset and maintaining the same accuracy will also help to reduce the build time of training dataset.

REFERENCES

[1] I. Idris, and A. Selamat, "Improved email spam detection model with negative selection algorithm and particle swarm optimization," Applied Soft Computing, vol. 22, pp. 11-27, 2014.

[2] F. Gillani, E. Al-Shaer, and B. AsSadhan, "Economic metric to improve spam detectors," Journal of Network and Computer Applications, vol.65, pp. 131-143, 2016.

[3] M. Luckner, M. Gad, and P Sobkowiak, "Stable web spam detection using features based on lexical items," Computers & Security, vol. 46, pp. 79-93, 2014.

[4] S. Maldonado, and G. L'Huillier, "SVM-based feature selection and classification for email filtering," Pattern Recognition-Applications and Methods, Springer Berlin Heidelberg, pp.135-148, 2013.

[5] B. Zhou, Y. Yao, and J. Luo, "Cost-sensitive three-way email spam filtering," Journal of Intelligent Information Systems, vol. 42, pp. 19-45, 2014.

[6] M. Mohamad, and A. Selamat, "An evaluation on the efficiency of hybrid feature selection in spam email classification," in International Conference of Computer, IEEE Communications, and Control Technology (I4CT), 2015, pp. 227-231.

[7] M. Zavvar, M. Rezaei, and S. Garavand, "Email spam detection using combination of particle swarm optimization and artificial neural network and support vector machine," International Journal of Modern Education & Computer Science, vol. 8, pp. 68, 2016.

[8] C. Neelavathi, and S.M. Jagatheesan, "Improving Spam Mail Filtering Using Classification Algorithms with Partition Membership Filter," International Research Journal of Engineering and Technology (IRJET), vol. 03, pp. 380-383, 2016.

[9] S.S. Hong, W. Lee, and M.M. Han, "The Feature Selection Method based on Genetic Algorithm for Efficient of Text Clustering and Text Classification," International Journal of Advances in Soft Computing & Its Applications, vol. 7, pp. 22-40, 2015.

[10] D.K. Renuka, P. Visalakshi, and T. Sankar, "Improving E-Mail Spam Classification using Ant Colony Optimization Algorithm," in International Conference on Innovations in Computing Techniques (ICICT 2015), pp. 22-26, 2015.

[11] D. Hassan, "Investigating the Effect of Combining TextClustering with Classification on ImprovingSpam Email Detection," in Intelligent Systems Design and Applications,Advances in Intelligent Systems and Computing 557,Springer-Berlin Heidelberg, pp. 99-107, 2017.

[12] W.A. Awad, and S.M. ELseuofi, "Machine Learning methods for E-mail Classification," International Journal of Computer Applications, vol. 16, pp. 39-45, 2011.

[13] S. O. Olatunji, "Improved email spam detection model based on support vector machines," Neural Computing and Applications, pp. 1-9, 2017.

[14] R. Shams, and R. E. Mercer, "Supervised classification of spam emails with natural language stylometry," Neural Computing and Applications, vol. 27, pp. 2315-2331, 2016.

[15] X. Jiya and L. Shang, "Three-Way Decisions Versus Two-Way Decisionson Filtering Spam Email," in Transactions on Rough Sets XVIII, 2014, pp. 69-91.

[16] G.S. Tomar, S. Verma & Ashish Jha; "Web Page Classification using Modified naïve Baysian Approach", IEEE TENCON-2006, pp 1-4, 14-17 Nov 2006.