Novel Approach: Naïve Bayes with Vector Space Model for Spam Classification

Safvan Vahora
Dept of Information Technology
VGEC, Chandkheda
Ahmedabad, India
safvan465@gmail.com

Mosin Hasan
Dept of Computer Engg.
BVM Engg College
V.V.Nagar, India
mosin83@yahoo.com

Reshma Lakhani
Dept of Computer Engg.
CSPIT, Charusat
Changa, India
reshmalakhani21@yahoo.com

***Abstract*--We always see our normal mail goes into spam folder of the mail box. Interestingly 90% of the time the mail server classifies it perfectly but sometimes it fails due to spammer are getting highly technical. In this paper, we are using novel approach which uses Vector space model with Naïve Bayes to correctly classify mails as spam mail. Naïve Bayes method is used for spam classification but still binding with personalize word vector helps in increasing the accuracy of the system because user receives special type of message only. In this research work, we use vector space model with naïve bayes to classify spam mail. We got nearly 85% of accuracy in spam classification. We have used personalize mail classification option instead of standard global classification because people visiting subjective (i.e. pornographic) sites frequently get spam mail related to that subject (pornography) only and hence personalization shows improved result.**

***Index Terms*--Classification, Spam, Naïve Bayes, Vector space model, Spammer**

I. INTRODUCTION

Spam: An electronic communication containing material or references to material of a commercial, solicitation or illegal nature, directed as part of a bulk distribution to any address where the address holder has not given explicit prior consent to receive it. [1] It is always unwanted – the receiver has not shown or given any direct consent to receive it. It is typically commercial or promotional (or sometimes illegal like child pornography) in nature and sometimes fraud or scam. [2] It is usually sent in masses to cover more people compared to direct marketing. But Sometimes user does not care about it, as spam mail gives a lucrative scheme. And it sends in masses but user is not aware of whether it is sent to 1 or 100 or 1 million people [2] Like, "you won lottery" so from the user perspective only one attribute which is "Unwanted Mail" is considered as spam mail.

1) Types of Spam

If there is no product to sell, then there is no spam. Ultimately, spam is about marketing a product or service, and for certain types of product or service; it is a very effective medium. [2] In this category we find the true villains of the piece. These are people who have made a career out of sending spam, almost always on behalf of vendors. There is the old joke that "guns don't kill people, people kill people", which is often used by weapons manufacturers to absolve themselves of associative guilt. [2] In the same way that a market demanding weapons will always find plenty of people who have no ethical problem supplying them, so spam has spawned an entire industry that revolves around the development and supply of specialist tools for harvesting addresses, sending spam and covering tracks.

2) Cost of Spam

Spam is not similar to direct marketing or telemarketing. In direct or telemarketing, cost of marketing is bear by the company, while in the spam cost is borne by the recipient. The recipient uses connection time to retrieve the message. Message also wastes space in the recipient's disk allocation. The ISP uses bandwidth to receive the spam, and pays for administrator time and licensing fees to maintain software to trap and remove the spam before it gets to the recipient. Spam may be cheap for the people who send it, but it can be a serious expense for your business. According to a study conducted earlier this year by Nucleus Research Inc., spam management costs U.S. businesses more than $71 billion annually in lost productivity — $712 per employee. [3]

There are few ways that spam drains your company's bank account like Anti-spam technology, loss of productivity, space wastage and few intangible cost. [4] Macfee has recently published the report on carbon footprint of spam mail which says [4], an estimated worldwide total of 62 trillion spam emails were sent in 2008. Globally, annual spam energy use totals 33 billion kilowatt- hours (KWh), or 33 terawatt hours (TWh). That is equivalent to the electricity used in 2.4 million homes in the United States, with the same GHG emissions as 3.1 million passenger cars using two billion United States gallons of gasoline. [4]

3) Important of Spam in the Perspective of Advertiser

E-mail is nearly free – the costs involved in sending a million e-mail messages are very less, which means that spammers can make a profit even on very low response rates. In a traditional direct-mail marketing campaign, a response rate of 2 – 4% is considered good, and is usually profitable

978-1-4577-2168-7/11/$26.00 ©2011 IEEE

enough to justify the campaign.

4) Statistics of Spam

The number of unwanted e-mail (commercial or scam) messages transmitted through the internet has been increasing nowadays. There was only 8% of spam of network e-mail traffic in 2001; however it reaches 70% in 2005 and increases exponentially showing high insecurity of internet technologies. [5] Some facts on spam statistics, 90% of spam is in English. A year ago it was 96%, so spam is getting more "international." [6]

II. PROBLEMS AREA IN SPAM CLASSIFICATION

Recently due to highly accurate spam classification tool now the spammers are switched to few interesting area from where it is quite difficult to track the spam. Following are the few well known emerging technique used by spammer.

Image: Most of the spam classification tool used the text classification technique and hence if entire or partial spam mail is presented in the form of image than it is quite difficult for the classifying tool to identify the mail as spam.

Foreign Language: Nowadays spammer used language like Spanish, German, and French to send the mail. They try to get localized and hence normal blacklist word approaches do not work for the spam mail which uses this type of techniques.

This two are recent problems that need to be tackle down by the researcher. A number of statistical classification methods and machine learning techniques have been applied to text categorization, including techniques based on decision trees [7, 16], neural networks [7, 16] Bayes probabilistic approaches. [7, 16]

III. TEXT CLASSIFICATION

The number of statistical classification methods and machine learning techniques have been applied to text categorization, including techniques based on decision trees, [15] neural networks [16], Bayes probabilistic approaches [15]. However, there is still need more accurate text classifiers based on new learning approaches. The purpose of the current work is to describe ways in which hybrid approach can be applied to the problem of text categorization, and to test its performance relative to a number of other text categorization algorithms. [7]

Several attempts, some of them quite successful, have been made at applying standard text classification techniques to spam filtering, for applications involving either personal mail [8, 9] or mailing lists [10]. However, operational spam filters must rely not only on standard machine learning techniques, but also on manually selected features.

1) Vector Space Model for Term Weighting

Vector space model is used to represent the document as vector based on feature weighting. Traditional Vector space model algorithm has local weight, there is no globalization. Globalization having number of advantages that described as below in global term weighting. This algorithm is used in many traditional information retrieval tasks such as text search, text categorization, text clustering. For example in spam mail, categorization vector space model can be used which identify whether the incoming email is spam and if yes then sending it to the spam folder in your mail box. [11] In this algorithm, first from the spam mails vector space will be generated which contains the feature of spam mails based on the term weighting and second it will generate vector space for the non-spam mails. In the testing phase, new mail will be and from that vector space will be identified and it will classify whether the new mail is Spam or Non Spam [11].

There are two phases in designing Spam Detector namely training and testing phase describes as in subsequent section.

A. Training Phase

In this phase our software is trained with dataset of nearly 50 Spam mails and 50 non spam mails. Spam mails are gathered of various categories. Table 1 shows the distribution of spam mails across various categories.

TABLE I
DISTRIBUTION OF SPAM MAILS

Spam Distribution (Dataset)	**Percentage (%)**
Adult	12
Health	4
Scam	18
Financial	10
Product /Services	24
Miscellaneous	32
Total	100

In this phase our software will generate Vector Space that will used to provide the result. Figure 1 shows diagrammatically representation of Training Phase. Training phase contains three modules are Feature extraction, Weight Assignment and Storage.

Feature Extraction: This module is divided into two sub-modules as Feature Identification and Stemming.

Feature Identification: In this module, Mail dataset is given as input. Each E-mail message contains number of words. Features are identified form that E-mail and given it to next module.

Stemming: A single word can be represented by many ways, for example "offer", "offering", "offers". This all should be stem and considered as a single word "offer". Stemming reduces the vector space by reducing set of word to one common word. Each word has its own importance in the mail and hence to assign the weight to the feature, features are passed to weight assignment module.

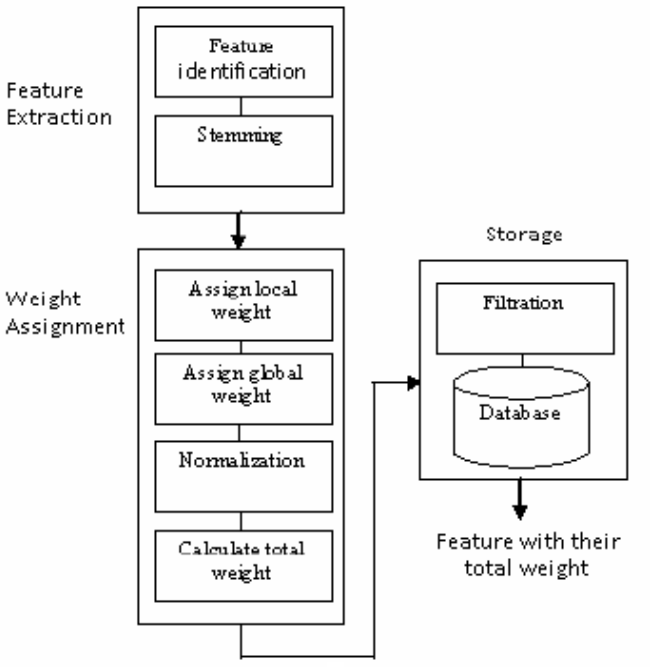

Fig. 1. Training Phase for Spam Detection

Weight Assignment: There are three components used in weighting scheme to calculate the total weight of feature:

$$a_{ij} = g_i * t_{ij} * d_j \quad (1)$$

where g_i is global weight of the i^{th} term, t_{ij} is the local weight of the i^{th} term in the j^{th} E-mail , d_j is the normalization factor for the j^{th} E-mail.

Local Term-Weighting: Logarithmic Term Frequency (TF)

This weight depends on the frequency of feature within the E-mail. When the frequency of any word is high at that time logarithmic term frequency formula is used. This method is to de-emphasize the effect of high frequency. [12] For example if "Free" word appears ten times in the E-mail than it doesn't mean that this word is ten times more important in Spam Mails. If we don't do this then any time if "Free" word appears in normal mail then also it gets classified as Spam due to high local frequency. The formula for Logarithmic term frequency is

$$\log(f_{ij} + \quad (2)$$

where f_{ij} is frequency of i^{th} feature in j^{th} Spam. [11]

Global Term-Weighting: Inverse Document Frequency (IDF)
Stop-list: it contains words those doesn't having any potential as feature in any document like "is", "There", "Subject", "Those", "That". Global term weighting is used to remove the necessity of stop-list. We have used IDF formula that is defined as logarithm of the ratio of total number of documents to the number of documents contains that word. [13] Thus common words having low IDF. For example we have total 100 E-mail and out of those only 20 contain a particular feature than Global Weight for that particular feature is log(100/20)=0.698. If all 100 E-mail contain that feature than Global Weight for that feature is log(100/100)=0.

Normalization: Normalization is required for 2 purposes. Normally spam mail will not be lengthy and normalization is not used.

Storage: Storage module contains Filtration and Database sub-modules.
Filtration: After the completion of Weight assignment there are number of features that have total weight zero or not contains any potential value of total weight. This all features remove out in this phase. This phase is called filtration phase as it removes unnecessary features from the vector space.
Database: Feature those having potential value of total weight are stored in database and form vector space contains total weight, state either spam or non-spam E-mail for each feature.

B. *Testing Phase*

Using Vector Space that we have created during Training Phase when a new E-mail is submitted to the system (Spam Detector), it will generate the result and shows that whether that E-mail categorized as spam or non-spam.

IV. NAÏVE BAYES FOR SPAM CLASSIFICATION

1) Basic of Naïve Bayes

There are two steps in Naïve bayes classifier. First step is to build vector of words that founds in the training set of document. [14] Second step is to calculate the probabilities for each category for example Spam and Non-Spam in the naive bayes classifier. This can be accomplished by first calculating the prior probability P(doc), that is for each category; which is 1/(number of documents) for each category. Next we had calculate the probability given a word from the set of all no repeated words that for each category. Hash table has been created for each category with key as word. And the values in the hash table are the number of times the word occurred in all the documents in that category. For example if the category was Spam and "free" occurred 3 times in each of the 20 training documents then total number of times word occurred would be 60 for that word. The total word count n (including repeated words) for each category was also calculated. From these values we can now calculate p(wi/spam) with the following equation.[14]

$$P(w_i/spam) = (nspam+1) / (ntotal + nr) \quad (3)$$

where nspam is number of occurrence of word w_i in spam category, ntotal is number of total words including repeated words in spam category, nr is number of total words in all category excluding repeated words.

With this calculation another hash table created having word w_i with its probability in each category. Thus we calculate P(category/document) as

$$P(category/document) = P(category) * \Pi\ (w_i/category)\ P(category) \quad (4)$$

The category of given testing document was max out of both categories Spam and Non-Spam category.

2) Experimental dataset and result

We have gathered 50 Non-spam mails and 50 Spam mails for the training and testing purpose. Following table shows the training and testing phase result for the given data.

TABLE I
EXPERIMENTAL RESULTS

Class	Training	Trained Data	Testing	Success	% of success
Non Spam	10	20	43	19	44.18
Spam	10		40	34	85
Non Spam	20	40	33	22	66.66
Spam	20		30	25	83.33
Non Spam	30	60	23	18	78.26
Spam	30		20	14	70
Non Spam	40	80	13	11	84.61
Spam	40		10	8	80
Non Spam	45	90	8	7	87.5
Spam	45		5	4	80

From the above statistics, we can conclude that when less data is given as training and more in testing, we will have the less success rate. As we increase the training set the success ratio also increases and after sometimes it become steady. Figure 2 show that success ratio reaches to saturation point around 85%. Performance of the spam detection can be increase considerably by providing personalize vector for each user which state specific word list. This wordlist can be used to identify legitimate e-mail for that specific user.

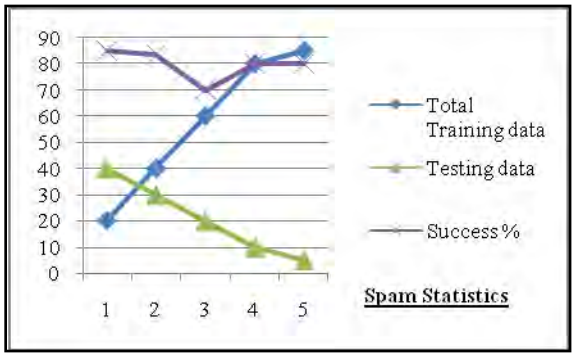

Figure 2: Statistics of the results : Spam Statistics

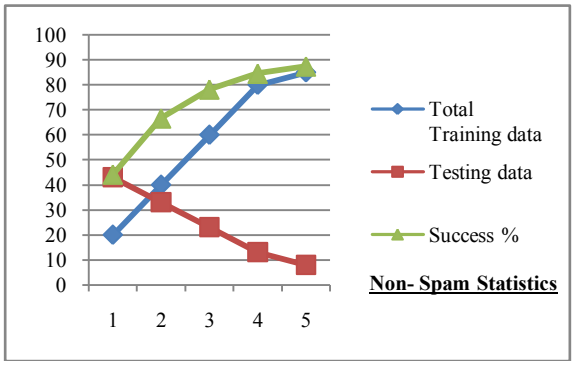

Figure 3: Statistics of the results : Non-Spam Statistics

V. CONCLUSION

The fusion of vector space model with naïve bayes provides the reasonable accuracy to classify spam mails. We check the possibility whether we can identify an E-mail as spam or non-spam on the bases of Vector (word) appears in normal Spam mails. We can also enhance this software by adding the prevention technique of Naïve Bayes Poisoning, as well as there are many mail which contains only image and link by adding one more module. Even today numbers of spam are appearing having foreign language. The use of E-mail body for classification, in the described methods improves the performance to classify the e-mails to spam, non-spam. The process to choose stop-list is slow the performance of the software. Here, by using the global weight, the problem of stop list is resolved and the performance of the system improves. We have used personalize mail classification option instead of standard global classification because people visiting subjective sites frequently get spam mail related to that subject only and hence personalization shows improved result. The performance figures are based on data set from different individuals. As the data set for training would be increase, the spam detector system becomes more robust. We need such effective Spam Detector which can be put on the ISPs so that, it would not only, checks for miscellaneous activity but also check the Scam mail and prevent customer from getting phished.

VI. REFERENCES

[1] W. W. Cohen, "Learning rules that classify e-mail," in AAAI Spring Symposium on Machine Learning in Information Access, Stanford, March 25–27.1996

[2] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk email," in AAAI'98 Wkshp. Learning for Text Categorization, Madison, WI, July 27, 1998.

[3] Cost of Spam : http://www.focus.com/briefs/it-security/real-cost-spam/

[4] The Carbon Footprint of Email Spam Report. http://www.resources.mcafee.com/content/NACarbonFootprintSpam

[5] MessageLabs Ltd, Spam review // Research report. – MessageLabs Ltd.; http://www.messagelabs.com

[6] Pingdom Report :Diving deep in to email spam statistics, Jan 19, 2011

[7] Nerijus Remeikis, Ignas Skučas and Vida Melninkaitė "Text Categorization Using Neural Networks Initialized with Decision Trees", Journal Informatica, Volume 15 Issue 4, December 2004

[8] Ion Androutsopoulos , John Koutsias , Konstantinos V. Chandrinos , Constantine D. Spyropoulos, An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages, Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, p.160-167, July 24-28, 2000, Athens, Greece

[9] José María Gómez Hidalgo, Evaluating cost-sensitive Unsolicited Bulk Email categorization, Proceedings of the 2002 ACM symposium on Applied computing, March 11-14, 2002, Madrid, Spain

[10] Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Spyropoulos, C.D. & Stamatopoulos, P., A memory-based approach to anti-spam filtering for mailing lists. Information Retrieval, 6(1), pp. 49–73, 2003.

[11] Nicola Polettini "The Vector Space Model in Information Retrieval – Term Weighting Problem". Department of Information and Communication Technology, University of Trento.2004

[12] Tamara Gibson Kolda. "Limited-Memory Matrix Methods with Applications. Applied Mathematics Program". University of Maryland at College Park, 1997.

[13] K. Sparck Jones. "A statistical interpretation of term specificity and its application in retrieval." J. Documentation,1972.

[14] Pattern Recognition Techniques and Applications by Rajjan Shinghal, Oxford Press.304 p,2005.

[15] D.Lewis and M. Ringuette "A comparison of two learning algorithms for text categorization. In Third Annual Symposium on Document Analysis and Information Retrieval, 81-93. 1994

[16] E.wiener, J. Pedersen,and A. Weigend, A neural network approach to topic spotting. In Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval 317-332. 1995