International Journal of Information Technology and Knowledge Management
January-June 2012, Volume 5, No. 1, pp. 169-175

AN IMPROVED STATISTICAL FILTER FOR SPAM DETECTION COMBINING BAYESIAN METHOD AND REGRESSION ANALYSIS

K. Srikanth[1], S. Ramakrishna[2] and K. V. S. Sarma[3]

ABSTRACT: The Naive Bayesian filter is the most popular statistical filter used for email filtering. The design of the filter depends however on the training data and the word corpus used by the filter designer. A new mail with unknown nature is classified into spam (unsolicited mail) or ham (legitimate mail) basing on a score by combining conditional probabilities of tokens in the mail. The statistical behavior of this score indicates some interesting features, which can be explored to improve performance of the filter. We propose a new method that utilizes the correlation structure between the number of words in the mail and the Bayesian score. We report the results of an experiment using Enron data set and highlight the advantages of the new filter. We also propose a new method of testing the model using random data sets.

Keywords: Statistical filter, spam, bayesian.

1. INTRODUCTION

Unsolicited email called spam is commonly found in the inbox of the recipients. According to Bowers and Harnett (2008) nearly 80% of email received by users is spam. A significant amount of time and productivity are lost for identifying and deleting spam mails.

Spam mails usually do not have a stable style and features. Spammers who send such mails, go on changing the features. For instance, the word Lottery will be written as Lot_tery or lotterie, so that filters cannot detect them. Legitimate mails are often called hams. Filters based on specific words or special characters (often called tokens) are commonly adopted by the end user to design their own filters. Content based filtering is a scientific approach to study the properties of spams. Since the spam features are not exactly known, the only way of handling them is by using probability theory. Bayesian filter is most popularly used statistical filter, which was first published by Sahami et al (1998) and known as Navie Bayesian filter. Paul graham (2002) and Tim Peter (2002) made several improvements of this version focusing on the estimation of token probabilities.

There is voluminous research contribution in this area, which is a common interest in the fields of data mining, text mining, machine learning and classification studies. Prabhakaran Raghavan and Christopher Manning (2003) have studied the problem of text classification using Bayesian filtering. Other important references includes Spam Assasin (2005) and Spam Bayes (2002).

Development of a filter is based on a corpus of emails (training data) each of which is already known to be spam or ham. The filter is expected to classify a new mail into spam or ham basing on a statistic (score obtained from the probabilities) derived from the training data. The following are the phases in designing a filter.

1. Developing Email corpus.
2. Tokenization (splitting the contents into words and special characters).
3. Developing a token corpus (table of tokens) by counting the number of times a token appears in spam and ham class. This table is called hash table.
4. Estimating token probabilities for each token in the hash table.
5. Combining the probabilities of individual tokens using Bayes formula.

A good filter should classify all the mails of the training data with zero misclassifications, by predicting the likelihood of a mail being spam. Classification errors can be displayed as a matrix given below.

		Predicted	
Actual		Spam	Ham
	Spam	True Positive (TP)	False Negative (FN)
	Ham	False Positive (FP)	True Negative (TN)

While the percentage of misclassification is a simple measure of performance of the filter, there are other measures like Sensitivity, Specificity, odds radio and ROC curve analysis.

[1] Department of Computer Science, SV University, Tirupati. India, E-mail: srikanth.kadainti@gmail.com.
[2] Department of Computer Science, SV University, Tirupati. India, E-mail: drsramakrishna@yahoo.com.
[3] Department of Statistics, SV University, Tirupati. India, E-mail: sarma_kvs@rediffmail.com.

In this paper, we re-examine the Naïve Bayesian filter by identifying some important statistical features of the filter. We also report the results of our experiment with Enron data set and carry out statistical analysis and establish the merits of the new method.

2. THE BAYESIAN FILTER

Let E denote the email having n tokens w_1, w_2,...w_n. Assuming that all the tokens are independent the probability of receiving mail E denoted by P(E) is equal to the probability of receiving all the tokens. It means P(E) = P(w_1,w_2, ...w_n). The unconditional probability of receiving the mail given by $P(E) = \prod_{i=1}^{n} P(w_i)$. Define two classes of mails S and H indicating spam and ham mails available in the training data.

Let P(E|S) = P(E comes from class S) and P(E|H) = P(E comes from class H) denote the conditional probability that mail E has come from S and H respectively. Thus

$$P(E|S) = \prod_{i=1}^{n} P(w_1|S)$$

and

$$P(E|H) = \prod_{i=1}^{n} P(w_1|H)$$

These two conditional probabilities help in estimating the token probabilities in the hash table.

Let P(S) and P(H) denote the probability of having spam or ham in the dataset. Let n_s and n_h denote the number of spam and ham mails in the training data. Then the estimates of probabilities are given by

$$P(S) = \frac{n_s}{n_s + n_h} \text{ and } P(H) = \frac{n_h}{n_s + n_h} \quad ...(1)$$

By Bayes theorem we have

$$P(w_i|s) = \frac{P(w_i \cap S)}{P(S)} \text{ and } P(w_i|H) = \frac{P(w_i \cap H)}{P(H)} \quad ...(2)$$

We now need the posterior probability of a spam mail given the overall probability of the mail, denoted by

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

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

and

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

Now P(S|E) > P(H|E) implies that the mail E is more likely to be a spam than a ham.

Consider the ratio $Z = \frac{P(S|E)}{P(H|E)}$

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

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

Taking logarithms on both sides, we get a linear function of the form

$$Z = Z_1 + Z_2 \quad ...(3)$$

where $Z_1 = \log\left\{\frac{P(S)}{P(H)}\right\}$ and $Z_2 = \sum_{i=1}^{n} \log\left\{\frac{P(w_i|S)}{P(w_i|H)}\right\}$.

Given the mail E we compute Z and classify E it into spam if Z > 0 and ham otherwise. We first note that Z_1 does not depend the individual token probabilities of E and remains constant for a given training data. The value of Z_2, however changes from mail to mail and thus it is possible to create data on Z_2 for each mail of the training set. The statistical properties of Z_2 have some impact on the classification.

A token w_i in a mail is said to be more spammy than hammy, if $P(w_i|S) > P(w_i|H)$. Sometimes a token w_i may not be found in the list of the training data. Graham (2002) calls them innocent tokens and for each such token it is taken that $P(w_i|S) = P(w_i|H) = 0.5$. Such probabilities are called hapaxial probabilities. In general if $P(w_i|S) = P(w_i|H)$ for all the tokens then $\log\left\{\frac{P(w_i|S)}{P(w_i|H)}\right\}$ becomes zero for that mail and such tokens do not contribute to the score. Consider the following propositions.

Proposition-1

Let w_H and w_S denote the set of spammy words and hammy words in the given mail E. Then $Z_{21} = \sum_{i \in w_S} \log\left\{\frac{a_i}{b_i}\right\}$ and $Z_{22} = \sum_{i \in w_H} \log\left\{\frac{a_i}{b_i}\right\}$ where $a_i = P(w_i|S)$ and $b_i = P(w_i|H)$. If $Z_{21} > Z_{22}$ then E tends to be spammy.

Proposition-2

Let n_w denote the number of tokens in the mail E. Then as n_w increases, the score Z_2 also increases leading to a positive correlation between n_w and Z_2. The marginal contribution of n_w to Z_2 can be estimated by simple linear regression model of the form $Z_2 = b_0 + b_1 n_w$ where b_0 and b_1 are regression coefficients.

Proposition-3

The classification score Z can be restrained from becoming large positive or negative, by replacing Z_2 with $\hat{Z}_2 = (Z_2 - b_1 n_w)$.

Then the modified Bayesian score is of the form $Z^{\wedge} = Z_1 + Z^{\wedge}_2$. If $Z^{\wedge} > 0$ the mail is classified as spam and ham otherwise.

3. DATA AND THE EXPERIMENT

The proposed method is studied with an experiment using Enron email data sets having 1324 mails in which 322 were spam and 1002 were hams. Each mail contained the content only and not the subject, address or the java script.

An Access database was created to store each mail with its mail number and known class (Spam = 1, Ham = 0). A visual basic code was developed to tokenize each mail along with the frequency for each token. Tokens having only numerals were not considered for inclusion in the list. This has lead to a hash table with 3271 tokens along with number of times a token appeared in spam or ham groups. Since $n_s = 322$ and $n_h = 1022$ we get $P(S) = \left\{\frac{322}{(322+1022)}\right\}$ = 0.239583 and P(H) = 1 – P(S) = 0.760417. For each token, the joint probability of appearance in the mail, given that it is a spam is calculated. The conditional probability of w_i given S and W_i given H are computed using (2). This table is the basis for classification and a partial list of tokens is shown in Table 1.

Table 1
Partial List of Tokens and Their Probabilities Obtained from the Training Data.

Mail No.	Token	Appeared in Spam	Appeared in Ham	P(W∩S)	P(W∩H)	P(W|S)	P(W|H)
1	a	296	134	.01143	.00518	.04700	.00684
2	call	214	40	.00827	.00155	.03400	.00205
3	you	188	51	.00726	.00197	.02985	.00260
2300	idiot	1	1	.00004	.00004	.00016	.00005
2301	I;m	1	1	.00004	.00004	.00016	.00005

All the 1324 mails have been used to build the model with Bayesian rule. Each mail E is split into tokens and the token probabilities are captured from table 1. If a token is not found in the list then $P(W_i|S) = P(W_i|H)$ is taken as 0.5. The following algorithm is used to classify the mails.

Algorithm

1. Read i^{th} mail from the training data.
2. Tokenize and count the number of distinct words (n_{wi}) in the i^{th} mail.
3. Calculate Z_{1i}, Z_{2i} and $Z_i = Z_{1i} + Z_{2i}$.
4. If $Z_{2i} > 0$ then define Predicted Class = 1 (Spam) else Predicted Class = 0 (Ham). Post these values into a table of statistics in the access database.
5. Calculate True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) counts for further analysis.
6. Find the percentage of misclassification over the mails classified in the training data.

For instance, the mail with id 1324 has the following content.

"Congratulations- Thanks to a good friend U have WON the £2,000 Xmas prize. 2 claim is easy, just call 08712103738 NOW! Only 10p per minute. BT-national-rate".

The tokens and their probabilities for this mail are shown in Table 2.

Table 2
Tokenized Mail with Probabilities.

Token	Appeared in Spam	Appeared in Ham	P(Wi|S)	P(Wi|H)	Radio = $\log\left\{\frac{P(w_1 \mid S)}{P(w_1 + H)}\right\}$
-	0	0	0.50	0.50	0.00
08712103738	0	0	0.50	0.50	0.00
10p	0	0	0.50	0.50	0.00
2	0	0	0.50	0.50	0.00
BT-national-rate	0	0	0.50	0.50	0.00
Congratulations	0	0	0.50	0.50	0.00
NOW	168	89	0.03	0.00	1.77

contd.

Only	62	22	0.01	0.00	2.17
Thanks	0	0	0.50	0.50	0.00
U	0	0	0.50	0.50	0.00
WON	0	0	0.50	0.50	0.00
Xmas	0	0	0.50	0.50	0.00
a	296	134	0.05	0.01	1.93
call	214	40	0.03	0.00	2.81
claim	214	40	0.03	0.00	2.81
easy	214	40	0.03	0.00	2.81
friend	0	0	0.50	0.50	0.00
good	0	0	0.50	0.50	0.00
have	101	34	0.02	0.00	2.23
is	163	82	0.03	0.00	1.82
just	0	0	0.50	0.50	0.00
minute.	0	0	0.50	0.50	0.00
per	0	0	0.50	0.50	0.00
prize.	0	0	0.50	0.50	0.00
the	149	78	0.02	0.00	1.78
to	149	78	0.02	0.00	1.78
£2000	149	78	0.02	0.00	1.78

The sum of ratios from the last column gives Z_2 = 23.6927. Since Z_1 = –1.1352 we get Z = –1.1352 + 23.6927 = 22.55755, which is positive and the mail is classified as spam. Now a partial list of statistics of Z scores along with the actual and predicted class of each mail is shown in Table 3.

Table 3
Z Scores for Different Mails.

Mail No.	Actual Class	Predicted Class	Token Count	Z1	Z2	Z
1	1	1	24	–1.1352	26.93261	25.7974
2	1	1	23	–1.1352	43.88958	42.75437
3	1	1	28	–1.1352	26.94681	25.81161
4	1	1	27	–1.1352	27.60934	26.47414
5	1	1	27	–1.1352	26.94572	25.81052
951	0	1	7	–1.1352	2.617787	1.482585
952	0	1	14	–1.1352	6.334186	5.198984
953	0	1	10	–1.1352	7.146559	6.011358
954	0	1	14	–1.1352	14.2507	13.11549
1321	1	1	19	–1.1352	39.02078	37.88558
1322	1	0	5	–1.1352	0	–1.1352
1323	1	1	19	–1.1352	18.59558	17.46038
1324	1	1	27	–1.1352	23.69275	22.55755

In the following section the statistical behavior of Z and its relationship with token count is explored using SPSS. Analyzing Z is equivalent to analyzing Z_2 since Z_1 is a constant for a given training data.

4. ANALYSIS OF Z_2 STATISTIC

The frequency distribution shown in figure 1(a) indicates that Z_2 scores of spam mails are nearly normally distributed between 0 and 73.7 and has a mean of 32.03 and standard deviation of 13.7. In the ham class the distribution is not normal (more likely exponential) but lies between 0 and 28.6 with a mean of 5.6 and a standard deviation of 4.7. Thus spam mails appear to have higher positive score than ham mails, leading to a predicted class dominantly as spam. Hence the token count could be an influencing factor for the filter. The results of classification are stored into another table named as ROC table in the same database.

With this behavior of Z_2, the Naïve Bayesian method leads to about 64% of misclassification among which 849 are false positives (hams as spams) and only 2 spams are misclassified as hams. The proportion of positive cases classified by the rule among all the positive cases is called sensitivity of the filter. This is given by SN = TP/(TP + FN). Similarly SP = TN/(TN + FP) is the proportion of negative cases classified by the rule among all the negative cases. Both SN and SP take values between 0 and 1. They independently measure the classification ability of the filter. However if SN = 0 then 0.5 is added so SN and if SN = 1 then 0.1 is subtracted from SN to obtain feasible values of SN. Similarly if SP = 0 then 0.1 is added to SP and when SP = 1 then 0.5 is subtracted from SP. This correction helps avoid overflow error in calculations.

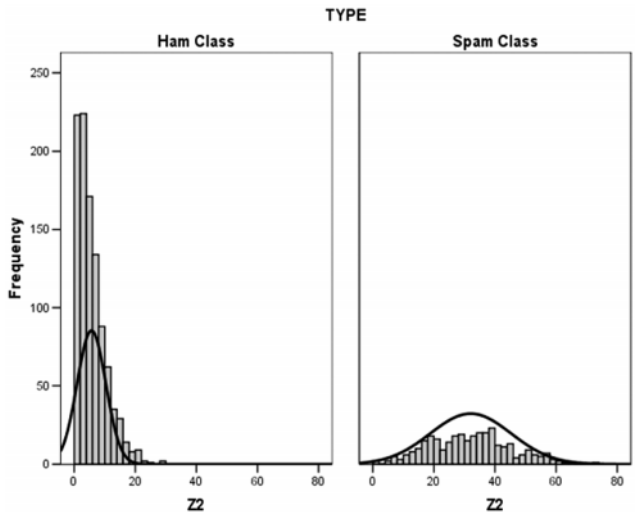

Figure 1(a): Distribution of Z_2 Scores.

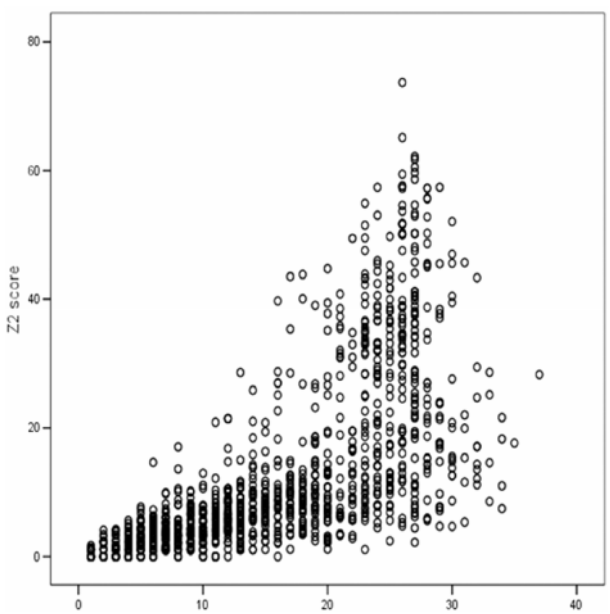

Figure 1(b): Scatter of Z_2 Scores Against Token Count.

A combined measure of SN and SP is called Receiver Operating Characteristic (ROC) curve analysis in which, for different filters, a plot of SN against (1-SP) is drawn and the area covered under the curve above the diagonal line is called the Area Under the Curve (AUC). Higher the AUC more will be the discriminating power of the filter. In the case of binary classification a different measure called Diagnostic Odds Ratio (DOR) is adopted which is given by

$$DOR = \frac{TP/FN}{FP/TN} \quad ...(4)$$

It can be shown that the AUC is related to DOR by the following formula (Afina et al (2003))

$$AUC = \frac{\mu(\mu-1)-\log(\mu)}{(\mu-1)^2}, \text{ where } \mu = DOR. \quad ...(5)$$

Afina et al (2003) have shown that the DOR is another indicator of the performance of the classification filter and its value ranges from 0 to ∞, with higher values indicating better discriminatory power of the filter. A value of 1 means that the filter does not discriminate between spams and hams and a value less than 1 indicates improper filter with more hams among the spams. Two other measures of performance are Precision = TP/(TP + FP) and Recall = TP/(TP + FN). In ROC curve analysis AUC is a popularly used measure of performance and in this case AUC = 0.9108.

With the available Z scores, the Naïve Bayesian filter has a high sensitivity of 0.9937 indicating that spams are accurately predicted. The low specificity of 0.1526 indicates that several ham mails are misclassified into spams leading to a poor prediction of hams. A comparison of these measures between the Naïve Bayesian method and the Modified Bayesian method is given in Table 4.

A close look at the Z_2 scores given in Table 3 indicate that Z_2 is higher for mails having higher the token count and vice versa.

The scatter diagram between token count and Z_2 score shown in figure 1(b) indicates a positive relationship between token count, n_w and Z_2.

A linear regression model has been fitted and the estimated model is $Z_2 = -4.372 + 1.098\ (n_w)$ with R-square = 0.485. The model is statistically significant (p = 0.00012). The regression coefficient 1.098 indicates that for every increase of one token in the mail E, the Z_2 score marginally increases by 1.098.

In the following section we make use of this relationship and propose a modified Z_2 statistic.

5. MODIFIED Z_2 STATISTIC

The linear effect of token count on the Z_2 can be removed by defining $Z^{\wedge}_2 = (Z_2 - b_1 n_w)$. In the classification code, the following change has been made and program is executed.

```
rs4.Open "select * from temp1", db1, adOpenStatic, adLockOptimistic
z2 = 0
Do While Not rs4.EOF()
   Dim a, b As Double
   a = rs4!pwgs
   b = rs4!pwgh
   z2 = z2 + rs4!ratio
        'z3 = z1 + z2 – 1.098 * wc' wc is the word count
   If z3 > 0 Then
        rs1!New_Type = "1"
   Else
        rs1!New_Type = "0"
   End If
  rs4.Update
  rs4.MoveNext
Loop
rs4.Close
```

The results of this modified Bayesian procedure classification as shown in Table 4.

Table 4
Summary Measures of Z_2 Statistic for Naïve Bayesian and Modified Naïve Bayesian Methods.

Parameter	Naïve Bayesian	Modified Bayesian
TP	320	217
FN	2	105
FP	849	11
TN	153	991
Total mails	1324	1324
% Misclassified	64.27	8.761
Sensitivity	0.9937	0.6739
Specificity	0.1526	0.9890
DOR	28.8339	186.1878
AUC	0.9108	0.9770

Interestingly the misclassification percent has come down to 8.76% and the FP has drastically come down. False negatives however, have increased. The new model has 67% of sensitivity and 98% of specificity. The odds ratio is also very high in this case compared to the results given in Table 4. The AUC indicates that the probability is 0.977 that a randomly selected spam mail has Z_2 value higher than that of a randomly selected ham mail.

In the following section the model is tested on randomly selected subsets from the training data.

6. RANDOM TESTING

A portion of training data is usually set apart for checking the validity of the model. If n mails are available only n' are used for training and remaining (n – n') are used for checking the validity of the model. Instead of using a predetermined sub set of data we propose random sets of various sizes drawn from the training data for checking the model testing with different sample sizes. The following code has been added to accommodate this new module in our main program.

```
Private Sub Command1_Click() 'generating random samples
db1.Execute "delete * from enron1"
'rs1.Open "select * from enron1" , db1, adOpenStatic, adLockOptimistic
Dim x1, x2, samples As Integer
Dim cont As String
Dim rn, i As Integer
samples = InputBox("How many samples?")
'MsgBox samples
For i = 1 To samples
   If Rnd(1000) = 0 Then
   rn = Int(Rnd(1000) * 1324) + 1
   Else
   rn = Int(Rnd(1000) * 1324)
End If
If rn > 0 Then
   rs0.Open "select * from enron where mno = val('" & rn & "')", db1, adOpenStatic, adLockOptimistic
   Open "select * from enron1 where mno = val('" & rn & "')", db1, adOpenStatic, adLockOptimistic
   If rs1.RecordCount > 0 Then 'drop this condition if duplicates are allowed in the samples
      rs1.AddNew
      rs1!mno = x1
      rs1!Type = x2
      rs1!content = x3
      rs1.Update
      Print i, rn, rs0!mno, rs0!Type
   End If
   rs0.Close
 rs1.Close
```

End If

Next

End Sub

With a target sample of size n, the usual procedure for randomization may get repeated random numbers leading to redundancy. As an option, we can avoid these duplicate mails in which case, the actual set will be n* (≤ n). The Modified Bayesian method is illustrated with different sample sizes calculated with and without redundancy and the results are shown in Table 5. The mean, standard deviation (sd) and the coefficient of variation (CV = sd/ mean * 100) are found for all the sample results.

From the Table 5 it can be seen that the average misclassification is 10.38% (sd = 0.93%, CV = 8.99) without redundancy while it is 10.39% (sd = 1.23%, CV = 11.80)) when redundancy is allowed. The percentage of misclassification against random sample size is shown in figure 2.

Table 5
Comparison of Performance Statistics Under Random Sampling.

	Without redundancy									With redundancy							
Trial	n	n*	TP	FP	FN	TN	% mis	DOR	AUC	n*	TP	FP	FN	TN	% mis	DOR	AUC
1	50	49	9	4	0	36	8.13	2.25	0.6322	50	10	3	0	37	6.00	3.33	0.6914
2	100	96	9	8	0	79	8.33	1.125	0.5196	100	9	8	0	83	8.00	1.125	0.5196
3	200	187	18	18	2	149	10.69	74.500	0.9541	200	17	20	1	162	10.50	137.70	0.9710
4	250	229	20	23	3	183	11.35	53.043	0.9414	250	20	25	2	203	10.80	81.19	0.9569
5	300	270	27	25	3	215	10.37	77.399	0.9554	300	26	29	2	243	10.33	108.93	0.9654
6	350	308	30	29	2	246	10.39	84.827	0.9583	350	30	32	2	286	9.71	134.062	0.9704
7	400	339	38	31	3	267	10.03	109.096	0.9654	400	37	39	2	322	10.25	152.743	0.9732
8	500	404	49	37	6	312	10.64	68.864	0.9514	500	51	50	7	392	11.4	57.1199	0.9444
9	600	473	61	43	10	359	11.20	50.927	0.9397	600	66	61	12	461	12.16	41.566	0.9305

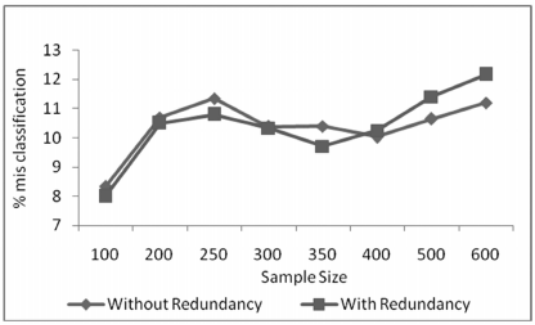

Figure 2: Percentage Misclassification During Random Testing.

Due to lower CV, the method of sampling without redundancy is relatively more consistent than the other method. The DOR in all cases is larger than unity indicating that this method has a high discriminating power between spams and hams. It may be recalled that the DOR was 186.18 with the Modified Bayesian method when all the 1324 mails are considered whereas it was only 28.83 with the Naïve Bayesian method.

7. CONCLUSIONS

In this paper we observe that in case of content based filtering the Naïve Bayesian rule can be modified by taking into account the correlation structure between the number of tokens and the sum of log ratio of the conditional probabilities (Z_2). The average misclassification is around 10% and the AUC is more than 0.90 for the new method.

REFERENCES

[1] Afina S. Glas, Jeroen G. Lijmer, Martin H. Prins, Gouke J. Bonsel, Patrick M.M. Bossuyt, (2003), "The Diagnostic Odds Ratio: A Single Indicator of Test Performance", Journal of Clinical Epidemiology, 56, pp. 29-1135

[2] Prabhakaran Raghavan and Christopher Manning., "Text Classification The Naïve Byaes Algorithm", 2003.

[3] Sahami, M. Dumais, S., Heckerman, D., and Horvitz, E. (1998), "A Bayesian Approach to Filtering Junk Email", In learning for Text Classification – Papers from AAAI workshop, pp. 55-62, Madision Wisconsin. AAAI Technical Report WS-98-05, 1998.

[4] Spam Assassin, (2005) http://www.spamassassin.org/index.html.

[5] SpamBayes, (2002) http://spambayes.sourceforge.net/.

[6] Tim Peter, August 2002. http://mail.ptyhon.org/pipermail/pythondev/2002-August/028216.html.

[7] Paul Graham., "A Plan for Spam", 2002. www. paulgraham.com/spam.html.