

## FEATURE SELECTION AND COMPARISON OF TWO NAÏVE BAYES CLASSIFICATION METHODS IN THE CONTEXT OF SPAM FILTERING

Liny Varghese<sup>1</sup>, Supriya M.H<sup>1</sup> and K. Poullose Jacob<sup>1</sup>

<sup>1</sup>Cochin University of Science and Technology, Cochin, India,

E-mail: liny@cusat.ac.in, E-mail: supriya@cusat.ac.in, E-mail: kpj@cusat.ac.in.

### ABSTRACT

Treating e-mail filtering as a binary text classification problem, researchers have applied several statistical learning algorithms to email corpora with promising results. This paper examines the performance of a Naive Bayes classifier using different approaches to feature selection and tokenization on different email corpora.

**Keywords:** Spam filtering, Naïve Bayes, Bernoulli model.

### 1. INTRODUCTION

Spam, also known as “unsolicited commercial e-mail” or “junk e-mail,” pollutes the communication medium of electronic mail. With the proliferation of direct marketers on the Internet and the increased availability of massive email address mailing lists, the volume of junk mail has grown enormously in the past few years. Recipients of spam have to waste time for deleting such annoying and possibly disgusting messages. When a user is troubled with a large amount of spam, the chance of overlooking a legitimate message increases as also spam creates overload on mail servers and Internet traffic. Legislative efforts to curb spam have been ineffective or counter-productive as spam accounts for more than two thirds of the mails received in a year.

This brings in the relevance of a spam filter which is a program that can be used to detect unsolicited and unwanted email and prevent those messages from getting to a user’s inbox. A spam filter looks for certain criteria on which it bases judgments but need not be effective, too often omitting perfectly legitimate messages (these are called false positives) and letting actual spam through. There are several different approaches to spam filtering, like firewalls, mail servers & email clients filters. Good results are found with similar approaches. But as spammers are finding new ways to send spam, maintenance of this lists/rulebase becomes very tricky. Another approach examines the content of an incoming message for features which indicate its status as spam or legitimate. This approach is mostly applied at the user level, and can incorporate facts about each user’s legitimate mail.

Treating e-mail filtering as a binary text classification problem, researchers have applied several statistical learning algorithms to email corpora with promising results [2, 3, 4]. This paper examines the performance of a Naive Bayes classifier using two different approaches to

feature selection and tokenization as well as on different corpus sizes.

### 2. APPROACH

Naive Bayes has several advantageous properties than other algorithms due to their simplicity, linear computational complexity, and their accuracy. This classifier can be constructed by a single scan through the training data and classification requires just a single table lookup per token, plus a final product or sum over each token. Most other approaches require iterated evaluation. Storage requirements are small in Naive Bayes because we need to store only the token counts, rather than whole messages. The classifier can be updated incrementally as new messages arrive. This study examines two variants of Naive Bayes text classification algorithm. Each method makes the independence assumption that the probability of tokens occurring in a message is independent.

### 3. MODELS FOR NB CLASSIFIER

Among different ways to setup an NB classifier, Multinomial and Bernoulli model models are analysed in this study. Multinomial model, which generates one term from vocabulary in each position of the document, and assumes generative model. Bernoulli model generates an indicator for each term of the vocabulary, 1 for presence of term and 0 for absence. The Bernoulli model has the same time complexity as the multinomial model.

### 4. FEATURE SELECTION

The three methods considered basic feature selection algorithm are mutual information, the  $X^2$  test and frequency models; the frequency model was selected since this is the simplest method. Frequency is defined as the document frequency for Bernoulli model and collection frequency for the multinomial model. No dimensionality reduction methods are studied.

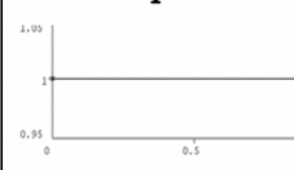
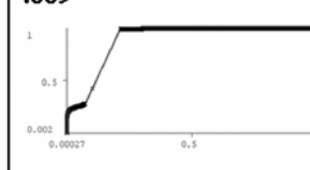
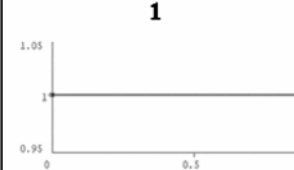
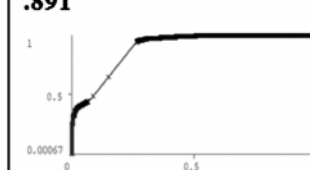
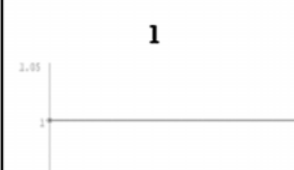
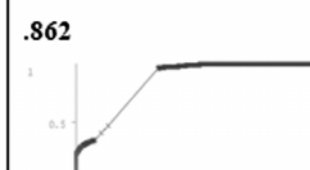
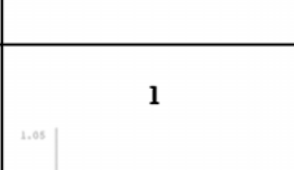

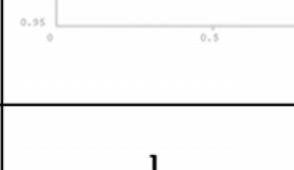
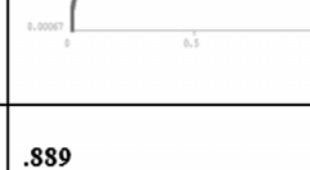


Data sets	Bernoulli model			Multinomial model		
	Spam Recall (Sensitivity)	Ham Recall (Specificity)	ROC curve	Spam Recall (Sensitivity)	(Specificity)	ROC curve
<b>Dataset 1 (enron1)</b> Legitimate : 3672 Spam : 1500	<b>1</b>	<b>1</b>	<b>1</b> 	<b>.242</b>	<b>.966</b>	<b>.889</b> 
<b>Dataset 2 (enron2)</b> Legitimate: 4361 Spam: 1496	<b>1</b>	<b>1</b>	<b>1</b> 	<b>.392</b>	<b>.973</b>	<b>.891</b> 
<b>Dataset 3 (enron3)</b> Legitimate: 4012 Spam: 1500	<b>1</b>	<b>1</b>	<b>1</b> 	<b>.304</b>	<b>.441</b>	<b>.862</b> 
<b>Dataset 4 (enron4)</b> Legitimate: 1500 Spam: 4500	<b>1</b>	<b>1</b>	<b>1</b> 	<b>.996</b>	<b>.734</b>	<b>.91</b> 
<b>Dataset 5 (enron5)</b> Legitimate: 1500 Spam: 3675	<b>1</b>	<b>1</b>	<b>1</b> 	<b>.97</b>	<b>.616</b>	<b>.889</b> 
<b>Dataset 6 (enron6)</b> Legitimate: 1500 Spam: 4500	<b>1</b>	<b>1</b>	<b>1</b> 	<b>.979</b>	<b>.629</b>	<b>.887</b> 

Figure 1: Naïve Bayes Implementation Results: Bernoulli Model and Multinomial Model.

## 5. TOKENIZATION

While tokenising the following factors are considered Extract header information: From, time, subject and body of the message. The body is unrestricted in its content. (b) Attachments are ignored (c) HTML tags are stripped off from the message. (d) Punctuation marks are ignored. (5) Stop words removed, No stemming applied (e) hyphens spaces are replaced with space character (f) Special characters (@,\$,! ..) are retained. Finally 'space' is used as the delimiting character to tokenize<sup>5</sup>.

Each mail is considered as a single document. In the experiments, each message is represented as a vector  $(t_1, t_2, \dots, t_m)$ , where  $t_1, \dots, t_m$  are the values of attributes  $T_1, \dots, T_m$  and  $m$  is the total number of tokens. In the Bernoulli model, all the attributes are Boolean:  $X_i = 1$  if the message contains the token; otherwise,  $X_i = 0$ . In multinomial model, attribute values are term frequencies (TF), showing the token frequency. Attributes with TF values carry more information than Boolean ones. Sometimes NB with TF attributes outperforms than those with Boolean attributes. A third alternative is called normalized TF, is to divide term frequencies by the total number of token occurrences in the message.

The document frequencies/collection frequency are computed and for feature selection (to create the final term-frequency matrix), only the tokens with document frequency greater than 10 (for Bernoulli) and collection frequency greater than 10 (for multinomial) are considered for training. This number is selected by heuristics, to reduce the number of attributes.

## 6. NAÏVE BAYES CLASSIFIER

Applying Bayes theorem in spam filtering context,

$$P(\text{Class} + \text{Email}) = \prod P(\text{Class} | \text{token})$$

For each token find:

$$P(\text{Class} | \text{token}) = \frac{P(\text{token} | \text{Class}) \times P(\text{Class})}{P(\text{token})}$$

where

Class = {Spam, legitimate} Email = {tokens}.

$P(\text{Class} | \text{Email})$  = Probability of the spam/legitimate given the tokens in an email.

$P(\text{token} | \text{Class})$  = Probability of the tokens in an spam/legitimate email (from training set).

$P(\text{Class})$  = Probability of the spam/legitimate emails.

$P(\text{token})$  = Probability of the tokens.

## 7. METHODOLOGY

This study evaluates two versions of *Naïve Bayes* mentioned above experimentally on six non-encoded

datasets [6]. The methodology used in this paper consisted of practical work involving several experiments which is supported by some theoretical background. To carry out the experiments test environments in Perl and WEKA were used. The attributes selected for each dataset are tested against Bernoulli NB and Multinomial NB. The Naïve Bayes is run with 10 fold cross-validation.

## 8. EXPERIMENTAL RESULTS

In the experiment, spam recall ( $TP / (TP + FN)$ ) and ham recall ( $TN / (TN + FP)$ ) are used for evaluation. Spam recall is the proportion of spam messages that the filter managed to identify correctly (how much spam it blocked), whereas ham recall is the proportion of ham messages that passed the filter. The results are shown in the following pages. A ROC space is defined by FPR and TPR as  $x$  and  $y$  axes respectively.

In Enron datasets 4, 5 and 6 spam recall is much higher than the ham recall, for multinomial method. The ham recall can be increased if we change the feature selection cutoff values and the equation to incorporate the difference between collection frequencies of spam and ham. In datasets 1 and 2 ham recall is greater than spam recall. But for Bernoulli model, ham recall = spam recall = 1 for all the six datasets. Based on the results from these six datasets we can say that Bernoulli model outperforms multinomial model in the context of spam filtering. The results show that, the Bernoulli model performs very well than multinomial model. Boolean values are used for Bernoulli model. This indicates that the number of times a word repeats in a message is not important, its presence is important to detect its class.

## 9. CONCLUSION AND FUTURE WORK

Spam filtering with two different versions of the Naive Bayes (NB) classifier are discussed and evaluated experimentally. The Bernoulli model and multinomial model are included in the analysis. To accommodate the current trends in future and to filter spam efficiently, periodic updation of corpora is required. To do the experiments around 1000 words are used. This number can be reduced by applying some dimensionality reduction methods like PCA. Improving on stop words and selecting document frequency, the system can perform very well on the problem of spam.

## REFERENCES

- [1] Robert Hall., "A Countermeasure to Duplicate-Detecting Anti-Spam Techniques", Technical Report 99.9.1, AT&T Research Labs, 1999.
- [2] Sahami, M., Dumais S., Heckerman D., and Horvitz, E. (1998), "A Bayesian Approach to Filtering Junk e-mail", *Learning for Text Categorization*, Papers from the AAAI Workshop, Madison Wisconsin, pp. 55-62.

- [3] Paul Graham, "A Plan for Spam", 2002. [www.paulgraham.com/spam.html](http://www.paulgraham.com/spam.html).
- [4] Paul Graham, "Better Bayesian Filtering", *In Proceedings of the First Annual Spam Conference*, MIT, 2003. [www.paulgraham.com/better.html](http://www.paulgraham.com/better.html).
- [5] Jon Kågström, "Improving Naive Bayesian Spam", Mid Sweden University, *Department for Information Technology and Media*, Spring 2005.
- [6] "The Enron-Spam Datasets are Available from <http://www.iit.demokritos.gr/skel/i-config/> and <http://www.aueb.gr/users/ion/publications.html> in both raw and pre-processed form.
- [7] Vangelis Metsis, Ion Androutsopoulos, Georgios Paliouras, "Spam Filtering with Naive Bayes - Which Naive Bayes?", *CEAS 2006 Third Conference on Email and AntiSpam*, July 27-28, 2006, Mountain View, California USA.