Email Classification with Naïve Bayes

Mitchell Della Marta, Shu Han Bor

May 3, 2014

1 Introduction

1.1 Aim

To implement text classification of emails using the Naïve Bayes classifier for spam detection.

1.2 Importance

As email is one of the main forms of communication, spam detection to remove spam is important. Email spam is an unsolved global issue that negatively affects productivity and uses up valuable resources. Spam emails are usually sent with malicious intent to users for some monetary gain. Email users spend a large amount of time regularly deleting these spam emails, which needlessly occupy storage space and consumes bandwidth. Hence the development of classifiers that are able to separate legitimate from spam emails is required. Classification of emails to separate and remove spam from other emails, aims to save the user from having to spend time and effort sorting through and deleting spam, and protect unaware users from malware.

Traditional non-machine learning based techniques are easily fooled by spammers and require periodic manual updates. Therefore, machine learning based methods are required, which are able to automatically analyse email contents and dynamically update themselves to cope with new spamming techniques.

2 Data Preprocessing

Before classification can be performed, we must represent the files in our sample set *Lingpspam-mini600* appropriately. Using the "bag-of-words" model, words are extracted from the file and treated as features. There are two main characteristics in an email; the subject and the body. Thus, we will construct two "bags-of-words", one for each component. To determine which corpus a feature will belong to, we if the line begins with "Subject:". If it does the words within the line will be added to the subject corpus, otherwise they will be added to the body corpus.

To each file, we performed the following steps:

1. Replace all punctuation and special symbols with a space.
2. Remove stop words, using a list of stop words "english.stop".
3. Remove numbers or words containing digits.

The tokenisation schemes used in the example data are different to that used in our stop words list. The data detaches clitics and contraction affices, whereas the stop words list is tokenised based on whitespace and punctuation characters. Therefore, we chose to replace punctuation with a space (e.g. "n't" becomes "n", "t") instead of simply removing them. This meant that the stop word list would remove clitics as it removes single characters, instead of keeping it as a feature as ("nt" isn't in the stop words list).

The *document frequency* feature selection method was then used to select the 200 words which occured in the most documents, and used to build a classifier for the given documents. The following steps were taken to accomplish this:

1. Counter is created for each email, which keeps track of the number of times each word appears in that email.
2. For each word that appeared in the counter for an email, it is added once to the counter for the corresponding subcorpus.
3. In each corpus the words with the top 200 document frequency score are selected as features to represent that corpus.

Note that different runs of our preprocessing script can result in different classifier accuracies. As the counter is stored in a hashmap, words with the same frequency before and after the 200 word cutoff are chosen arbitrarily. Depending on the words chosen, the accuracy of the classifier can be affected.

Each selected feature was then weighted using its *td-idf* score [1].

$$tfidf(t_k, d_j) = \#(t_k, d_j) \times \log \frac{|Tr|}{\#Tr(t_k)}$$

Then the scores were normalised, using *cosine normalisation* [1].

$$w_{kj} = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|}(tfidf(t_s, d_j))^2}}$$

2.1 Data Characteristics

Table I: Characteristics of dataset

	Subject	**Body**
# Features before removing stop words	1074	19886
# Features after removing stop words	915	19386
# $Class_{nonspam}$	600	600
# $Class_{spam}$	200	200

The “bag-of-words" model produced 19886 and 1074 features in the body and subject corpora respectively. After the removal of stop words, there were 19386 unique words remaining in the body corpus, and 915 words in the subject corpus (see Table I).

3 Feature Selection

Feature selection is performed using document frequency. This method involves computing the number of documents a word occurs in, for every word, and selecting the top 200 words with the highest scores to build a classifier.

Shown in Table II and Table III are the top 100 words for the subject and body corpora respectively and their document frequency score. Removing stop words filters out extremely common words that have little value in classification. It is beneficial in text processing, as it removes low quality features, allowing more significant features to have precedence. Most resulting words are sensible, although there are some aren't (e.g. “qs”, “ll”, “ca”). However, removing these words results in a lower accuracy. Therefore, we have chosen to retain these as some of them may be help the classifier distinguish between spam and non-spam emails. Thus, given our task to process natural language text, the selection of words shown makes sense as it gives a better representation of the contents of the emails, and helps improve the accuracy of the classifier.

A comparison of Table II and Table III shows significant disparities in the document frequency of features and word distribution. The frequencies for the subject is significantly lower than that of the body. Whilst some features are shared between subject and body, most features selected are different.

Table II: Top 100 words in subject corpus and corresponding document frequency (DF) scores

Rank	Word	Score	Rank	Word	Score	Rank	Word	Score
1	sum	30	35	speaker	6	69	intuitions	4
2	summary	26	36	german	6	70	banning	4
3	english	24	37	internet	6	71	school	3
4	language	21	38	business	6	72	resolution	3
5	free	20	39	list	5	73	ary	3
6	disc	19	40	resources	5	74	adjectives	3
7	query	18	41	native	5	75	verbal	3
8	linguistics	15	42	research	5	76	teaching	3
9	comparative	13	43	word	5	77	future	3
10	sex	13	44	spanish	5	78	lists	3
11	opposites	12	45	linguist	5	79	background	3
12	words	12	46	jobs	5	80	synthetic	3
13	book	10	47	needed	5	81	credit	3
14	email	10	48	grammar	5	82	home	3
15	call	9	49	software	5	83	live	3
16	job	9	50	languages	5	84	youthese	3
17	method	9	51	time	5	85	uniformitarianism	3
18	japanese	8	52	fwd	4	86	released	3
19	correction	8	53	summer	4	87	names	3
20	syntax	7	54	address	4	88	opportunity	3
21	program	7	55	books	4	89	decimal	3
22	qs	7	56	information	4	90	world	3
23	chinese	7	57	request	4	91	misc	3
24	announcement	7	58	phonetics	4	92	sites	3
25	million	7	59	pig	4	93	double	3
26	part	6	60	american	4	94	acquisition	3
27	slip	6	61	programs	4	95	site	3
28	workshop	6	62	unlimited	4	96	policy	3
29	armey	6	63	web	4	97	fall	3
30	money	6	64	www	4	98	teach	3
31	lang	6	65	secrets	4	99	hey	3
32	conference	6	66	great	4	100	line	3
33	dick	6	67	read	4			
34	mail	6	68	systems	4			

Table III: Top 100 words in body corpus and corresponding document frequency (DF) scores

Rank	Word	Score	Rank	Word	Score	Rank	Word	Score
1	information	205	35	message	91	69	full	74
2	language	192	36	ll	89	70	system	74
3	mail	183	37	receive	88	71	ac	73
4	university	179	38	check	88	72	today	73
5	time	178	39	phone	88	73	questions	72
6	list	171	40	good	87	74	remove	72
7	address	165	41	day	86	75	interest	72
8	english	159	42	interested	86	76	john	71
9	linguistics	156	43	year	86	77	found	70
10	http	156	44	include	85	78	related	70
11	people	146	45	working	85	79	site	69
12	send	146	46	case	85	80	linguist	69
13	free	144	47	based	84	81	usa	69
14	make	140	48	ve	84	82	text	68
15	email	133	49	note	83	83	point	68
16	number	128	50	home	83	84	read	68
17	work	128	51	made	83	85	ago	67
18	www	122	52	part	83	86	book	67
19	languages	119	53	including	81	87	week	67
20	find	118	54	mailing	81	88	making	66
21	fax	116	55	type	80	89	dear	66
22	order	108	56	give	79	90	cost	66
23	call	103	57	program	79	91	question	65
24	form	101	58	web	79	92	simply	65
25	research	100	59	place	79	93	received	63
26	state	99	60	special	78	94	offer	63
27	linguistic	99	61	line	78	95	general	63
28	subject	98	62	date	78	96	important	62
29	years	98	63	days	77	97	data	62
30	world	98	64	back	76	98	ca	61
31	contact	97	65	internet	76	99	long	61
32	de	96	66	american	75	100	summary	61
33	money	94	67	service	75			
34	word	91	68	business	74			

4 Subject vs Body Analysis

4.1 Results

We ran out preprocessed data against a variety of Weka classifiers (ZeroR, OneR, 1-NN, 3-NN, NB, DT and MLP) tested with their 10 fold cross validation, as well as against our Naïve Bayes classifier (MyNB) tested with our 10 fold cross validation method (Table IV).

Table IV: Various classifiers tested with 10 fold cross validation for both the subject and body corpora

	Accuracy (%)	
Classifier	**Subject**	**Body**
ZeroR	66.67	66.67
OneR	70.00	82.00
1-NN	79.00	87.17
3-NN	68.17	85.00
NB	68.50	94.83
DT	66.67	92.50
MLP	76.17	96.67
MyNB	80.83	95.33

4.2 Discussion

4.2.1 Comparison of Classifiers

The ZeroR classifier returns the majority class given from the training data. As a result ZeroR returns "nonspam" as there are 400 non-spam and 200 spam cases. Thus ZeroR produces a correct answer for and only for test data of the majority class, "nonspam", which is as obtained.

The multilayer perceptron (MLP) defined as "consists of multiple layers of simple, two-state, sigmoid processing elements (nodes) or neurons that interact using weighted connections." (Pal 1992 p,.684). The Weka multilayer perceptron assigns the features as output neurons with a random weight. Weka's default number of hidden neurons was used, $(attributes + classes)/2 = (200+2)/2 = 101$[2]. The network is then trained using back propagation, correcting the weights to minimize the error in the entire output.

is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs. A MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function.

In contrast, Multilayer perceptron (MLP) returns the best accuracy in both cases.

OneR generates one rule for each item in the dataset. After calculating rules for each, it will select the rule that gives the smallest total error. Although it is simple, it only performs slightly worse than more complicated algorithms such as 1-NN or 3-NN.

1-NN and 3-NN are both types of the k-nearest neighbour algorithm, where $k = 1$ and $k = 3$ respectively. Although they use similar algorithms, 1-NN performs better than 3-NN. 1-NN takes one neighbour, whereas 3-NN takes three. Therefore, this disparity is probably because in most cases the closest neighbour correctly classifies the new example, but the next two closest do not. Hence 1-NN gets a greater accuracy than 3-NN.

4.2.2 Comparison of myNB vs Weka's NB

We conduct a paired t-test in our comparison to determine whether the differences in accuracies between our's and Weka's Naïve Bayes was statistically significant, with a confience level of 95%. To carry out the paired t-test:

1. Calculate the difference between the two classifiers in each fold

$$|c_{1_i} - c_{2_i}| \text{ for } i \in [1, \ldots, N]$$

   where $c_{1_i} \in C_1$ and $c_{2_i} \in C_2$ for all i.
2. Calculate the sample standard deviation of the differences:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

   where $X = \{x_1, \ldots, x_N\}$ are the observed differences.
3. Calculate the $(1-\alpha)$-upper confidence interval (UCL) of the mean:

$$\text{UCL}_{1-\alpha} = \bar{X} \pm t_{(1-\alpha)(k-1)} s_{N-1}$$

   for the confidence level $(1-\alpha) = 0.95$.
4. If the interval $\text{UCL}_{0.95}$ contains 0 then the difference in accuracies is not statistically significant; otherwise, it is.

Table V: Comparison of fold accuracies between our Naïve Bayes classifier (C_1) and Weka's (C_2) for the subject corpus

Fold Number	C1 (%)	C2 (%)	Difference
Fold 1	71.67	58.33	13.34
Fold 2	66.67	56.67	10.00
Fold 3	58.33	66.67	8.34
Fold 4	53.33	71.67	18.34
Fold 5	63.33	65.00	1.67
Fold 6	70.00	78.33	8.33
Fold 7	66.67	80.00	13.33
Fold 8	71.67	65.00	6.67
Fold 9	70.00	76.67	6.67
Fold 10	70.00	78.33	8.33
Mean	66.17	69.67	9.5

Table VI: Comparison of fold accuracies between our Naïve Bayes classifier (C_1) and Weka's (C_2) for the body corpus

Fold Number	C1 (%)	C2 (%)	Difference
Fold 1	98.33	98.33	0.00
Fold 2	98.33	93.33	5.00
Fold 3	96.67	95.00	1.67
Fold 4	90.00	100.00	10.00
Fold 5	91.67	95.00	3.33
Fold 6	96.67	93.33	3.33
Fold 7	93.33	95.00	1.67
Fold 8	95.00	91.67	3.33
Fold 9	96.67	91.67	5.00
Fold 10	96.67	95.00	1.67
Mean	95.33	94.83	3.50

Using the differences calculated in Table V for the subject corpus we find that:

$$
\begin{aligned}
s &= 6.86 \\
\text{UCL}_{0.95} &= \bar{X} \pm t_{0.95,9} \cdot s_{N-1} \\
&= 12.34 \pm 2.262 \times 6.86 \\
&= 12.34 \pm 15.52 \\
&= [-3.18, 27.85]
\end{aligned}
$$

Similarly, given the differences calculated in Table VI, we obtain for body corpus:

$$\begin{aligned} s &= 2.77 \\ \text{UCL}_{0.95} &= \bar{X} \pm t_{0.95,9} \cdot s_{N-1} \\ &= 3.50 \pm 2.262 \times 2.77 \\ &= 3.50 \pm 6.27 \\ &= [-2.77, 9.77] \end{aligned}$$

For the subject corpus $\text{UCL}_{0.95} = [-3.18, 27.85]$ and for body corpus $\text{UCL}_{0.95} = [-2.77, 9.77]$. In both cases the interval $\text{UCL}_{0.95}$ contains 0. Therefore, the difference between the two Naïve Bayes classifiers for both subject and body are not statistically significant.

4.2.3 Comparison Between Subject and Body Corpora

The overall performance of the classifiers regardless of type, perform significantly better on the body than the subject corpus. This suggests that the email body is a better indicator of content.

5 Challenge Analysis

5.1 Results

5.2 Discussion

6 Conclusion and Future Work

The tokens used in our classifier are formed from single words. Therefore, it will not analyse common consecutive words that are found in spam emails, leading to the failure of our classifier from detecting these emails.

By taking into account permutations of consecutive words, or words that appear within a specified distance of each other, the accuracy of our Bayesian classifier could be increased.

7 Reflection

8 Instructions: How to run code

References

[1] Fabrizio Sebastiani, *Machine learning in automated text categorization*. ACM Computing Surveys, 34(1):1-47, 2002.

[2] Ware, M n.d., *Class MultilayerPerceptron*, Revision 10169, University of Waikato, New Zealand, viewed 30 April 2014, <http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/MultilayerPerceptron.html>.