TREC 2005 Spam Track Overview

Gordon Cormack and Thomas Lynam
University of Waterloo
Waterloo, Ontario, Canada

Abstract

TREC's *Spam Track* introduces a standard testing framework that presents a chronological sequence of email messages, one at a time, to a spam filter for classification. The filter yields a binary judgement (*spam* or *ham* [i.e. non-spam]) which is compared to a human-adjudicated *gold standard*. The filter also yields a *spamminess* score, intended to reflect the likelihood that the classified message is spam, which is the subject of post-hoc ROC (Receiver Operating Characteristic) analysis. The gold standard for each message is communicated to the filter immediately following classification. Eight test corpora – email messages plus gold standard judgements – were used to evaluate 53 subject filters. Five of the corpora (the *public* corpora) were distributed to participants, who ran their filters on the corpora using a track-supplied toolkit implementing the framework. Three of the corpora (the *private* corpora) were not distributed to participants; rather, participants submitted filter implementations that were run, using the toolkit, on the private data. Twelve groups participated in the track, submitting 44 filters for evaluation. The other nine subject filters were variants of popular open-source implementations adapted for use in the toolkit in consultation with their authors.

1 Introduction

The spam track's purpose is to model an email spam filter's usage as closely as possible, to measure quantities that reflect the filter's effectiveness for its intended purpose, and to yield repeatable (i.e. controlled and statistically valid) results.

Figure 1 characterizes an email filter's actual usage. Incoming email messages are received by the filter, which puts them into one of two files - the ham[1] file (*in box*) and the spam file (*quarantine*). The user regularly reads the ham file, rejects any spam messages (which have been misfiled by the filter), and reads or otherwise deals with the remaining ham messages. The human may also report the misfiled spam to the filter. Occasionally (perhaps rarely or never) the spam file is searched for ham messages that have been misfiled. The human may also report such ham misfilings to the filter. The filter may use this feedback, as well as external resources such as blacklists, to improve its effectiveness.

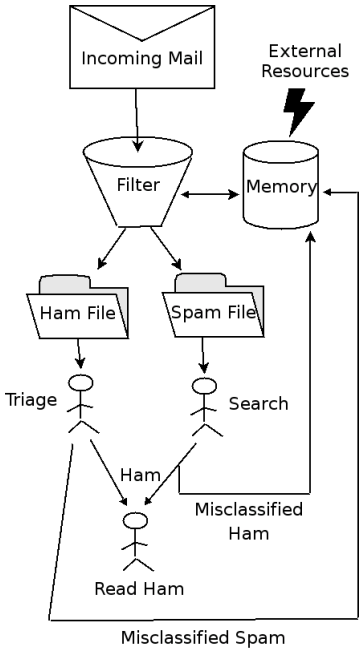

Figure 1: Real Filter Usage

The filter's effectiveness for its intended purpose has two principal aspects: the extent to which ham is placed in the ham file (not the spam file) and the extent to which spam is placed in the spam file (not the ham file). It is convenient to quantify the filter's failures in these two aspects: the *ham misclassification percentage* (*hm%*) is the fraction of all ham delivered to the spam file; the *spam misclassification percentage* (*sm%*) is the fraction of all spam delivered to the ham file. A filter is effective to the extent that it minimizes both ham and spam misclassification; however, the two have disparate impacts on the user. Spam misclassification reflects directly the extent to which the filter falls short of its intended purpose – to detect spam. Spam misclassification inconveniences and annoys the user, and may, by cluttering the ham file, cause the user to overlook important messages. Ham misclassification, on the other hand, is an undesirable side-effect of spam filtering. Ham misclassification inconveniences the user and risks loss of important messages. This risk is difficult to quantify as it depends on (1) how likely the user is to notice a ham misclassification, and (2) the importance to the user of the misclassified ham. In general, ham

[1] *Ham* denotes non-spam. Spam is defined to be "Unsolicited, unwanted email that was sent indiscriminately, directly or indirectly, by a sender having no current relationship with the recipient."

[2] An analogy may be drawn with automobile safety and fuel efficiency standards. *Deaths per 100 million km* and *litres per 100 km* are used to measure these aspects of automobile design. It is desirable to minimize both, but dimensionally meaningless to sum them or to combine them by some other linear formula.

misclassification is considerably more deleterious than spam misclassification. Because they measure qualitatively different aspects of spam filtering[2], the spam track avoids quantifying the relative importance of ham and spam misclassification.

There is a natural tension between ham and spam misclassification percentages. A filter may improve one at the expense of the other. Most filters, either internally or externally, compute a score that reflects the filter's estimate of the likelihood that a message is spam. This score is compared against some fixed threshold t to determine the ham/spam classification. Increasing t reduces *hm%* while increasing *sm%* and vice versa. Given the score for each message, it is possible to compute *sm%* as a function of *hm%* (that is, *sm%* when t is adjusted to as to achieve a specific *hm%*) or vice versa. The graphical representation of this function is a Receiver Operating Characteristic (ROC) curve; alternatively a recall-fallout curve. The area under the ROC curve is a cumulative measure of the effectiveness of the filter over all possible values. ROC area also has a probabilistic interpretation: the probability that a random ham will receive a lower score than a random spam. For consistency with *hm%* and *sm%*, which measure failure rather than effectiveness, spam track reports the area *above* the ROC curve, as a percentage ($(1 - ROCA)\%$).

For the reasons stated above, accuracy (percentage of correctly classified mail, whether ham or spam) is inconsistent with the effectiveness of a filter for its intended purpose[3], and is not reported here. A single quality measure, based only on the filter's binary ham/spam classifications, is nonetheless a desirable objective. To this end, spam track reports *logistic average misclassification percentage* (*lam%*) defined as $lam\% = logit^{-1}(\frac{logit(hm\%)+logit(sm\%)}{2})$ where $logit(x) = log(\frac{x}{100\%-x})$. That is, *lam%* is the geometric mean of the *odds* of ham and spam misclassification, converted back to a proportion[4]. This measure imposes no a priori relative importance on ham or spam misclassification, and rewards equally a fixed-factor improvement in the odds of either.

In addition to $(1 - ROCA)\%$ and *lam%*, which are threshold-neutral, the appendix reports *sm%* for various values of *hm%*, and *hm%* for various values of *sm%*. One of these statistics – *sm%* at $hamm\% = 0.1$ (denoted $h = .1$) – was chosen as indicative of overall filter effectiveness and included in comparative summary tables.

It may be argued that the filter's behaviour and the user's expectation evolve during filter use. A filter's classification performance may improve (or degrade) with use. A user may be more tolerant of errors that are made early in the filter's deployment. The spam track includes two approaches to measuring the filter's learning curve: (1) piecewise approximation and logistic regression are used to model *hm%* and *sm%* as a function of the number of messages processed; (2) cumulative *(1-ROCA)%* is given as a function of the number of messages processed.

In support of repeatability, the incoming email sequence and gold standard adjudications are fixed before filter testing. External resources are not available to the filters[5] during testing. For each measure and each corpus, 95% confidence limits are computed based on the assumption that the corpus was randomly selected from some source population with the same characteristics. *hm%* and *sm%* limits are computed using exact binomial probabilities. *lam%* limits are computed using logistic regression. *(1-ROCA)%* limits are computed using 100 bootstrap samples to estimate the standard error of $(1 - ROCA)\%$.

2 Spam Filter Evaluation Tool Kit

All filter evaluations were performed using the *TREC Spam Filter Evaluation Toolkit*, developed for this purpose. The toolkit is free software and is readily portable.

TREC 2005 participants were required to provide filter implementations for Linux or Windows implementing five command-line operations mandated by the toolkit:

- **initialize** – creates any files or servers necessary for the operation of the filter
- **classify** *message* – returns ham/spam classification and spamminess score for *message*
- **train ham** *message* – informs filter of correct (ham) classification for previously classified *message*
- **train spam** *message* – informs filter of correct (spam) classification for previously classified *message*
- **finalize** – removes any files or servers created by the filter.

[3] Optimizing accuracy incents filters to use threshold values that are clearly at odds with the their intended purpose.[3]

[4] For small values, odds and proportion are essentially equal. Therefore *lam%* shares much with the geometric mean average precision used in the robust track.

[5] Nevertheless, participants are at liberty to embed an unbounded quantity of prior data in their filter submissions. Within the framework it would be possible to capture and include blacklists, DNS servers, known-spam signatures, and so on, thus simulating many external resources.

Track guidelines prohibited filters from using network resources, and constrained temporary disk storage (1 GB), RAM (1 GB), and run-time (2 sec/message, amortized). These constraints were not rigidly enforced and, in the case of run-time, exceeded by orders of magnitude by some filters. Track guidelines indicated that the largest email sequence would not exceed 100,000 messages. This limit was exceeded as well – the largest consisted of 172,000 messages – but all filters appeared to be able to handle this size, given sufficient time. All but two participant filters – tamSPAM3 and tamSPAM4, which took 22 days and 12 days respectively to process the 49,000-message Mr. X corpus – were run on all corpora.

The toolkit takes as input a test corpus consisting of a set of email messages, one per file, and an index file indicating the chronological sequence and gold standard judgements for the messages. It calls on the filter to classify each message in turn, records the result, and communicates the gold standard judgement to the filter before proceeding to the next message.

The recorded results are post-processed by an evaluation component supplied with the toolkit. This component computes statistics, confidence intervals, and graphs summarizing the filter's performance.

3 Test Corpora

It is a simple matter to capture all the email delivered to a recipient or a set of recipients. Using this captured email in a public corpus, as for the other TREC tasks, is not so simple. Few individuals are willing to publish their email, because doing so would compromise their privacy and the privacy of their correspondents. A choice must be made between using a somewhat artificial public collection of messages and using a more realistic collection that must be kept private. The 2005 spam track explores this tradeoff by using both public and private collections. Participants ran their filters on the public data and submitted their results, in accordance with TREC tradition. In addition, participants submitted their filter implementations, which were run on private data by the proprietors of the data.

To form a test corpus, captured email must be augmented with gold-standard judgements. The track's definition of spam is *"Unsolicited, unwanted email that was sent indiscriminately, directly or indirectly, by a sender having no current relationship with the recipient."* The gold standard represents, as accurately as is practicable, the result of applying this definition to each message in the collection. The gold standard plays two distinct roles in the testing framework. One role is as a basis for evaluation. The gold standard is assumed to be *truth* and the filter is deemed correct when it agrees with the gold standard. The second role is as a source of user feedback. The toolkit communicates the gold standard to the filter for each message after the filter has been run on that message.

Human adjudication is a necessary component of gold standard creation. Exhaustive adjudication is tedious and error-prone; therefore we use a bootstrap method to improve both efficiency and accuracy. The bootstrap method begins with an initial gold standard G_0. One or more filters is run, using the toolkit and G_0 for feedback. The evaluation component reports all messages for which the filter and G_0 disagree. Each such message is re-adjudicated by the human and, where G_0 is found to be wrong, it is corrected. The result of all corrections is a new standard G_1. This process is repeated, using different filters, to form G_2, and so on, to G_n.

One way to construct G_0 is to have the recipient, in the ordinary course of reading his or her email, flag spam; unflagged email would be assumed to be ham. Or the recipient could use a spam filter and flag the spam filter's errors; unflagged messages would be assumed to be correctly classified by the filter. Where it is not possible to capture judgements in real time – as for all public collections to which we have access – it is necessary to construct G_0 without help from the recipient. This can be done by training a filter on a subset of the messages (or by using a filter that requires no training) and running the filter with no feedback.

3.1 Public Corpus – trec05p-1

Public Corpora

	Ham	Spam	Total
trec05p-1/full	39399	52790	92189
trec05p-1/ham25	9751	52790	62541
trec05p-1/ham50	19586	52790	72376
trec05p-1/spam25	39399	13179	52578
trec05p-1/spam50	39399	26283	65682

Private Corpora

	Ham	Spam	Total
Mr X	9038	40048	49086
S B	6231	775	7006
T M	150685	19516	170201
Total	165954	60339	226293

Table 1: Corpus Statistics

In the course of the Federal Energy Regulatory Commission's investigation, more than 1 million messages and files from the email folders of 150 Enron employees were released to the public. A digest of these files[6] was investigated as an email collection, but proved unsuitable as a large number of files did not appear to be email messages; those that were had been reformatted, deleting headers, markup, and attachments, and replacing original *message-ids* with synthetic ones. The files used in the collection were fetched directly from FERC **[4].** Of these files, some 100,000 were email messages with headers; however, only 43,000 had had a "Received:" line indicating that the headers were (more-or-less) complete. These 43,000 messages form the core of the trec05p-1 public corpus.

G_0 was constructed using Spamassassin 2.63 with user feedback disabled. Subsequent iterations used a number of filters – Spamassassin, Bogofilter, Spamprobe and crm114, interleaved with human assessments for all cases in which the filter disagreed with the current gold standard. This process identified about 5% of the messages as spam.

It was problematic to adjudicate many messages because it was difficult to glean the relationship between the sender and the receiver. In particular, the collection has a preponderance of sports betting pool announcements, stock market tips, and religious bulk mail that was initially adjudicated as spam but later re-adjudicated as ham. Advertising from vendors whose relationship with the recipient was tenuous presented an adjudication challenge.

During this process, the need arose to view the messages by sender; for example, once the adjudicator decides that a particular sports pool is indeed by subscription, it is more efficient and probably more accurate to adjudicate all messages from the same sender at one time. Similarly, in determining whether or not a particular "newsletter" is spam, it is desirable to identify all of its recipients. This observation occasioned the design and use of a new tool for adjudication – one that allows the adjudicator to use full-text retrieval to look for evidence and to ensure consistent judgements.

The 43,000 Enron messages were augmented by approximately 50,000 spam messages collected in 2005. The headers of these messages were altered so as to appear that they were delivered to the Enron mail server during the same time frame (summer 2001 through summer 2002). "To:" and "From:" headers, as well as the message bodies, were altered to substitute the names and email addresses of Enron employees for those of the original recipients. Spamassassin and Bogofilter were run on the corpora, and their dictionaries examined, to identify artifacts that might identify these messages. A handful were detected and removed; for example, incorrect uses of daylight saving time, and incorrect versions of server software in header information.

A final iteration of bootstrap process was effected to produce the final gold standard.

In addition to the full public corpus, four subsets were defined. These subsets use the same email collection and gold standard judgements, but include only a subset of the index entries so as to reflect different proportions of ham and spam. *trec05p-1/spam50* contains all of the ham and 50% of the spam from the full corpus; *trec05p-1/spam25* contains all of the ham and 25% of the spam. Similarly *trec05p-1/ham50* contains all of the spam and 50% of the ham, while *trec05p-1/ham25* contains all of the spam and 25% of the ham. All subsets were chosen at random. The numbers of ham and spam in each corpus are reported in table 1.

3.2 Private Corpus – Mr. X

The Mr. X corpus was created by Cormack and Lynam in 2004[3]. The email collection consists of the 49086 messages received by an individual, X, from August 2003 through March 2004. X has had the same email address for twenty years; variants of X's email address appear on the Web and in Usenet archives. X has subscribed to services and purchased goods on the Internet. X used a spam filter – Spamassassin 2.60 – during the period in question, and reported observed misclassifications to the filter. G_0 was captured from the filter's database. Table 2 illustrates the five revision steps forming G_1 through G_5, the final gold standard. $S \rightarrow H$ is the number of message classifications revised from spam to ham; $H \rightarrow S$ is the opposite. Note that G_0 had 421 spam messages incorrectly classified as ham. Left uncorrected, these errors would cause the evaluation kit to over-report the false positive rate of the filters by this a mount – more than an order of magnitude for the best filters. In other words, the results captured from user feedback alone – G_0 – were not accurate enough to form a useful gold standard. G_5, on the other hand, appears to be sufficiently accurate; systematic inspection of the 2004 results and of the 2005 spam track results reveals no gold standard errors – any that may persist do not contribute materially to the results.

3.3 Private Corpus – S. B.

The S. B. corpus consists of 7,006 messages (89% ham, 11% spam) received by an individual in 2005. The majority of all ham messages stems from 4 mailing lists (23%, 10%, 9%, and 6% of all ham messages) and private messages received from 3 frequent correspondents (7%, 3%, and 2%, respectively), while the vast majority of the spam messages (80%) are traditional spam: viruses, phishing, pornography, and Viagra ads.

	$S \to H$	$H \to S$
$G_0 \to G_1$	0	278
$G_1 \to G_2$	4	83
$G_2 \to G_3$	0	56
$G_3 \to G_4$	10	15
$G_4 \to G_5$	0	0
$G_0 \to G_5$	8	421
G_5	$\|H\| = 9038$	$\|S\| = 40048$

Table 2: Mr. X Bootstrap Gold Standard Iterations

Starting from a manual preclassification of all emails, performed when each message arrived in the mailbox, the gold standard was created by running at least one spam filter from each participating group and manually reclassifying all messages for which at least one of the filters disagreed with the preclassification. During this process, 95% of all spam messages and 15% of all ham messages were manually re-adjudicated, and reclassified as necessary. Genre classification was done using a mixture of email header pattern matching (for mailing lists and newsletters) and manual classification.

3.4 Private Corpus – T. M.

The T. M. corpus [7] includes personal email, from all accounts owned by an individual, including all mail received (except for spam filtered out by gmail to the gmail address). There are 170,201 messages in total. Messages were manually classified as they arrived, and the classifications were verified them by running his filter over the corpus and manually examining all false positives, false negatives and unsures until there were no more errors. Further verification was effected by running Bogofilter, SpamProbe, SpamBayes and CRM114 (in the TREC setup) over the corpus, manually examining all false positives and false negatives. The corpus ranges from Tue, 30 Apr 2002 to Wed, 6 Apr 2005.

4 Spam Track Participation

Group	Filter Prefixes
Beijing University of Posts and Telecommunications	kidSPAM1, kidSPAM2, kidSPAM3, kidSPAM4
Chinese Academy of Sciences (ICT)	ICTSPAM1, ICTSPAM2, ICTSPAM3, ICTSPAM4
Dalhousie University	dalSPAM1, dalSPAM2, dalSPAM3, dalSPAM4
IBM Research (Segal)	621SPAM1, 621SPAM2, 621SPAM3
Indiana University	indSPAM1, indSPAM2, indSPAM3, indSPAM4
Jozef Stefan Institute	ijsSPAM1, ijsSPAM2, ijsSPAM3, ijsSPAM4
Laird Breyer	lbSPAM1, lbSPAM2, lbSPAM3, lbSPAM4
Tony Meyer (Massey University in appendix)	tamSPAM1, tamSPAM2, tamSPAM3, tamSPAM4
Mitsubishi Electric Research Labs (CRM114)	crmSPAM1, crmSPAM2, crmSPAM3, crmSPAM4
Pontificia Universidade Catolica Do Rio Grande Do Sul	pucSPAM1, pucSPAM2, pucSPAM3
Universite Paris-Sud	azeSPAM1, azeSPAM2
York University	yorSPAM1, yorSPAM2, yorSPAM3, yorSPAM4

Table 3: Participant filters

The filter evaluation toolkit was made available in advance to participating groups. In addition to the testing and evaluation components detailed above, the toolkit included a sample public corpus, derived from the Spamassassin Corpus [10], and eight open-source sample filter implementations: Bogofilter [9], CRM114 [12], DSPAM [13], dbacl [1], Popfile [5], Spamassassin [11], SpamBayes [8], and Spamprobe [2].

Participating groups were required to configure their filters to conform to the toolkit, and to submit a pilot implementation which was run by the track coordinators on the supplied corpus and also on a 150-message sample of Enron email. Thirteen groups submitted pilot filters; results and problems with the pilot runs were reported back to these groups.

Each group was invited to submit up to four filter implementations for final evaluation; twelve groups submitted a total of 44 filters for final evaluation. Groups were asked to prioritize their submissions in case insufficient resources were available

Filter	*Run Prefix*	*Configuration*
Bogofilter	bogofilter	0.92.2
DSPAM	dspam-tum	3.4.9, train-until-mature
	dspam-toe	3.4.9, train-on-errors
	dspam-teft	3.4.9, train-on-everything
Popfile	popfile	0.22.2
Spamassassin	spamasasb	3.0.2, Bayes component only
	spamasasv	3.0.2, Vanilla (out of the box)
	spamasasx	3.0.2, Mr. X configuration
Spamprobe	spamprobe	1.0a

Table 4: Non-participant filters

to test all filters on all corpora, but it was not necessary to use this information – all but two of the 44 filters, mentioned above, were run on all private corpora.

Following the filter submissions, the public corpus trec05p-1 was made available to participants, who were required to run their filters, as submitted, on trec05p-1/full and submit the results. Participants were also encouraged to run their filters on the subset corpora.

All test runs are labelled with an identifier whose prefix indicates the group and filter priority and whose suffix indicates the corpus to which the filter is applied. Table 3 shows the identifier prefix for each submitted filter.

4.1 Non-participant Runs

For comparison, revised versions of the open-source filters supplied with the toolkit were run on the spam track corpora. The authors of three – crm114, dbacl, and Spambayes – were spam track participants. The authors of the remaining five – Bogofilter, DSPAM, Popfile, Spamassassin, and Spamprobe were approached to suggest revisions or variants of their filters. These versions were tested in the same manner as the participant runs. Table 4 illustrates each non-participant filter.

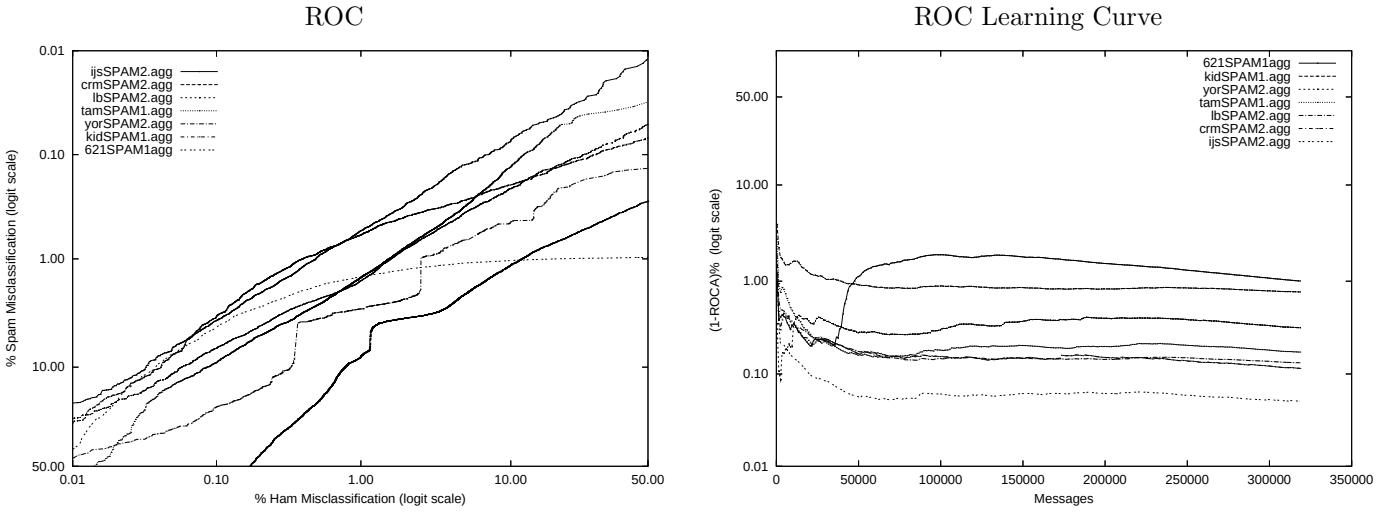

Figure 2: Aggregate

4.2 Aggregate Runs

The subject filters were run separately on the various corpora. That is, each filter was subject to (up to) eight test runs. The four full corpora – trec05p-1/full, mrx, sb, and tm – provide the primary results for comparison. For each filter, and *aggregate run* was created combining its results on the four corpora as if they were one. The evaluation component of the

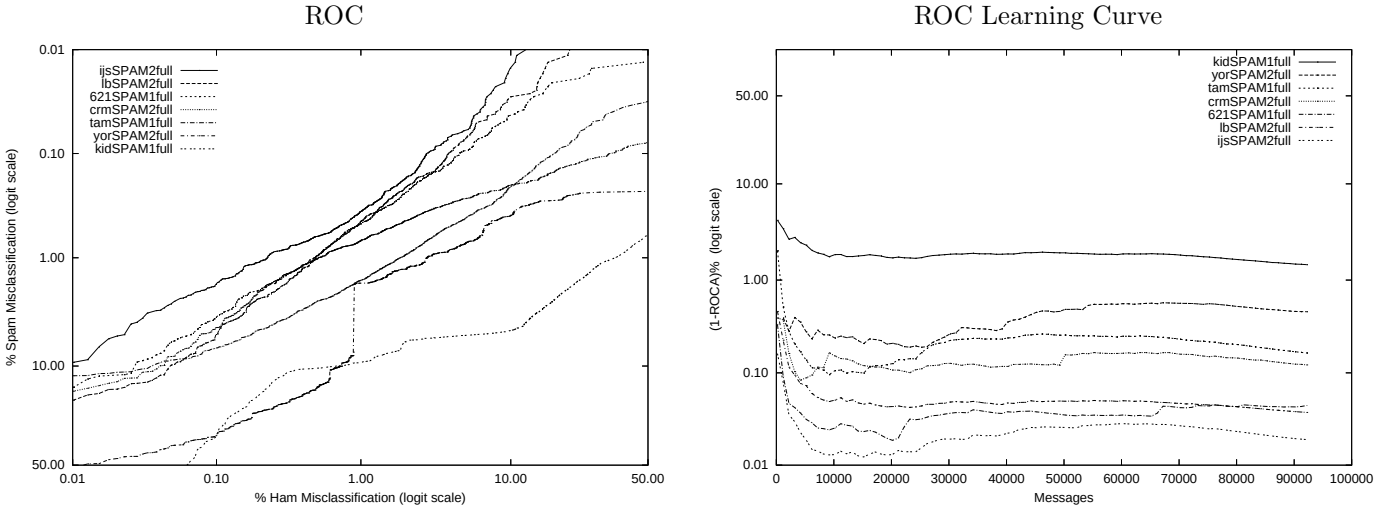

Figure 3: trec05–1/full

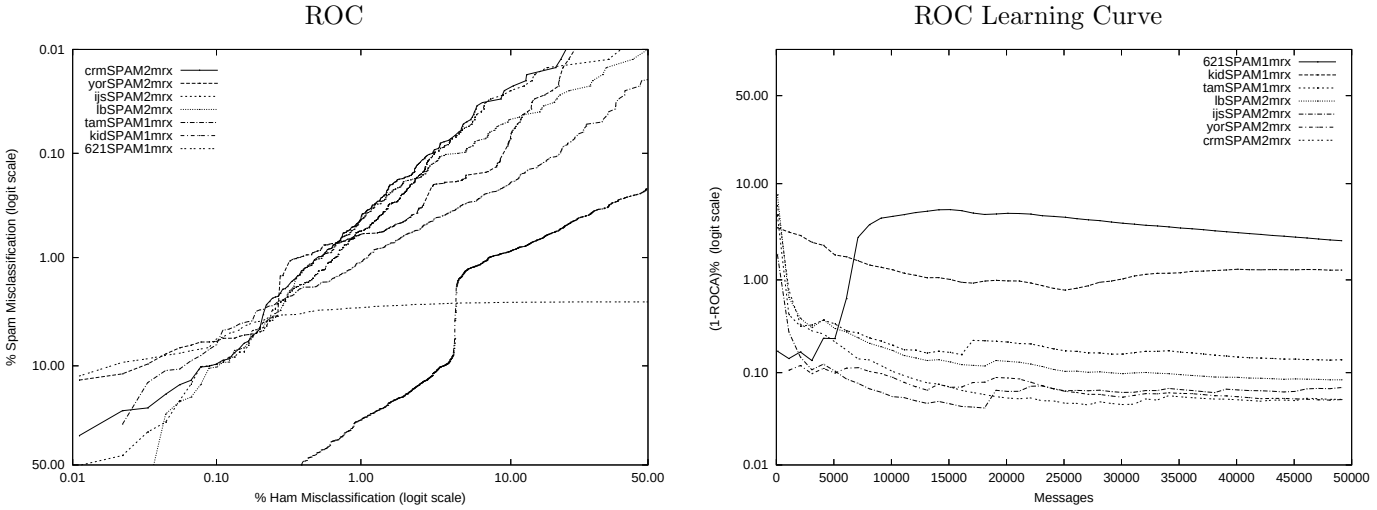

Figure 4: Mr X

toolkit was run on the aggregate results, consisting of 318,482 messages in total – 113,129 spam and 205,253 ham. The summary results on the aggregate runs provide a composite view of the performance on all corpora.

5 Results

Table 5 presents the three measures of the binary classification measures: *hm%, sm%,* and *lam%.* Table 6 presents three summary measurements of filter quality – *(1-ROCA)%, h=.1%, and lam%.* Table 7 shows the relative ranks achieved by the filters according to each of the fifteen summary measures. The tables show each filter's performance on each of the four full corpora, and in the aggregate, ordered by aggregate *(1-ROCA)%.* More detailed results for each run, including confidence limits and graphs, may be found in the notebook appendix.

Figure 2 shows the ROC curves for the best seven participant runs ranked by (1-ROCA)%, and restricted to one run (the best) per participant. ijsSPAM2 dominates the other curves over most regions. However, if one considers the intercept with the 0.10% ham misclassification line, crmSPAM2 is slightly (but not significantly) higher. This difference is reflected in the different rankings shown in table 7. It may be argued that this intercept accurately reflects the usefulness of the filter for its intended purpose. On the other hand, a broad ROC curve may be argued to reflect good filtering performance. Indeed, the crm group indicated that the falloff of the curve was due to a bug they discovered in the course of their TREC

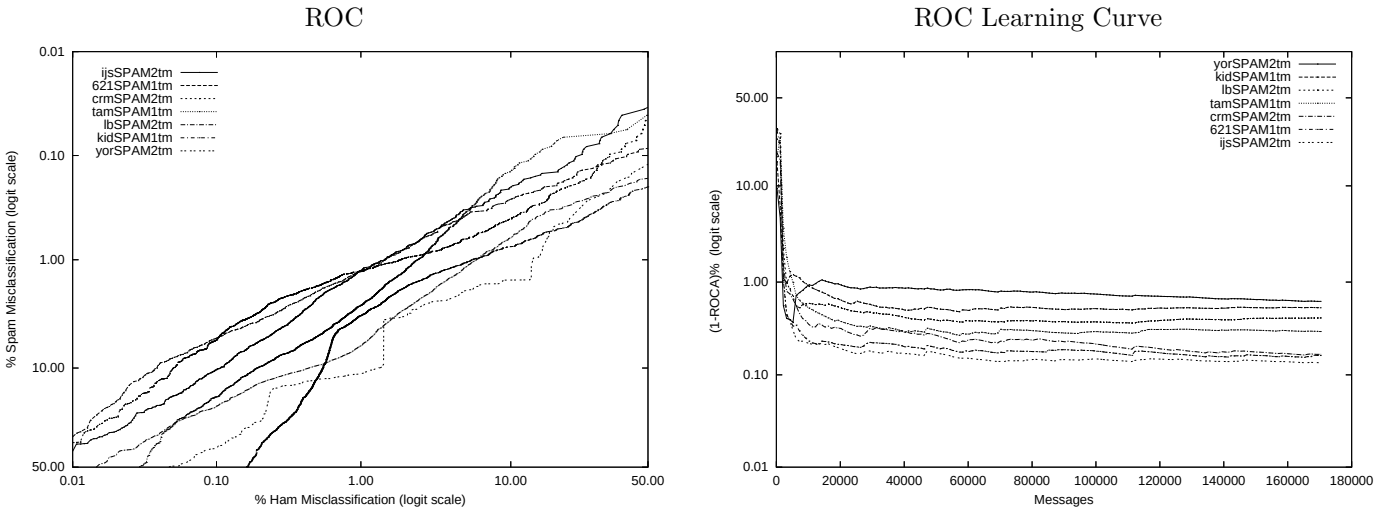

Figure 5: T. M.

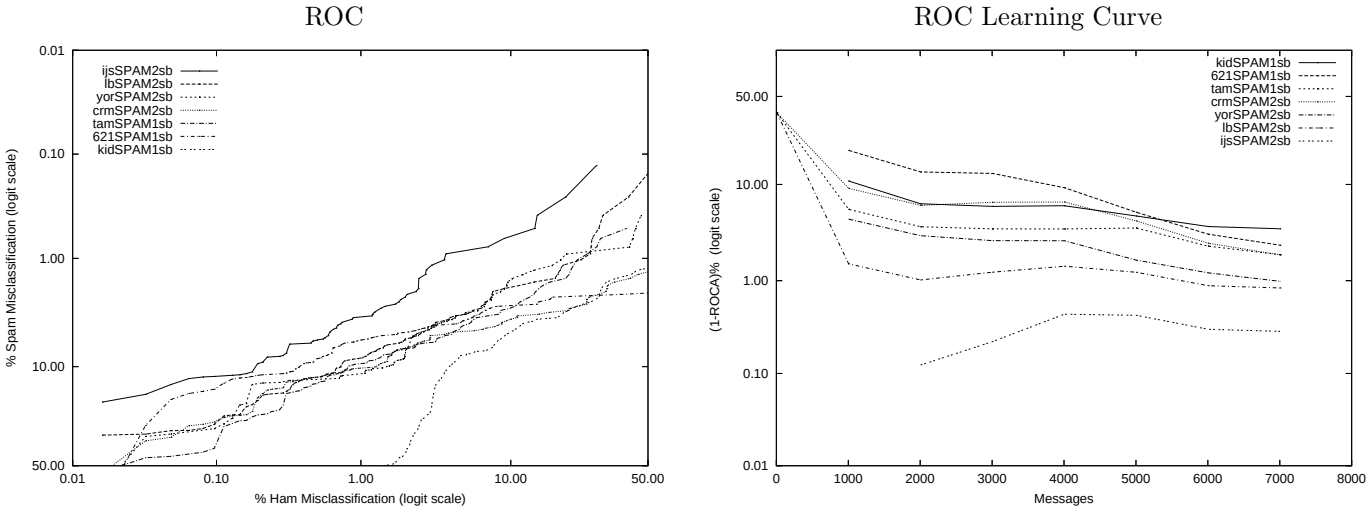

Figure 6: S. B.

participation. 621SPAM1 demonstrates a severe falloff, also due to a bug – this filter failed on every message larger than 100KB. Figures 3 through 6 show the curves for the same filters on the four primary corpora. Figure 7 shows the ROC curves for the non-participant aggregate runs; additionally, for comparison, the best participant run.

Learning curves for the aggregate and four major corpora are also shown in figures 2 through 6. These curves show (1-ROCA)% as a function of the number of messages classified. The curves appear to indicate that the filters have reached steady-state performance. Instantaneous ham and spam learning curves for each run are given in the notebook appendix.

Table 11 gives a *genre* classification for each misclassified message in the S. B. Corpus. Genre classification may be useful to assess the impact of misclassification; for instance, a misclassified personal message or a message from a frequent correspondent is more likely to have serious negative consequences than a misclassified newsletter article. In addition, genre classification may give insight into the nature of messages that are difficult to classify. The ham genres are:

- *Automated.* Sent by software to the recipient, perhaps as part of an Internet transaction.
- *Commercial.* Commercial email not considered spam.
- *Encrypted.* Personal or other sensitive email, sent in an encrypted format.
- *Frequent.* Email from a frequent correspondent.

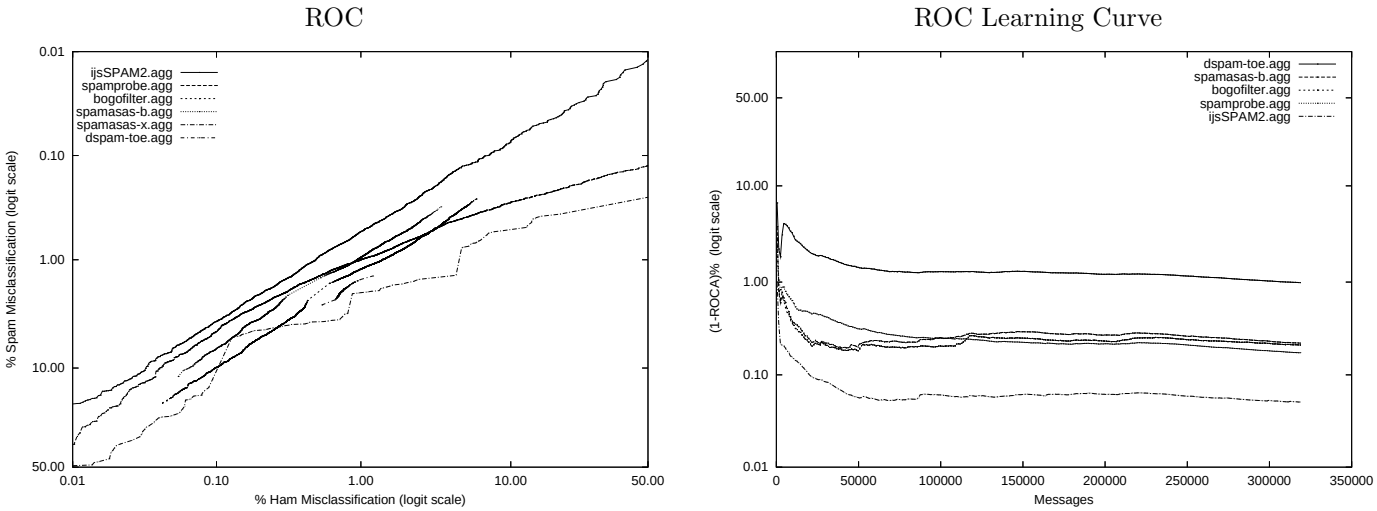

Figure 7: Non-participant Aggregate

- *List.* Email from a mailing list
- *Newsletter.* Message from a subscribed-to news service.
- *Personal.* Personal individual correspondence.

The spam genres are:

- *Automated.* Unwelcome messages sent automatically to the recipient.
- *List.* Spam delivered via a mailing list to which the recipient is subscribed.
- *Newsletter.* An unwelcome newsletter to which the recipient did not subscribe.
- *Phishing.* Fraudulent email misrepresenting its origin or purpose.
- *Sex.* Pornography or other sexually-related spam.
- *Virus.* An email message containing a virus.

6 Conclusions

Notwithstanding a few operational issues which occasioned extensions to deadlines, relaxation of limits, and patches to filters, the submission mechanism worked satisfactorily. Participants submitted filters to the track, and also ran the same filters on public data received by the track. The public corpus appears to have yielded comparable results to those achieved on the private corpora – preliminary analysis shows that the statistical differences between the results on private and public corpora appear to be no larger than those among the private corpora. This observation contradicts the authors' prior prediction, which was that large anomalies would be apparent in the public corpus results. Further post-hoc analysis will likely uncover some artifacts of the public corpus that worked either to the filters' advantage or disadvantage.

The results presented here indicate that content-based spam filters can be quite effective, but not a panacea. Misclassification rates are easily observable, even with the smallest corpus of about 8,000 messages. The results call into question a number of public claims both as to the effectiveness and ineffectiveness of "Bayesian" and "statistical" spam filters.

The filters did not, in general, appear to be seriously disadvantaged by the lack of an explicit training set. Their error rates converged quickly, and the overall misclassification percentages were not dominated by early errors. In any event, the use of a training set would have been inconsistent with the track objective of modelling real usage as closely as possible.

TREC 2005 did not afford the filters on-line access to external resources, such as black lists, name servers, and the like. Participants could have included, but did not, archived versions of these resources with their submissions. No aspect of

the toolkit or evaluation measures precludes the use of on-line resources; privacy and repeatability considerations excluded them at TREC. The efficacy of these resources remains an open question, notwithstanding public claims in this regard.

The public corpus will be made generally available, subject to a standard TREC usage agreement that proscribes disclosure of information that would compromise its utility as a test corpus. It may be desirable, before the corpus is made generally available, to use it in another round of blind testing.

7 Acknowledgements

The authors thank Tony Meyer and Stefan Buettcher for their invaluable contributions to this effort.

References

[1] Breyer, L. Laird breyer's free software - dbacl. http://www.lbreyer.com/gpl.html, 2005.

[2] Burton, B. Spamprobe - a fast bayesian spam filter. http://spamprobe.sourceforge.net, 2002.

[3] Cormack, G., and Lynam, T. A study of supervised spam detection applied to eight months of personal email. http://http://plg.uwaterloo.ca/ gvcormac/spamcormack.html, 2004.

[4] FERC. Information released in enron investigation. http://fercic.aspensys.com/members/manager.asp, 2003.

[5] Graham-Cumming, J. Popfile. http://popfile.sourceforge.net/, 2004.

[6] Klimt, B., and Yang, Y. Introducing the enron corpus. In *First Conference on Email and Anti-Spam (CEAS)* (2004).

[7] Meyer, T. Email classification. http://www.massey.ac.nz/ tameyer/research/spambayes/index.html, 2005.

[8] Peters, T. Spambayes: Bayesian anti-spam classifier in python. http://spambayes.sourceforge.net/, 2004.

[9] Raymond, E. S., Relson, D., Andree, M., and Louis, G. Bogofilter. http://bogofilter.sourceforge.net/, 2004.

[10] spamassassin.org. The spamassassin public mail corpus. http://spamassassin.apache.org/publiccorpus, 2003.

[11] spamassassin.org. Welcome to spamassassin. http://spamassassin.apache.org, 2005.

[12] Yerazunis, W. S. CRM114 - the controllable regex mutilator. http://crm114.sourceforge.net/, 2004.

[13] Zdziarski, J. A. The DSPAM project. http://www.nuclearelephant.com/projects/dspam/, 2004.

	Aggregate			trec05p-1/full			Mr. X			S. B.			T. M.		
Filters	hm%	sm%	lam%	hm%	sm%	lam%	hm%	sm%	lam%	hm%	sm%	lam%	hm%	sm%	lam%
ijsSPAM2	0.38	1.24	0.69	0.23	0.95	0.47	1.52	0.34	0.72	0.16	11.74	1.44	0.36	3.43	1.12
ijsSPAM1	0.39	1.22	0.69	0.25	0.93	0.48	1.54	0.39	0.77	0.35	11.35	2.09	0.36	3.31	1.10
ijsSPAM4	0.46	1.28	0.77	0.37	0.91	0.58	1.44	0.54	0.88	0.30	12.77	2.07	0.43	3.34	1.21
ijsSPAM3	0.64	1.32	0.92	0.26	0.97	0.51	1.33	0.33	0.66	0.59	11.10	2.66	0.70	3.87	1.66
crmSPAM2	0.35	1.08	0.62	0.62	0.87	0.73	1.50	0.24	0.60	0.30	13.55	2.14	0.21	2.91	0.79
crmSPAM3	0.73	1.40	1.01	2.56	0.15	0.63	0.58	1.66	0.98	0.34	7.61	1.64	0.28	3.99	1.08
crmSPAM4	0.37	1.05	0.62	0.91	0.25	0.47	0.67	0.91	0.79	0.39	6.97	1.67	0.21	3.26	0.83
lbSPAM2	0.38	2.75	1.03	0.51	0.93	0.69	1.63	0.23	0.62	0.03	33.16	1.25	0.29	11.63	1.90
lbSPAM1	0.34	2.46	0.91	0.41	0.90	0.61	1.14	0.28	0.57	0.05	36.13	1.62	0.28	9.80	1.72
tamSPAM1	0.25	4.43	1.07	0.26	4.10	1.05	0.28	2.55	0.84	0.14	27.48	2.29	0.25	8.25	1.49
spamprobe	0.14	3.35	0.70	0.15	2.11	0.57	0.41	0.85	0.59	0.05	42.06	1.84	0.13	10.31	1.21
tamSPAM2	0.45	2.63	1.10	0.85	1.45	1.11	1.43	1.75	1.58	1.04	9.29	3.18	0.27	7.36	1.44
bogofilter	0.09	10.86	1.02	0.01	10.47	0.30	0.08	6.51	0.73	0.00	73.03	1.15	0.11	18.41	1.57
spamasas-b	0.31	2.17	0.83	0.25	1.29	0.57	0.49	1.00	0.70	0.06	25.68	1.47	0.33	6.00	1.43
lbSPAM3	0.68	2.81	1.39	0.83	1.05	0.94	6.84	0.35	1.57	0.47	42.45	5.55	0.29	11.04	1.85
crmSPAM1	0.80	3.75	1.74	1.84	1.65	1.74	4.22	0.50	1.46	0.37	13.42	2.34	0.33	15.72	2.44
lbSPAM4	0.59	4.66	1.67	0.91	3.87	1.89	4.86	1.21	2.44	0.26	49.94	4.82	0.26	12.06	1.87
yorSPAM2	0.38	3.91	1.23	0.92	1.74	1.27	0.34	1.03	0.60	0.14	23.64	2.07	0.25	14.90	2.05
spamasas-x	0.13	5.39	0.85	0.15	3.16	0.70	0.14	2.28	0.58	0.00	14.84	0.29	0.13	17.43	1.61
kidSPAM1	0.93	8.60	2.88	0.91	9.40	2.99	4.02	9.10	6.08	3.37	13.57	6.89	0.65	5.24	1.86
dspam-toe	0.58	1.88	1.05	1.04	0.99	1.01	1.94	0.59	1.07	0.05	30.97	1.45	0.40	5.78	1.55
621SPAM1	2.20	1.23	1.65	2.38	0.20	0.69	2.31	2.77	2.53	1.57	5.29	2.90	2.17	0.74	1.27
621SPAM3	0.70	12.58	3.08	3.14	0.17	0.73	1.73	2.87	2.23	0.56	7.48	2.09	0.00	66.27	0.95
yorSPAM4	1.29	2.98	1.96	2.99	1.36	2.02	5.20	0.45	1.55	0.77	91.35	22.26	0.63	9.04	2.45
dspam-tum	0.31	2.57	0.89	0.26	1.79	0.69	1.81	0.57	1.02	0.05	35.23	1.59	0.24	7.48	1.37
dspam-teft	0.26	2.93	0.87	0.26	1.79	0.69	1.85	0.53	0.99	0.00	44.26	0.63	0.17	9.32	1.31
yorSPAM3	1.16	2.29	1.63	1.29	1.20	1.25	4.41	0.65	1.71	1.36	15.32	4.76	0.92	8.11	2.78
dalSPAM3	5.44	8.65	6.87	6.80	6.23	6.51	3.44	9.79	5.86	4.03	13.29	7.43	5.27	12.63	8.23
yorSPAM1	1.32	2.85	1.94	2.44	2.43	2.44	4.96	0.55	1.67	1.19	13.81	4.20	0.82	8.28	2.65
dalSPAM1	0.92	18.93	4.44	1.17	21.07	5.33	1.17	13.83	4.18	1.35	38.19	8.42	0.82	22.82	4.70
dalSPAM2	5.40	9.64	7.24	5.34	7.52	6.34	3.12	11.33	6.03	4.73	12.39	7.73	5.58	11.83	8.17
kidSPAM4	2.94	5.05	3.86	9.74	6.57	8.01	5.31	2.39	3.57	5.75	18.09	10.40	0.91	5.88	2.34
kidSPAM3	0.75	11.11	2.99	0.82	12.49	3.33	3.03	10.27	5.64	2.86	24.42	8.89	0.51	8.58	2.15
kidSPAM2	0.84	9.71	2.92	0.87	10.53	3.11	2.71	9.89	5.24	3.40	16.15	7.62	0.61	6.85	2.08
ICTSPAM2	4.31	9.80	6.54	8.33	8.03	8.18	4.51	3.42	3.93	1.08	15.74	4.31	3.38	27.41	10.31
dalSPAM4	2.92	11.66	5.93	2.69	4.50	3.49	2.18	14.40	5.77	1.89	40.90	10.36	3.07	24.23	9.14
indSPAM3	2.49	8.74	4.71	1.09	7.66	2.93	1.81	4.62	2.90	1.83	37.03	9.48	2.92	18.98	7.74
pucSPAM0	2.21	8.06	4.27	3.41	5.10	4.18	4.93	2.26	3.35	1.44	24.13	6.39	1.77	27.34	7.61
indSPAM1	2.54	13.57	6.01	0.82	15.16	3.70	1.47	5.83	2.95	1.83	41.03	10.22	3.08	24.05	9.11
pucSPAM1	2.30	8.38	4.44	3.57	5.33	4.36	5.97	2.72	4.05	1.03	17.29	4.45	1.80	27.87	7.77
621SPAM2	14.59	4.50	8.23	55.06	1.07	10.32	25.82	2.87	9.21	2.47	6.84	4.13	3.83	17.02	8.29
pucSPAM2	2.58	7.17	4.33	3.35	5.00	4.10	6.07	2.77	4.12	2.47	40.90	11.70	2.17	20.73	7.08
ICTSPAM1	23.16	15.20	18.86	5.69	20.85	11.19	4.47	2.37	3.26	1.01	18.06	4.53	29.76	26.12	27.91
ICTSPAM3	13.11	27.33	19.24	14.10	28.22	20.26	9.57	3.91	6.16	6.61	19.35	11.53	13.33	73.29	39.38
ICTSPAM4	62.14	16.02	35.88	8.18	24.89	14.66	6.68	4.10	5.24	6.31	14.97	9.82	81.88	16.54	48.62
azeSPAM1	30.78	4.21	12.26	64.84	4.57	22.92	47.81	2.28	12.76	57.90	9.16	27.14	19.73	6.97	11.95
spamasas-v	-	-	-	0.06	39.51	1.87	0.02	11.70	0.54	0.02	72.13	2.00	-	-	-
popfile	-	-	-	0.92	1.26	0.94	0.96	0.49	0.69	0.14	22.97	2.03	-	-	-
tamSPAM4	-	-	-	-	-	-	0.96	0.89	0.92	3.92	6.19	4.93	-	-	-
tamSPAM3	-	-	-	0.22	4.46	1.01	0.82	1.85	1.23	6.29	3.64	4.79	-	-	-
indSPAM4	-	-	-	-	-	-	1.28	7.49	3.14	0.93	35.23	6.67	0.34	16.74	2.54
indSPAM2	-	-	-	-	-	-	2.66	3.09	2.86	0.03	100.00	99.99	2.87	21.41	8.24
azeSPAM2	-	-	-	-	-	-	8.54	25.35	15.12	8.04	59.48	26.38	0.63	36.84	5.75

Table 5: Misclassification Summary

	Aggregate			trec05p-1/full			Mr. X			S. B.			T. M.		
Filters	ROCA	h=.1	lam%	ROCA	h=.1	lam%	ROCA	h=.1	lam%	ROCA	h=.1	lam%	ROCA	h=.1	lam%
ijsSPAM2	0.051	3.78	0.69	0.019	1.78	0.47	0.069	9.72	0.72	0.285	12.13	1.44	0.135	10.31	1.12
ijsSPAM1	0.054	3.73	0.69	0.021	1.84	0.48	0.069	13.63	0.77	0.372	15.87	2.09	0.155	9.88	1.10
ijsSPAM4	0.058	4.91	0.77	0.025	2.22	0.58	0.063	8.68	0.88	0.422	17.03	2.07	0.167	12.66	1.21
ijsSPAM3	0.064	5.54	0.92	0.022	1.84	0.51	0.050	3.56	0.66	0.475	21.29	2.66	0.181	14.49	1.66
crmSPAM2	0.115	3.46	0.62	0.122	4.52	0.73	0.051	9.65	0.60	1.888	27.48	2.14	0.166	5.64	0.79
crmSPAM3	0.116	10.50	1.01	0.042	2.63	0.63	0.177	40.82	0.98	0.231	11.23	1.64	0.195	13.06	1.08
crmSPAM4	0.128	5.90	0.62	0.049	1.96	0.47	0.218	82.36	0.79	0.393	15.23	1.67	0.272	8.59	0.83
lbSPAM2	0.132	6.75	1.03	0.037	5.19	0.69	0.083	10.24	0.62	0.835	28.52	1.25	0.411	20.53	1.90
lbSPAM1	0.136	6.19	0.91	0.039	4.56	0.61	0.103	20.67	0.57	0.778	31.61	1.62	0.443	17.94	1.72
tamSPAM1	0.172	9.10	1.07	0.164	6.92	1.05	0.138	6.51	0.84	1.892	40.52	2.29	0.294	17.40	1.49
spamprobe	0.173	4.71	0.70	0.059	2.77	0.57	0.097	15.54	0.59	2.039	28.77	1.84	0.445	12.02	1.21
tamSPAM2	0.209	15.50	1.10	0.178	27.38	1.11	0.349	78.36	1.58	1.127	66.06	3.18	0.416	19.41	1.44
bogofilter	0.210	9.86	1.02	0.048	3.41	0.30	0.045	3.90	0.73	1.426	30.97	1.15	0.792	19.86	1.57
spamasas-b	0.220	6.72	0.83	0.059	2.56	0.57	0.097	6.19	0.70	1.620	19.87	1.47	0.736	15.58	1.43
lbSPAM3	0.262	29.95	1.39	0.122	22.38	0.94	0.875	95.73	1.57	2.727	98.32	5.55	0.456	22.38	1.85
crmSPAM1	0.263	12.79	1.74	0.169	10.53	1.74	0.311	81.61	1.46	2.393	23.48	2.34	0.790	23.12	2.44
lbSPAM4	0.302	17.23	1.67	0.238	22.94	1.89	0.492	58.36	2.44	1.988	52.65	4.82	0.588	19.67	1.87
yorSPAM2	0.316	21.14	1.23	0.457	34.21	1.27	0.051	6.08	0.60	0.983	30.52	2.07	0.619	39.19	2.05
spamasas-x	0.380	11.17	0.85	0.345	16.59	0.70	0.065	2.50	0.58	0.558	10.84	0.29	1.123	29.50	1.61
kidSPAM1	0.768	66.13	2.88	1.463	34.93	2.99	1.274	83.55	6.08	3.553	99.22	6.89	0.530	62.56	1.86
dspam-toe	0.987	83.68	1.05	0.773	88.76	1.01	1.109	96.23	1.07	14.149	31.61	1.45	2.626	77.16	1.55
621SPAM1	1.008	4.36	1.65	0.044	3.63	0.69	2.616	5.71	2.53	2.389	15.48	2.90	0.161	5.42	1.27
621SPAM3	1.090	7.89	3.08	0.060	7.02	0.73	2.692	4.55	2.23	2.604	17.16	2.09	0.332	6.15	0.95
yorSPAM4	1.122	81.80	1.96	0.688	84.92	2.02	1.407	96.18	1.55	58.165	98.06	22.26	1.081	78.66	2.45
dspam-tum	1.274	51.43	0.89	0.827	47.09	0.69	0.997	95.18	1.02	19.384	40.77	1.59	3.700	37.22	1.37
dspam-teft	1.383	51.60	0.87	0.827	47.09	0.69	0.942	95.17	0.99	21.428	43.35	0.63	4.263	33.79	1.31
yorSPAM3	1.491	70.88	1.63	0.861	62.13	1.25	1.993	92.07	1.71	8.234	70.13	4.76	4.366	78.42	2.78
dalSPAM3	1.873	59.50	6.87	1.491	41.00	6.51	1.613	70.03	5.86	2.845	77.16	7.43	3.090	59.70	8.23
yorSPAM1	1.917	84.38	1.94	2.032	87.24	2.44	2.632	95.76	1.67	7.237	77.16	4.20	4.400	78.76	2.65
dalSPAM1	2.097	99.15	4.44	2.348	99.75	5.33	2.240	99.31	4.18	4.614	100.00	8.42	3.085	52.08	4.70
dalSPAM2	2.100	60.64	7.24	1.674	41.92	6.34	1.824	69.41	6.03	3.293	83.48	7.73	2.898	59.84	8.17
kidSPAM4	2.606	89.15	3.86	3.990	93.74	8.01	2.326	98.23	3.57	8.042	95.22	10.40	2.473	85.34	2.34
kidSPAM3	2.741	88.23	2.99	4.167	90.62	3.33	2.822	97.67	5.64	6.360	93.67	8.89	2.653	82.11	2.15
kidSPAM2	3.003	88.29	2.92	4.544	91.65	3.11	2.738	97.64	5.24	7.020	97.29	7.62	2.749	85.16	2.08
ICTSPAM2	3.048	60.29	6.54	2.643	79.51	8.18	0.943	37.43	3.93	3.110	99.35	4.31	8.298	86.36	10.31
dalSPAM4	3.115	79.14	5.93	1.370	76.58	3.49	4.282	96.93	5.77	9.002	100.00	10.36	6.294	58.51	9.14
indSPAM3	3.168	96.99	4.71	2.822	97.35	2.93	2.321	99.31	2.90	12.454	91.10	9.48	5.843	99.41	7.74
pucSPAM0	4.030	59.56	4.27	2.083	59.71	4.18	1.910	51.00	3.35	1.408	61.81	6.39	2.925	88.94	7.61
indSPAM1	4.302	96.06	6.01	5.346	93.19	3.70	2.471	99.10	2.95	13.507	93.16	10.22	8.382	99.44	9.11
pucSPAM1	5.746	57.60	4.44	2.185	52.58	4.36	3.081	55.92	4.05	1.585	56.52	4.45	2.712	88.48	7.77
621SPAM2	6.064	54.21	8.23	11.362	28.85	10.32	6.814	59.16	9.21	3.169	61.94	4.13	2.647	47.89	8.29
pucSPAM2	6.107	99.98	4.33	1.967	51.28	4.10	3.454	78.25	4.12	5.437	73.42	11.70	3.688	99.99	7.08
ICTSPAM1	15.115	67.60	18.86	4.659	72.26	11.19	0.748	41.24	3.26	3.023	97.55	4.53	34.208	98.13	27.91
ICTSPAM3	17.637	99.17	19.24	20.485	99.39	20.26	5.328	98.50	6.16	9.985	98.71	11.53	36.233	99.75	39.38
ICTSPAM4	33.879	99.84	35.88	10.952	98.44	14.66	4.114	97.95	5.24	6.112	97.03	9.82	42.893	99.83	48.62
azeSPAM1	34.079	99.76	12.26	28.887	99.50	22.92	34.048	99.69	12.76	44.502	99.48	27.14	39.082	99.72	11.95
spamasas-v	-	-	-	0.516	31.31	1.87	0.091	4.97	0.54	5.736	68.26	2.00	-	-	-
popfile	-	-	-	0.325	7.35	0.94	0.326	86.94	0.69	2.199	24.65	2.03	-	-	-
tamSPAM4	-	-	-	-	-	-	0.159	46.24	0.92	1.421	89.68	4.93	-	-	-
tamSPAM3	-	-	-	0.183	7.64	1.01	0.257	58.80	1.23	1.934	96.49	4.79	-	-	-
indSPAM4	-	-	-	-	-	-	1.757	97.33	3.14	9.588	100.00	6.67	3.388	96.77	2.54
indSPAM2	-	-	-	-	-	-	2.804	99.75	2.86	68.572	99.87	99.99	13.462	99.44	8.24
azeSPAM2	-	-	-	-	-	-	29.765	99.95	15.12	37.739	100.00	26.38	22.625	99.89	5.75

Table 6: Summary Results

	Aggregate			trec05p-1/full			Mr. X			S. B.			T. M.		
Filters	ROCA	h=.1	lam%	ROCA	h=.1	lam%	ROCA	h=.1	lam%	ROCA	h=.1	lam%	ROCA	h=.1	lam%
ijsSPAM2	1	3	3	1	1	2	7	12	11	2	3	5	1	6	6
ijsSPAM1	2	2	3	2	2	4	7	14	13	3	6	17	2	5	5
ijsSPAM4	3	6	6	4	5	8	5	10	16	5	7	15	5	8	7
ijsSPAM3	4	7	12	3	2	5	2	2	8	6	10	22	6	10	18
crmSPAM2	5	1	1	14	11	16	3	11	5	17	13	19	4	2	1
crmSPAM3	6	15	13	7	7	10	16	18	18	1	2	10	7	9	4
crmSPAM4	7	8	1	10	4	2	17	31	14	4	4	11	8	4	2
lbSPAM2	8	11	15	5	13	11	9	13	7	9	14	4	11	17	23
lbSPAM1	9	9	11	6	12	9	13	16	2	8	18	9	13	13	19
tamSPAM1	10	13	17	16	14	22	14	9	15	18	20	20	9	12	14
spamprobe	11	5	5	11	8	6	11	15	4	21	15	12	14	7	7
tamSPAM2	12	18	18	18	22	23	21	29	26	11	27	24	12	14	13
bogofilter	13	14	14	9	9	1	1	3	12	14	17	3	21	16	16
spamasas-b	14	10	7	11	6	6	11	8	10	16	9	7	19	11	12
lbSPAM3	15	21	20	14	20	18	24	37	25	26	44	34	15	18	20
crmSPAM1	16	17	24	17	18	26	19	30	23	24	11	21	20	19	28
lbSPAM4	17	19	23	20	21	28	22	23	30	20	23	32	17	15	22
yorSPAM2	18	20	19	23	25	25	3	7	5	10	16	15	18	23	24
spamasas-x	19	16	8	22	19	15	6	1	3	7	1	1	23	20	17
kidSPAM1	20	30	27	31	26	32	29	32	49	32	46	37	16	29	21
dspam-toe	21	35	16	26	40	20	28	40	21	47	18	6	25	30	15
621SPAM1	22	4	22	8	10	11	40	6	31	23	5	23	3	1	9
621SPAM3	23	12	30	13	15	16	42	4	29	25	8	17	10	3	3
yorSPAM4	24	34	26	25	38	29	30	39	24	52	43	50	22	32	29
dspam-tum	25	22	10	27	29	11	27	36	20	48	21	8	36	22	11
dspam-teft	26	23	9	27	29	11	25	35	19	49	22	2	37	21	10
yorSPAM3	27	32	21	29	34	24	35	34	28	41	29	30	38	31	32
dalSPAM3	28	26	40	32	27	42	31	27	47	27	31	38	33	27	40
yorSPAM1	29	36	25	35	39	30	41	38	27	39	31	26	39	33	31
dalSPAM1	30	42	34	38	49	40	36	49	42	33	50	41	32	25	33
dalSPAM2	31	29	41	33	28	41	33	26	48	31	33	40	30	28	39
kidSPAM4	32	39	31	41	44	43	38	46	38	40	38	47	24	36	27
kidSPAM3	33	37	29	42	41	34	45	44	45	37	37	42	27	34	26
kidSPAM2	34	38	28	43	42	33	43	43	43	38	41	39	29	35	25
ICTSPAM2	35	28	39	39	37	44	26	17	39	29	47	27	42	37	45
dalSPAM4	36	33	37	30	36	35	49	41	46	42	50	46	41	26	44
indSPAM3	37	41	36	40	45	31	37	49	33	45	35	43	40	42	37
pucSPAM0	38	27	32	36	33	38	34	21	37	12	25	35	31	39	36
indSPAM1	39	40	38	45	43	36	39	48	34	46	36	45	43	43	43
pucSPAM1	40	25	34	37	32	39	46	22	40	15	24	28	28	38	38
621SPAM2	41	24	42	47	23	45	51	25	51	30	26	25	26	24	42
pucSPAM2	42	46	33	34	31	37	47	28	41	34	30	49	35	49	35
ICTSPAM1	43	31	44	44	35	46	23	19	36	28	42	29	46	41	47
ICTSPAM3	44	43	45	48	47	48	50	47	50	44	45	48	47	46	48
ICTSPAM4	45	45	46	46	46	47	48	45	43	36	40	44	49	47	49
azeSPAM1	46	44	43	49	48	49	53	51	52	51	48	52	48	45	46
spamasas-v	-	-	-	24	24	27	10	5	1	35	28	13	-	-	-
popfile	-	-	-	21	16	18	20	33	9	22	12	14	-	-	-
tamSPAM4	-	-	-	-	-	-	15	20	17	13	34	33	-	-	-
tamSPAM3	-	-	-	19	17	20	18	24	22	19	39	31	-	-	-
indSPAM4	-	-	-	-	-	-	32	42	35	43	50	36	34	40	30
indSPAM2	-	-	-	-	-	-	44	52	32	53	49	53	44	43	41
azeSPAM2	-	-	-	-	-	-	52	53	53	50	50	51	45	48	34

Table 7: Summary Result Rankings

	trec05p-1/full			trec05p-1/s25			trec05p-1/s50			trec05p-1/h25			trec05p-1/h50		
Filters	hm%	sm%	lam%	hm%	sm%	lam%	hm%	sm%	lam%	hm%	sm%	lam%	hm%	sm%	lam%
621SPAM1	2.38	0.20	0.69	3.45	0.42	1.22	2.51	0.27	0.83	3.94	0.17	0.83	2.78	0.19	0.72
621SPAM2	55.06	1.07	10.32	54.53	1.34	11.33	54.80	0.86	9.32	58.82	1.39	12.42	57.10	1.18	11.20
621SPAM3	3.14	0.17	0.73	5.28	0.17	0.98	4.32	0.16	0.84	3.33	0.16	0.74	2.84	0.16	0.68
ICTSPAM1	5.69	20.85	11.19	3.01	16.37	7.23	6.05	13.75	9.20	15.15	3.74	7.69	10.64	8.51	9.52
ICTSPAM2	8.33	8.03	8.18	6.91	17.98	11.31	5.57	15.47	9.42	11.32	14.29	12.73	7.73	15.66	11.09
ICTSPAM3	14.10	28.22	20.26	14.42	23.67	18.61	13.50	27.17	19.44	13.68	26.99	19.49	12.51	27.21	18.78
ICTSPAM4	8.18	24.89	14.66	1.60	64.28	14.61	1.60	64.24	14.60	19.51	9.65	13.86	8.31	18.46	12.53
azeSPAM1	64.84	4.57	22.92	-	-	-	-	-	-	-	-	-	-	-	-
crmSPAM1	1.84	1.65	1.74	0.22	6.76	1.26	0.68	3.79	1.61	5.98	0.59	1.91	3.47	1.00	1.87
crmSPAM2	0.62	0.87	0.73	0.28	2.67	0.87	0.27	49.18	4.84	2.11	0.38	0.89	0.97	0.53	0.71
crmSPAM3	2.56	0.15	0.63	2.41	0.33	0.89	2.48	0.23	0.76	4.12	0.16	0.82	3.17	0.15	0.70
crmSPAM4	0.91	0.25	0.47	0.61	0.72	0.66	0.73	0.40	0.54	3.56	0.09	0.57	1.96	0.13	0.51
dalSPAM1	1.17	21.07	5.33	1.17	22.66	5.57	1.09	21.26	5.18	2.27	20.67	7.21	1.54	17.57	5.46
dalSPAM2	5.34	7.52	6.34	5.69	8.97	7.16	5.65	7.88	6.68	5.88	7.11	6.47	5.34	7.31	6.25
dalSPAM3	6.80	6.23	6.51	6.96	7.58	7.27	6.94	6.48	6.71	7.11	5.88	6.47	7.02	6.00	6.49
dalSPAM4	2.69	4.50	3.49	2.47	6.19	3.93	2.28	4.88	3.35	4.66	5.44	5.03	3.58	3.42	3.50
ijsSPAM1	0.25	0.93	0.48	-	-	-	-	-	-	0.32	1.02	0.57	-	-	-
ijsSPAM2	0.23	0.95	0.47	-	-	-	-	-	-	0.30	1.04	0.56	-	-	-
ijsSPAM3	0.26	0.97	0.51	-	-	-	-	-	-	0.38	1.11	0.65	-	-	-
ijsSPAM4	0.37	0.91	0.58	-	-	-	-	-	-	0.45	1.05	0.69	-	-	-
indSPAM1	0.82	15.16	3.70	0.70	21.48	4.21	0.75	17.58	3.86	1.75	11.02	4.49	1.20	13.11	4.10
indSPAM3	1.09	7.66	2.93	0.89	9.32	2.95	1.18	7.02	2.92	2.27	5.56	3.56	1.70	6.95	3.46
kidSPAM1	0.91	9.40	2.99	1.99	6.74	3.69	1.44	8.01	3.45	0.40	13.24	2.42	0.36	12.01	2.16
kidSPAM2	0.87	10.53	3.11	-	-	-	-	-	-	-	-	-	-	-	-
kidSPAM3	0.82	12.49	3.33	-	-	-	-	-	-	-	-	-	-	-	-
kidSPAM4	9.74	6.57	8.01	-	-	-	-	-	-	-	-	-	-	-	-
lbSPAM1	0.41	0.90	0.61	0.16	4.33	0.84	0.28	1.95	0.74	1.68	0.31	0.73	0.85	0.58	0.71
lbSPAM2	0.51	0.93	0.69	-	-	-	-	-	-	-	-	-	-	-	-
lbSPAM3	0.83	1.05	0.94	-	-	-	-	-	-	-	-	-	-	-	-
lbSPAM4	0.91	3.87	1.89	0.58	11.69	2.71	0.71	6.94	2.26	2.96	1.46	2.08	1.44	2.57	1.93
pucSPAM0	3.41	5.10	4.18	1.62	9.70	4.04	2.28	6.86	3.98	9.62	3.57	5.91	5.82	4.32	5.02
pucSPAM1	3.57	5.33	4.36	1.71	10.25	4.27	2.44	7.31	4.25	10.07	3.74	6.19	6.06	4.50	5.22
pucSPAM2	3.35	5.00	4.10	1.50	8.97	3.73	2.15	6.47	3.76	10.51	3.92	6.47	6.00	4.46	5.18
tamSPAM1	0.26	4.10	1.05	0.22	9.05	1.45	0.07	13.94	1.08	0.47	4.55	1.48	0.37	3.15	1.08
tamSPAM2	0.85	1.45	1.11	0.73	3.03	1.49	0.72	2.39	1.31	1.97	1.56	1.75	1.42	1.51	1.46
tamSPAM3	0.22	4.46	1.01	0.34	69.17	8.05	-	-	-	-	-	-	-	-	-
yorSPAM1	2.44	2.43	2.44	1.00	6.36	2.56	1.62	3.89	2.51	7.22	1.08	2.84	4.55	1.70	2.79
yorSPAM2	0.92	1.74	1.27	0.48	3.60	1.32	0.72	2.43	1.32	2.26	1.17	1.63	1.45	1.44	1.44
yorSPAM3	1.29	1.20	1.25	0.47	2.60	1.11	0.80	1.86	1.22	3.75	0.72	1.65	2.26	0.95	1.47
yorSPAM4	2.99	1.36	2.02	0.96	3.87	1.94	1.74	2.32	2.01	9.98	0.48	2.26	5.66	0.77	2.11

Table 8: Public Corpora Misclassification Summary

	trec05p-1/full			trec05p-1/s25			trec05p-1/s50			trec05p-1/h25			trec05p-1/h50		
Filters	ROCA	h=.1	lam%	ROCA	h=.1	lam%	ROCA	h=.1	lam%	ROCA	h=.1	lam%	ROCA	h=.1	lam%
621SPAM1	0.044	3.63	0.69	0.091	4.14	1.22	0.048	2.72	0.83	0.070	6.65	0.83	0.054	5.34	0.72
621SPAM2	11.362	28.85	10.32	12.291	29.72	11.33	11.352	27.36	9.32	12.626	26.83	12.42	12.221	27.75	11.20
621SPAM3	0.060	7.02	0.73	0.085	6.72	0.98	0.061	7.07	0.84	0.068	7.58	0.74	0.058	6.28	0.68
ICTSPAM1	4.659	72.26	11.19	3.036	88.03	7.23	3.325	77.75	9.20	4.012	77.86	7.69	3.611	77.58	9.52
ICTSPAM2	2.643	79.51	8.18	4.571	89.15	11.31	2.741	85.34	9.42	6.140	95.18	12.73	3.777	83.79	11.09
ICTSPAM3	20.485	99.39	20.26	17.086	99.49	18.61	19.558	99.49	19.44	19.947	99.66	19.49	19.044	99.29	18.78
ICTSPAM4	10.952	98.44	14.66	27.891	97.00	14.61	27.506	96.29	14.60	10.821	99.58	13.86	8.995	98.67	12.53
azeSPAM1	28.887	99.50	22.92	-	-	-	-	-	-	-	-	-	-	-	-
crmSPAM1	0.169	10.53	1.74	0.236	9.64	1.26	0.194	10.58	1.61	0.383	43.87	1.91	0.219	18.37	1.87
crmSPAM2	0.122	4.52	0.73	0.343	5.23	0.87	41.915	50.14	4.84	0.097	22.25	0.89	0.067	7.59	0.71
crmSPAM3	0.042	2.63	0.63	0.051	2.96	0.89	0.044	2.64	0.76	0.066	6.42	0.82	0.051	2.11	0.70
crmSPAM4	0.049	1.96	0.47	0.089	1.90	0.66	0.055	1.36	0.54	0.069	11.63	0.57	0.059	3.26	0.51
dalSPAM1	2.348	99.75	5.33	2.662	99.73	5.57	2.183	99.76	5.18	2.997	99.47	7.21	2.026	99.50	5.46
dalSPAM2	1.674	41.92	6.34	1.970	56.39	7.16	1.827	49.81	6.68	1.713	41.31	6.47	1.694	40.60	6.25
dalSPAM3	1.491	41.00	6.51	1.814	51.35	7.27	1.635	47.24	6.71	1.453	40.80	6.47	1.459	38.51	6.49
dalSPAM4	1.370	76.58	3.49	1.854	85.10	3.93	1.430	82.06	3.35	2.087	82.63	5.03	1.217	71.00	3.50
ijsSPAM1	0.021	1.84	0.48	-	-	-	-	-	-	0.034	3.69	0.57	-	-	-
ijsSPAM2	0.019	1.78	0.47	-	-	-	-	-	-	0.031	3.15	0.56	-	-	-
ijsSPAM3	0.022	1.84	0.51	-	-	-	-	-	-	0.038	2.43	0.65	-	-	-
ijsSPAM4	0.025	2.22	0.58	-	-	-	-	-	-	0.041	4.03	0.69	-	-	-
indSPAM1	5.346	93.19	3.70	7.053	89.30	4.21	5.951	91.24	3.86	4.576	97.08	4.49	4.939	96.80	4.10
indSPAM3	2.822	97.35	2.93	2.844	97.56	2.95	2.471	98.18	2.92	3.210	98.53	3.56	3.012	97.59	3.46
kidSPAM1	1.463	34.93	2.99	1.589	55.96	3.69	1.546	46.36	3.45	1.812	26.90	2.42	1.586	27.99	2.16
kidSPAM2	4.544	91.65	3.11	-	-	-	-	-	-	-	-	-	-	-	-
kidSPAM3	4.167	90.62	3.33	-	-	-	-	-	-	-	-	-	-	-	-
kidSPAM4	3.990	93.74	8.01	-	-	-	-	-	-	-	-	-	-	-	-
lbSPAM1	0.039	4.56	0.61	0.092	5.71	0.84	0.054	4.75	0.74	0.081	14.26	0.73	0.056	10.10	0.71
lbSPAM2	0.037	5.19	0.69	-	-	-	-	-	-	-	-	-	-	-	-
lbSPAM3	0.122	22.38	0.94	-	-	-	-	-	-	-	-	-	-	-	-
lbSPAM4	0.238	22.94	1.89	0.588	20.53	2.71	0.347	21.61	2.26	0.332	46.84	2.08	0.261	33.31	1.93
pucSPAM0	2.083	59.71	4.18	2.200	65.46	4.04	2.083	61.97	3.98	2.600	59.58	5.91	2.314	56.50	5.02
pucSPAM1	2.185	52.58	4.36	2.623	48.07	4.27	2.367	51.10	4.25	2.618	49.60	6.19	2.409	55.12	5.22
pucSPAM2	1.967	51.28	4.10	1.788	54.12	3.73	1.853	52.57	3.76	3.274	52.25	6.47	2.358	54.32	5.18
tamSPAM1	0.164	6.92	1.05	0.483	12.33	1.45	1.004	11.97	1.08	0.234	10.71	1.48	0.123	6.10	1.08
tamSPAM2	0.178	27.38	1.11	0.268	15.61	1.49	0.225	16.81	1.31	0.326	60.91	1.75	0.323	54.26	1.46
tamSPAM3	0.183	7.64	1.01	22.663	71.24	8.05	-	-	-	-	-	-	-	-	-
yorSPAM1	2.032	87.24	2.44	3.234	92.91	2.56	2.369	89.66	2.51	3.292	91.80	2.84	2.564	89.86	2.79
yorSPAM2	0.457	34.21	1.27	0.420	24.46	1.32	0.426	29.56	1.32	0.669	38.53	1.63	0.530	35.53	1.44
yorSPAM3	0.861	62.13	1.25	1.176	56.54	1.11	1.025	64.74	1.22	1.382	72.84	1.65	1.082	70.10	1.47
yorSPAM4	0.688	84.92	2.02	0.586	78.32	1.94	0.537	84.19	2.01	1.975	90.82	2.26	1.117	89.26	2.11

Table 9: Public Corpora Summary Results

	Misclassified Spam (of 775 spams)							Misclassified Ham (of 6231 hams)							
	Automated	List	Newsletter	Phishing	Sex	Virus	**Total**	Automated	Commercial	Encrypted	Frequent	List	Newsletter	Personal	**Total**
621SPAM1	1	6	7	0	10	17	**41**	15	20	0	13	14	8	28	**98**
621SPAM2	1	9	7	3	15	18	**53**	20	15	18	29	15	9	48	**154**
621SPAM3	3	7	10	1	17	20	**58**	7	11	0	0	3	3	11	**35**
ICTSPAM1	11	21	14	5	83	6	**140**	6	6	0	6	27	9	9	**63**
ICTSPAM2	8	12	17	7	68	10	**122**	4	3	2	8	30	6	14	**67**
ICTSPAM3	5	17	11	1	114	2	**150**	14	29	45	56	154	64	50	**412**
ICTSPAM4	6	12	22	1	47	28	**116**	12	36	4	37	94	160	50	**393**
azeSPAM1	0	16	6	6	43	0	**71**	70	51	126	808	1938	255	360	**3608**
crmSPAM1	5	14	18	3	60	4	**104**	5	6	0	0	6	4	2	**23**
crmSPAM2	4	9	10	3	67	12	**105**	6	7	0	1	3	1	1	**19**
crmSPAM3	2	7	10	1	37	2	**59**	4	6	0	1	5	2	3	**21**
crmSPAM4	2	6	10	0	35	1	**54**	3	6	0	0	8	2	5	**24**
dalSPAM1	11	13	14	9	211	38	**296**	3	12	0	22	33	8	6	**84**
dalSPAM2	2	6	10	2	72	4	**96**	5	22	1	59	82	78	48	**295**
dalSPAM3	2	5	11	2	78	5	**103**	2	22	1	52	67	76	31	**251**
dalSPAM4	11	23	8	8	249	18	**317**	4	11	0	22	53	10	18	**118**
ijsSPAM1	3	9	4	1	66	5	**88**	6	6	0	1	6	2	1	**22**
ijsSPAM2	3	10	4	3	69	2	**91**	4	3	0	0	2	1	0	**10**
ijsSPAM3	2	7	3	0	69	5	**86**	9	10	0	1	12	3	2	**37**
ijsSPAM4	3	10	3	1	75	7	**99**	5	5	0	1	5	2	1	**19**
indSPAM1	5	18	19	6	251	19	**318**	4	5	0	10	34	55	6	**114**
indSPAM3	3	22	17	7	220	18	**287**	3	7	0	11	27	60	6	**114**
kidSPAM1	3	8	12	4	74	4	**105**	5	14	1	121	20	2	47	**210**
kidSPAM2	3	10	12	7	88	5	**125**	6	12	1	126	22	2	43	**212**
kidSPAM3	5	10	23	7	133	11	**189**	5	10	1	110	14	0	38	**178**
kidSPAM4	3	7	15	7	98	10	**140**	6	15	131	96	61	4	45	**358**
lbSPAM1	3	45	10	5	203	14	**280**	1	0	0	0	2	0	0	**3**
lbSPAM2	3	47	12	6	178	11	**257**	1	0	0	0	1	0	0	**2**
lbSPAM3	3	43	13	6	240	24	**329**	3	1	0	2	17	2	4	**29**
lbSPAM4	3	56	16	9	290	13	**387**	1	0	0	10	3	0	2	**16**
pucSPAM0	6	23	26	2	125	5	**187**	4	6	2	46	14	1	17	**90**
pucSPAM1	5	13	30	6	72	8	**134**	3	5	0	35	16	0	5	**64**
pucSPAM2	5	28	15	2	264	3	**317**	4	3	9	100	15	2	21	**154**
tamSPAM1	3	40	14	3	147	6	**213**	4	1	0	0	3	0	1	**9**
tamSPAM2	2	8	10	2	48	2	**72**	10	3	0	11	24	8	9	**65**
tamSPAM3	1	4	5	1	17	0	**28**	33	20	2	86	113	84	53	**392**
yorSPAM1	2	7	23	4	67	4	**107**	8	8	0	12	17	14	14	**74**
yorSPAM2	9	11	26	3	114	19	**182**	1	3	0	0	2	3	0	**9**
yorSPAM3	4	8	19	3	73	11	**118**	10	8	0	13	20	20	14	**85**
yorSPAM4	12	102	34	32	514	14	**708**	1	5	0	7	19	7	9	**48**

Table 10: Genre Classification of Misclassifications on S. B. Corpus

	Misclassified Spam (of 775 spams)							Misclassified Ham (of 6231 hams)							
	Automated	List	Newsletter	Phishing	Sex	Virus	**Total**	Automated	Commercial	Encrypted	Frequent	List	Newsletter	Personal	**Total**
ijsSPAM2	3	10	4	3	69	2	**91**	4	3	0	0	2	1	0	**10**
lbSPAM2	3	47	12	6	178	11	**257**	1	0	0	0	1	0	0	**2**
crmSPAM3	2	7	10	1	37	2	**59**	4	6	0	1	5	2	3	**21**
621SPAM1	1	6	7	0	10	17	**41**	15	20	0	13	14	8	28	**98**
tamSPAM1	3	40	14	3	147	6	**213**	4	1	0	0	3	0	1	**9**
yorSPAM2	9	11	26	3	114	19	**182**	1	3	0	0	2	3	0	**9**
dalSPAM4	11	23	8	8	249	18	**317**	4	11	0	22	53	10	18	**118**
kidSPAM1	3	8	12	4	74	4	**105**	5	14	1	121	20	2	47	**210**
pucSPAM2	5	28	15	2	264	3	**317**	4	3	9	100	15	2	21	**154**
ICTSPAM2	8	12	17	7	68	10	**122**	4	3	2	8	30	6	14	**67**
indSPAM3	3	22	17	7	220	18	**287**	3	7	0	11	27	60	6	**114**
azeSPAM1	0	16	6	6	43	0	**71**	70	51	126	808	1938	255	360	**3608**

Table 11: Genre Classification of Misclassifications on S. B. Corpus