Spam Filter Evaluation with Imprecise Ground Truth

Gordon V. Cormack
Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1, Canada

Aleksander Kołcz
Microsoft Live Labs
One Microsoft Way
Redmond, WA, USA

ABSTRACT

When trained and evaluated on accurately labeled datasets, online email spam filters are remarkably effective, achieving error rates an order of magnitude better than classifiers in similar applications. But labels acquired from user feedback or third-party adjudication exhibit higher error rates than the best filters – even filters trained using the same source of labels. It is appropriate to use naturally occuring labels – including errors – as training data in evaluating spam filters. Erroneous labels are problematic, however, when used as ground truth to measure filter effectiveness. Any measurement of the filter's error rate will be augmented and perhaps masked by the label error rate. Using two natural sources of labels, we demonstrate automatic and semi-automatic methods that reduce the influence of labeling errors on evaluation, yielding substantially more precise measurements of true filter error rates.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]:information filtering

General Terms: Experimentation, Measurement

Keywords: spam, email, filtering, classification, noise

1. INTRODUCTION

When trained and evaluated on accurately labeled datasets, online email spam filters achieve remarkably good performance. Gradient descent logistic regression [14], for example, yields an overall error rate of less than 0.5% when evaluated on the trec05p-1 corpus [10] using the TREC[1] Spam Track methodology [11]. This result betters by an order of magnitude those reported for similar applications of classifiers (cf. [28]). But is it good enough? Can it be improved? Can it be demonstrated beyond the laboratory? Imprecise knowledge of ground truth – the correct label for each filtered message – presents a substantial impediment to addressing these questions.

[1]http://trec.nist.gov

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.

The TREC methodology assumes a chronological sequence of accurately labeled email messages. The labels serve two distinct purposes in evaluation: as training examples for the filter, and as ground truth against which to measure filter error. Label errors therefore compromise both the filter's learning and the accuracy of the measured error.

The TREC datasets were carefully labeled by the organizers, adhering to a prescribed definition of spam:

> *Unsolicited, unwanted email that was sent indiscriminately, directly or indirectly, by a sender having no current relationship with the recipient.*

While there is no doubt that the TREC corpora contain some labeling errors, it is reasonable to assume that the error rate is not much worse than that of the best reported filter results using the labels; i.e. about 0.5%. Labels acquired in real-world deployment may be expected to exhibit much higher error rates. When explicitly asked to classify messages, human subjects have been reported to exhibit error rates of 3%-7% [30, 16]. Tacitly derived labels, such as those obtained from a "report spam" button, where it is assumed that unreported messages are ham, may have even higher rates.

It would be more appropriate to use naturally occuring labels – including errors – as training data, in conjunction with more accurate labels for measurement. Were the natural labels used for measurement, the estimated filter error rate would be augmented and perhaps masked by the label error rate, especially when the filter's true error rate is substantially lower than that of the labels. A filter with a true error rate of 0.5%, for example, might be estimated to have an error rate between 5.5% and 6.5% when evaluated using labels with an error rate of 6.0%, depending on the correlation between filter and label errors. A filter with a true error rate of 1.0% might show an error rate between 5.0% and 7.0%. These measurements would be insufficient to distinguish the two filters, and might even invert their relative performance.

An ideal filter evaluation would use a realistic sequence of messages, with a set of natural labels for training and a set of gold standard labels for evaluation. Accurately labeled sequences of messages are rare and somewhat unrealistic due to the costs and logistic challenges of acquiring and labeling them. The only public dataset for which both accurate and natural labels are available is *trec05p-1*, the TREC 2005 Public Spam corpus. The messages were acquired from Enron and released to the public in the course of a criminal investigation; accurate labels were created by the TREC coordinators [10]; natural labels were collected

later from Web users through the SpamOrHam effort [16]. Although the typicality of the messages and labels may be questioned, the messages are real and the two labeling efforts are clearly independent. The natural label error rate, at 6.7% [25], overwhelms that of the accurate label error rate.

The cost and difficulty of acquiring both natural and accurate labels for a realistic email sequence may be prohibitive. The SpamOrHam labels were acquired with much effort on the part of Web users, and predicated on the ability to publish the messages. It is unlikely that such an effort could be marshalled again, and, in any event, privacy considerations prevent almost any realistic email collection from being published. The gold standard labels were also acquired with much effort; privacy considerations impede similar efforts, as assessors are typically permitted to view messages only under restrictive and potentially unwieldy conditions.

2. RELATED WORK

Spam filtering as a research area is relatively new but nevertheless quite diverse. The TREC Spam Track is the largest and most realistic laboratory evaluation to date. In the three years it ran, ten test corpora with a total of 721,461 messages were used to test filters submitted by thirty-five participants. The results of more than one thousand experimental runs may be found in the summaries and appendices of the TREC proceedings [11, 6, 7].

The core task at TREC – the *immediate feedback* task – simulates the on-line deployment of a spam filter with idealized user feedback. The task, along with methods and evaluation measures embodied in an open-source toolkit [19], was developed by Cormack and Lynam [12].

This on-line evaluation strategy differs from traditional batch evaluation in that the messages are ordered and not explicitly partitioned into training and test examples. Every example is potentially available for training once the filter has rendered a verdict. The difference in methodology and results for state-of-the art filters has been studied by Cormack and Bratko [9]. The best results in these tests have been achieved by on-line support vector machines [26], logistic regression [14], and compression models [3], along with a number of open-source "Bayesian" filters modeled after the work of Graham and Robinson [15, 24], notably Bogofilter [22] and OSBF-Lua [2]. The best-performing approach demonstrated using TREC methods and datasets is the fusion all filters submitted for evaluation at TREC [20].

Variants of the task, including *delayed feedback, partial feedback,* and *active learning* explore more realistic models of user feedback; in particular, the impact of tardy, incomplete or restricted presentation of training examples after classification. The CEAS Live Challenge [1] integrates the TREC methodology with a live data stream, so as to be able to compare laboratory and in vivo filter deployment. Two new laboratory corpora were created; one captured in real time, and the other sampled from email delivered to clients of a large email service provider.

Sculley and Cormack [25] explore the effect of synthetic and natural label noise on filter training and conclude that both substantially compromise filter performance, natural noise more substantially so. Several approaches to mitigating the effect are found to improve filter performance, but not to anywhere near that achieved with noise-free training labels. A regularized version of an SVM classifier which performed well on synthetic noise yielded mediocre performance on natural noise. The effect of label noise on the measurement of filter error is not considered.

Labeling noise has been studied from the perspective of its impact on the accuracy machine learning algorithms, with the methodology of injecting artificial noise into a an initially noise-free dataset. Brodley and Friedl [4] proposed a data cleaning methodology whereby instances misclassified with high confidence by a learner (or ones for which there is a significant disagreement within an ensemble of learners) are removed from the training collection. They showed that models created over such cleaned data tend to be more accurate and demonstrated that ensemble based models are more resilient to class noise, at least when the level of noise is moderate. In addition to cleaning methods relying on inconsistent instance removal, alternatives based on label correction [25] and instance weighting [23] have also been considered.

While there exists a substantial body of published worked dedicated to different filtering and feature extraction techniques, our work is not so much to improve on these techniques as to improve the evaluation methodology; in particular, the ability to discriminate among and to measure the absolute performance of the best reported approaches for spam filtering, and for training label noise mitigation.

Lam amd Stork [18] discuss the problem of evaluating classifiers by means of test data with noisy labels. Given an estimate of the label noise and the assumption that label noise and filter error are uncorrelated, a precise estimate of filter error may be achieved, given enough examples. If, on the other hand, there may be correlation, wide bounds apply.

The issue of noisy and incomplete relevance judgements has been considered extensively within the context of the laboratory evaluation of information retrieval systems (cf. [5]). Although inter-adjudicator disagreement is large, system rankings are not particularly sensitive to it. Incompleteness is commonly resolved by the "pooling method," in which the top-ranked documents from each system are adjudicated as relevant or not, and all other documents are assumed to be irrelevant. More efficient methods select for adjudication more documents from top-performing systems, or documents most likely to discriminate among the filters under test; these methods are the subject of ongoing research interest. IR evaluation research is typically concerned only with ranking systems by relative performance on a particular corpus, as opposed to a calibrated effectiveness measure. IR evaluation further differs from spam filter evaluation in that relevant documents are extremely rare, and error rates (e.g. 1-recall, 1-precision) are quite large – of the order of 50% in a typical evaluation.

3. OBJECTIVES

The objective of this work is to determine the extent to which label error compromises filter performance measurement, and to validate a method of achieving better measurements without the intensive adjudication effort and access to messages normally associated with creating an accurate labeling. The automatic method requires no access whatsoever to the messages, while the semi-automatic method requires adjudication of a small fraction. The automatic method is applicable in situations where results are gathered by instrumenting an email system, but the content of mes-

		adj		
		ham	spam	all
nat	ham	28493	4058	32551
	spam	4055	43870	47925
	all	32548	47928	80476

Table 1: Agreement between nat and adj SpamOrHam label sets.

		trec		
		ham	spam	all
nat	ham	30612	1939	32551
	spam	3596	44329	47925
	all	34208	46268	80476

Table 2: Agreement between nat and trec label sets.

		trec		
		ham	spam	all
adj	ham	30622	1926	32548
	spam	3586	44342	47928
	all	34208	46268	80476

Table 3: Agreement between adj and trec label sets.

sages are not available to evaluators. The semi-automatic method is applicable in situations where it is possible to adjudicate some messages, provided the number is not too onerous. For example, the original recipient may be asked to review occasional messages, or to spend an hour or two participating in a review process [21].

We assume that it is possible to apply several filters to exactly the same sequence of messages with the same training labels and that each filter's result (either a categorical decision or a score or both) is available, along with a natural label, for each message in the sequence. The results of these filters are fused to form a set of labels which, we hypothesize, has a lower error rate and yields more accurate results than the natural labels when used as ground truth for evaluation. The hypothesis is tested in several ways:

- direct comparison between pseudo-gold and gold standard labels, if a gold standard is available;
- adjudication of differences between pseudo-gold and natural labels, using an independent adjudicator;
- comparison of performance measures (AUC and LAM);
- comparison of filter performance rankings;
- measuring the statistical power to discriminate between pairs of filters.

4. DATASETS

We used email messages from two separate corpora: the TREC 2005 Public Spam Corpus (trec05p-1) [11], and the CEAS 2008 Live Challenge private corpus [1]. Four independent sets of labels were used in total; three for the TREC messages and one for the CEAS messages.

The TREC corpus includes a gold standard label for each message. In addition, we acquired the labels collected by the SpamOrHam project [16]. The messages labeled by

Tag	Description
bayes	Naive Bayes, character 4-gram binary features, first 3000 bytes of message, including header. Per-message feature selection.
nobs	Naive Bayes, alphanumeric sequence "word" features, first 3000 words of message, including header. Per-message feature selection.,
bogo	Bogofilter version 1.1.5, default parameters [22].
dmc	Dynamic Markov Compression [3].
wat1	Gradient descent logistic regression, 4-gram character features [8].
osbf	OSBF-Lua [2].
tft1	Relaxed Online Support Vector Machine [27].

Table 4: Base filters used for evaluation.

Tag	Description
logbagf	**wat1**, modified to train only 10% of the features, selected at random for each example.
logbagm	ensemble of **wat1** filters, each modified to train only 10% of examples, selected at random.
lrslow	**wat1**, with the learning rate parameter reduced from 0.01 to 0.001
dmc	Dynamic Markov Compression [3].
tft1-0.5	**tft1**, with C parameter reduced from 100 to 0.1.

Table 5: Base filters altered to reduce sensitivity to label noise.

SpamOrHam were selected at random with replacement, resulting in a variable number of labels per message. From the messages having two or more labels, we selected one label at random to serve as the natural label, and a second (without replacement) to simulate the result of adjudication. Messages having fewer than two SpamOrHam labels were eliminated from the evaluation. The TREC labels were used as a reference standard, and also to simulate more reliable adjudication. The resulting dataset contains 80,476 messages, approximately 33,000 ham and 48,000 spam.

The CEAS corpus was collected from messages delivered to clients of a large service provider. The messages to be labeled were selected at random from those delivered to a set of volunteer clients; the labels are the responses to specific adjudication requests to the recipients. The corpus contains 198,574 messages, of which 89,451 are labeled ham, and 109,123 spam. There is exactly one label per message.

5. FILTER EVALUATION MEASURES

The TREC methodology requires that filters return both a hard result (ham or spam) and a soft result (a "spaminess" score) which may be compared after the fact to some threshold t. Hard results are evaluated as a pair of error rates (fpr, fnr) for ham and spam respectively. The labels fpr and fnr denote false positive and false negative rate from diagnostic test theory [13]. A classifier with lower fpr and fnr than another is superior. (Under the assumption that all messages have equal misclassification cost [17].) Whether a classifier with a lower fpr and higher fnr is superior or inferior depends on the user's sensitivity to each kind of error.

Tag	Description
bayesm	8 **bayes** ensemble, disjoint training examples.
nobsm	8 **nobs** ensemble, disjoint training examples.
bogom	8 **bogo** ensemble, disjoint training examples.
dmcm	8 **dmc** ensemble, disjoint training examples.
wat1m	8 **wat1** ensemble, disjoint training examples.
logbagfm	8 **logbagf** ensemble, disjoint training examples.
osbfm	8 osbf ensemble, disjoint training examples.
tft1m	8 **tft1** ensemble, disjoint training examples.

Table 6: Base filters in bagging configuration to reduce sensitivity to label noise.

A plethora of measures – including accuracy, weighted accuracy, total cost ratio, F-measure, and utility – attempt to quantify this sensitivity and to use this quantification to combine fpr and fnr, along with the corpus ham-to-spam ratio, into a one-dimensional measure.

The soft result may be characterized by the set of all distinguishable (fpr, fnr) pairs for different values of t. This set of points defines a receiver operating characteristic (ROC) curve [29]; a filter whose ROC curve is strictly above that of another is superior in all deployment situations, while a filter whose ROC curve crosses that of another is superior for some threshold settings and inferior for others.

The area under the ROC curve (AUC) provides an estimate of the effectiveness of a soft classifier over all threshold settings. AUC also has a probabilistic interpretation: it is the probability that the classifier will award a random spam message a higher score than a random ham message. In the spam filtering domain, typical AUC values are of the order of 0.999 or greater; following TREC, we report (1-AUC)% , the area above the ROC curve, as a percentage. So AUC=0.999 would be reported instead as (1-AUC)%=0.1.

While AUC provides an amenable score for ranking soft classifiers, the pair (fpr, fnr) does not serve this purpose for hard classifiers. It has been observed [13] that the diagnostic odds ratio, $dor = \frac{(1-fpr)\cdot(1-fnr)}{fpr\cdot fnr}$ is, for many diagnostic tests, effectively invariant over a large number of threshold settings. Intuitively, a change in threshold setting that increases the odds of misclassifying ham by some multiplicative factor tends to decrease the odds of misclassifying spam by the same factor. Therefore dor is a useful summary measure largely uninfluenced by threshold setting. The same effect has been observed at TREC [11], giving rise to the measure *logistic average misclassification* rate, $LAM = logit^{-1}(\frac{logit(fpr)+logit(fnr)}{2}) = logit^{-1}(log(dor^{-0.5}))$. Note that the value LAM is necessarily between fpr and fnr; when t is set to equalize error rates, we have $fpr = fnr = LAM$.

Under the assumption that dor is invariant, it is possible to estimate (fpr', fnr') from (fpr, fnr) by solving the equation

$$\frac{(1-fpr)\cdot(1-fnr)}{fpr\cdot fnr} \approx \frac{(1-fpr')\cdot(1-fnr')}{fpr'\cdot fnr'}.$$

It is further possible to estimate $AUC \approx \int_0^1 (fnr)\, d\,(fpr)$.

6. SPAM FILTERS

We chose to evaluate several base filters previously configured for the TREC Spam Filter Evaluation Toolkit. In addition, we included variants of these filters altered to mitigate training label error.Table 4 provides an identifying tag and a short description for each base filter. Table 5 describes altered versions of some of these filters. Table 6 identifies ensemble filters that we created by running a particular filter eight times, training each time on only one eighth of the examples, randomly partitioned. We were unable to complete some of the filter runs either because the filters failed or because they failed to complete in a reasonable amount of time. These runs were excluded from consideration; as a consequence 17 of the subject filters were used with the TREC data, and 11 on the CEAS data.

7. APPROXIMATIONS TO TRUTH

The overall objective of spam filter evaluation is to determine which spam filters better approximate truth, so that they may better serve their intended purpose. If the true class of each message is known, filter performance may be quantified by an amenable measure of the distance between the filter's result and truth. The TREC evaluations use receiver operating characteristic (ROC) area under the curve (expressed as $(1 - AUC)(\%)$, so that lower numbers are better) and logistic average misclassification rate (expressed as $(LAM)(\%)$) as threshold-insensitive measures of performance.

When a pair of filters exhibit similar or contradictory relative performance according to these measures, pairwise differential comparison may provide a more sensitive indication of which is closer to truth. In a differential comparison, only the cases of disagreement between filter results are compared to truth; a simple sign test determines the better approximation to truth. A tournament – in which each filter is differentially compared to each other filter – may be used to establish a ranking, but no quantitative measure of how close an approximation to truth is achieved by each filter.

Differential comparison, unlike the TREC measures, is very sensitive to the filters' threshold settings, and also to the prevalence of spam in the evaluation dataset. It therefore cannot reward, and is likely to penalize, a filter's ability – as may well be desirable – to identify ham with a lower error rate than spam. In the TREC setting, where filters report a confidence score in addition to a categorical classification, this shortcoming of differential comparison can be mitigated by threshold-adjusted differential comparison. Prior to comparison, each classifier's threshold is adjusted to achieve equal apparent ham and spam error rates (i.e. $fpr = fnr = LAM$). Threshold-adjusted differential comparison affords a consistent threshold-independent approximation, albeit one that fails to capture one aspect of filter performance.

Truth, at least with respect to spam filtering, is an abstraction. It may be approximated but never realized; the aptness of an approximation can only be estimated. Infor-

a	l_1	l_2	$l_1 = a$	$l_2 = a$	winner
adj	nat	trec	1467	4068	trec ($p \approx 0.00$)
nat	adj	trec	1467	4045	trec ($p \approx 0.00$)
trec	nat	adj	4045	4068	tie ($p \approx 0.8$)

Table 7: Adjudicated differential comparison between label sets l_1 and l_2 using a as adjudicator.

	adjudication		
rank	trec	nat	adj
1	tft1-05	tft1-05	tft1-05 ($p \approx 0.00$)
2	**dmcm**	**dmcm**	**logbagfm** ($p < 0.39$)
3	**logbagfm**	**logbagfm**	**lrslow** ($p < 0.87$)
4	**lrslow**	**logbagm**	**dmcm**($p < 0.96$)
5	**logbagm**	**lrslow**	**logbagm** ($p < 0.36$)
6	**wat1m**	**logbagf**	**logbagf** ($p < 0.67$)
7	**logbagf**	**wat1m**	**wat1m**($p < 0.03$)
8	**osbfm**	**osbfm**	**osbfm**($p < 0.66$)
9	**bogom**	**bogom**	**bogom**($p \approx 0.00$)
10	**dmc**	**dmc**	**dmc**($p \approx 0.34$)
11	**bogo**	**bogo**	**bogo**($p \approx 0.00$)
12	**nobs**	**nobs**	**wat1**($p \approx 0.00$)
13	**wat1**	**wat1**	**nobs**($p \approx 0.00$)
14	**bayes**	**bayes**	**bayes**($p \approx 0.00$)
15	**nobsm**	**nobsm**	**nobsm**($p \approx 0.00$)
16	**bayesm**	**bayesm**	**bayesm**($p \approx 0.00$)
17	**tft1**	**tft1**	**tft1**
power	0.82	0.90	0.88

Table 8: Adjudicated tournament ranking of subject filter performance, using trec, nat and adj labels as ground truth.

mally, through a combination of qualitative observations and statistical inference, we argue that the TREC labels better approximate truth than either of the two SpamOrHam label sets, which are different but equally good approximations. Pairwise agreement and disagreement between the three label sets (dubbed trec, nat, and adj) is quantified in tables 1 through 3. We see that nat and adj disagree on 10% of the messages, while each disagrees with trec on 6.9%. These agreement rates indicate that SpamOrHam judgements have 5.2% random error, and that the TREC judgements have considerably less. There may also be a systematic difference between the effective definition of spam applied by TREC and SpamOrHam assessors, or any number of other systematic differences. The net effect is that each approximation differs from truth by two factors: random error (noise) and systematic error (bias).

For the purposes of this evaluation we define the true class to be the majority opinion of the hypothetical infinite population of users from which the SpamOrHam judgements are drawn. That is, we deem the SpamOrHam labels to have no bias, and to differ from truth by random error alone. A spam filter's performance is therefore defined by how well it predicts the majority opinion, notwithstanding any quibbling about the definition of spam or the competence of members of the population to apply the definition.

From the perspective of this definition, the TREC labels exhibit some bias. Evidence of this bias is apparent from the prevalence of spam labels in each set: nat contains 59.6% (95% c.l.: $59.2 - 59.9$) spam, as does adj ($59.2 - 59.9$). On the other hand, trec contains 57.1% spam ($57.2 - 57.8$). nat and adj agree within the limits of chance (as we would expect, given that they are independent samples from the same population) while trec disagrees by 2.5%, a significant systematic error. While random error for the trec labels is difficult to quantify, we may infer from this bias estimate and the disagreement rate that trec label noise is smaller than, and positively correlated with, SpamOrHam label noise.

Adjudicated differential comparison may be used to compare pairs of labelings. But the "correct labels" are unknown, so instead we use a third independent labeling, or a live adjudicator, as a surrogate. Provided the third labeling is independent and yields the correct label more often than not, we may conclude that, of the labelings being compared, the one that agrees with the third more often better approximates truth. A sign test evaluates the overall significance of the comparison result. Adjudicated differential comparison may be used to compare and rank labelings, but offers no quantitative estimate of the error rates of the labelings being compared or, for that matter, of the adjudication labeling. Table 7 illustrates the result of adjudicated differential comparison among the trec labelings; demonstrating formally our observation that the trec labels are more accurate, while the SpamOrHam labelings are statistically indistinguishable.

8. RANKING FILTER PERFORMANCE

While the ultimate goal of this work is to accurately estimate filter performance using standard measures, we first consider the problem of ranking filters. A ranking is considered good to the extent that it orders filters consistently with the standard measures, without concern for quantitative estimates. Kendall's τ rank correlation is commonly used to compare rankings. If t is the true ranking of filters according to some measure, $\hat{t}_1$and $\hat{t}_2$ are approximations, $\hat{t}_1$is closer to the true ranking than $\hat{t}_2$ if $\tau(t, \hat{t}_1) < \tau(t, \hat{t}_2)$, but rank correlation gives no indication whether the difference between t and $\hat{t}_1$, or between t and $\hat{t}_2$, or between $\hat{t}_1$and $\hat{t}_2$ represent chance or significant differences. To this end, instead of $\tau(t, \hat{t})$ we report the proportion of *inversions* between t and $\hat{t}$, as well as the proportion of *significant inversions* and the *power* of the estimate $\hat{t}$. An inversion between filters f_1 and f_2 occurs if $f_1 > f_2$ in t while $f_1 < f_2$ in $\hat{t}$. An inversion is significant if a statistical test determines that $f_1 < f_2$ in $\hat{t}$ ($p < \alpha$), for some small α. The power of $\hat{t}$ is the fraction of pairs $f_1 < f_2$ such that $p < \alpha$. A good ranking would yield high power, and low significant inversions when compared to the true ranking. If the proportion of significant inversions is less than α, we cannot reject the null hypothesis that the difference is due to chance. Unless otherwise stated, we assume $\alpha = 0.05$.

Table 8 shows the rankings of 17 subject filters (described in table 5), ranked by tournament using threshold-adjusted differential comparison using each of the three labels for adjudication. Due to space limitations, we include statistical p-values for only one ranking, and then only for the differences between adjacent filters in the ranking (e.g., in the adj column, **osbfm**<**wat1m** ($p < 0.03$)). The power of the three rankings is given in the bottom row (e.g., the adj ranking has power 0.88 because 122 of the 136 pairings yield a significant difference). None of the inversions between any pair of rankings is significant ($p < 0.05$) according to either of the rankings.

9. QUANTIFYING FILTER PERFORMANCE

We first consider the problem of measuring the filters' threshold adjusted error rates that were used for ranking in the previous section. Later, we consider the TREC measures. Table 9 estimates threshold adjusted error rates, using adj and trec, respectively, as ground truth. We have not

	filter error (95% c.l.)	
filter	adj labels	trec labels
tft1-0.5	6.76 (6.61 - 6.91)	1.82 (1.71 - 1.93)
logbagfm	6.88 (6.70 - 7.07)	2.08 (1.99 - 2.19)
lrslow	6.90 (6.71 - 7.10)	2.13 (2.03 - 2.23)
dmcm	6.92 (6.72 - 7.12)	2.07 (1.98 - 2.18)
logbagm	6.92 (6.76 - 7.08)	2.10 (2.00 - 2.20)
logbagf	6.95 (6.77 - 7.13)	2.16 (2.05 - 2.27)
wat1m	6.96 (6.80 - 7.14)	2.18 (2.09 - 2.28)
osbfm	7.07 (6.91 - 7.24)	2.22 (2.11 - 2.33)
bogom	7.09 (6.91 - 7.28)	2.49 (2.39 - 2.60)
dmc	7.26 (7.08 - 7.44)	2.75 (2.63 - 2.87)
bogo	7.35 (7.18 - 7.52)	2.88 (2.76 - 3.01)
wat1	8.30 (8.12 - 8.48)	4.27 (4.14 - 4.39)
nobs	8.36 (8.17 - 8.54)	4.14 (3.99 - 4.30)
bayes	8.64 (8.44 - 8.83)	4.62 (4.48 - 4.76)
nobsm	8.79 (8.59 - 9.00)	4.70 (4.58 - 4.83)
bayesm	8.83 (8.64 - 9.02)	4.84 (4.69 - 5.00)
tft1	9.71 (9.50 - 9.94)	5.99 (5.79 - 6.19)

Table 9: Threshold adjusted error estimates using trec and adj labels as ground truth.

	LAM(%) (95% c.l.)	
filter	lmns labels	trec labels
tft1-0.5	0.85 (0.79 - 0.91)	1.82 (1.71 - 1.93)
dmcm	1.23 (1.15 - 1.31)	2.07 (1.98 - 2.18)
logbagfm	1.23 (1.16 - 1.30)	2.08 (1.99 - 2.19)
lrslow	1.26 (1.18 - 1.34)	2.13 (2.03 - 2.23)
wat1m	1.30 (1.23 - 1.38)	2.18 (2.09 - 2.28)
logbagm	1.31 (1.24 - 1.39)	2.10 (2.00 - 2.20)
logbagf	1.35 (1.26 - 1.43)	2.16 (2.05 - 2.27)
osbfm	1.50 (1.41 - 1.59)	2.22 (2.11 - 2.33)
bogom	1.65 (1.56 - 1.75)	2.49 (2.39 - 2.60)
dmc	1.79 (1.70 - 1.88)	2.75 (2.63 - 2.87)
bogo	2.01 (1.91 - 2.12)	2.88 (2.76 - 3.01)
nobs	3.36 (3.22 - 3.50)	4.14 (3.99 - 4.30)
bayes	3.60 (3.48 - 3.73)	4.62 (4.48 - 4.76)
wat1	3.61 (3.49 - 3.72)	4.27 (4.14 - 4.39)
nobsm	3.87 (3.75 - 4.00)	4.70 (4.58 - 4.83)
bayesm	3.88 (3.75 - 4.01)	4.84 (4.69 - 5.00)
tft1	5.37 (5.19 - 5.55)	5.99 (5.79 - 6.19)

Table 10: Logistic average misclassification estimates using lmns and trec labels as ground truth.

formally computed the power of the rankings resulting from these estimates; however, we may conclude from the overlapping confidence intervals of all but two adjacent filters, the power is, as expected, much lower than that achieved using differential comparison. That said, there are no significant inversions between the rankings. The error rates are, as predicted, substantially higher using the adj labels. For both labelings, the estimated filter error rates are comparable to known errors in the labels of 2.7% and 5.2%, respectively. The best we can determine from these results is that tft1-0.5, for example, likely has a true error rate between about 0.9% and 3.6%.

We can do better. A differential comparison between tft1-0.5 and the trec labels (adjudicated using adj) shows tft1-0.5 to be more accurate ($p < 0.003$). So the trec labeling serves better as an upper bound on the filter's error rate than as ground truth. And if tft1-0.5 better approximates truth, why not use its results as ground truth? One possible objection is that results may be biased in favor of similar filters, and against dissimilar ones. One way to reduce bias is use a committee of filters to create the labels. Lynam and Cormack [20] have shown that filter fusion can be expected to outperform any single filter, even when the performance of the filters varies by several orders of magnitude (although we know in this case from evaluation using existing labels, that none of the filters is terrible). The best reported fusion method first adjusts the scores to estimate log-odds, and then combines the scores with on-line logistic regression. The log-odds adjustment replaces the score s_i for the i^{th} message by

$$log\left(\frac{|\{j \mid s_j \leq s_i \, and \, label_j = spam\}|}{|\{j \mid s_j \geq s_i \, and \, label_j = ham\}|}\right).$$

The resulting scores, one per filter, comprise the input feature vector for adaptive logistic regression. We applied this technique, and also performed threshold adjustment on the result so that $fpr = fnr = LAM$. The categorical results of this effort were used to form the pseudo-gold standard lmns.

Adjudicated using adj, a differential comparison fails to show a significant difference between tft1-0.5 and lmns ($p < 0.8$). Adjudicated using trec, lmns is clearly superior ($p \approx 0.000$). The ranking yielded by lmns shows no significant inversions from those reported above. Table 11 reports threshold-adjusted error using lmns as ground truth, and also reports (1-AUC)(%). We note that the error reported for tft1-0.5 is at the low end of the range we guessed based on measurements using the trec and adj labels. The AUC measures appear reasonable – the noise-tolerant methods do much better than the others, but still not as well as the others when they are trained and evaluated on the trec labels. In fact, the AUC measures are remarkably similar to those reported by Sculley and Cormack [25], using synthetic noise.

For comparison, AUC measures using trec and adj ground truth are presented in table 12. While the larger magnitude of the estimates is to be expected, the difference in score and ranking of tft1-0.5 is remarkable. Using trec or nat labels, its performance appears mediocre, whereas by every other measure it is the best by a substantial margin. LAM scores, on the other hand, are uniformly higher with respect to the trec labels, and there are no significant inversions (see table 13).

The same method was used to prepare new ground-truth labels for the CEAS dataset, albeit with fewer filters. Figure 14 presents the AUC results, while table 15 presents the LAM results using as ground truth the original CEAS labels, as well as new labels constructed by fusion. The original values and the substantial improvement of all scores are consistent with high random error in the original labels – comparable to the level in the TREC dataset – which is abated in the new labels.

10. DISCUSSION AND CONCLUSION

A fully automatic method fuses the results of candidate filters to yield a pseudo-gold labeling that is used as ground truth in evaluating email spam filters. The pseudo-gold labels exhibit a lower error rate than labels obtained from natural sources including user labels and exhaustive adjudication by experts. Using the labeling as ground truth for

	error measure (95% c.l.)	
filter	thresh. adj.	(1-AUC)(%)
tft1-0.5	0.85 (0.79 - 0.91)	0.041 (0.034 - 0.051)
dmcm	1.23 (1.15 - 1.31)	0.074 (0.062 - 0.087)
logbagfm	1.23 (1.16 - 1.30)	0.051 (0.046 - 0.056)
lrslow	1.26 (1.18 - 1.34)	0.052 (0.047 - 0.057)
wat1m	1.30 (1.23 - 1.38)	0.056 (0.051 - 0.062)
logbagm	1.31 (1.24 - 1.39)	0.057 (0.052 - 0.063)
logbagf	1.35 (1.26 - 1.43)	0.057 (0.051 - 0.062)
osbfm	1.50 (1.41 - 1.59)	0.096 (0.086 - 0.107)
bogom	1.65 (1.56 - 1.75)	0.10 (0.09 - 0.11)
dmc	1.79 (1.70 - 1.88)	0.24 (0.22 - 0.26)
bogo	2.01 (1.91 - 2.12)	0.21 (0.19 - 0.23)
nobs	3.36 (3.22 - 3.50)	0.45 (0.42 - 0.48)
bayes	3.60 (3.48 - 3.73)	0.52 (0.50 - 0.56)
wat1	3.61 (3.49 - 3.72)	0.90 (0.85 - 0.96)
nobsm	3.87 (3.75 - 4.00)	0.69 (0.65 - 0.74)
bayesm	3.88 (3.75 - 4.01)	0.63 (0.59 - 0.67)
tft1	5.37 (5.19 - 5.55)	1.5 (1.4 - 1.6)

Table 11: Threshold adjusted error and ROC area estimates using pseudo-gold labels as ground truth.

	(1-AUC)(%) (95% c.l.)	
filter	trec labels	nat labels
logbagm	0.14 (0.13 - 0.15)	3.49 (3.37 - 3.63)
logbagfm	0.14 (0.13 - 0.15)	3.40 (3.25 - 3.55)
lrslow	0.14 (0.13 - 0.15)	3.33 (3.22 - 3.45)
wat1m	0.15 (0.14 - 0.16)	3.53 (3.38 - 3.67)
logbagf	0.15 (0.14 - 0.16)	3.35 (3.24 - 3.48)
dmcm	0.22 (0.20 - 0.24)	4.04 (3.92 - 4.17)
osbfm	0.24 (0.22 - 0.26)	3.90 (3.77 - 4.04)
bogom	0.30 (0.29 - 0.32)	3.99 (3.86 - 4.13)
tft1-0.5	0.30 (0.27 - 0.34)	4.02 (3.88 - 4.17)
bogo	0.53 (0.49 - 0.57)	4.51 (4.35 - 4.67)
dmc	0.59 (0.55 - 0.63)	4.33 (4.18 - 4.50)
nobs	0.74 (0.70 - 0.78)	4.26 (4.11 - 4.41)
bayes	0.84 (0.81 - 0.88)	4.45 (4.32 - 4.59)
bayesm	1.02 (0.97 - 1.08)	4.75 (4.59 - 4.93)
nobsm	1.11 (1.06 - 1.17)	4.69 (4.53 - 4.85)
wat1	1.25 (1.19 - 1.32)	4.83 (4.68 - 4.98)
tft1	2.06 (1.96 - 2.15)	5.56 (5.41 - 5.71)

Table 12: ROC area estimates using trec and nat labels as ground truth.

evaluation results in lower estimated filter error rates across the board, according to standard measures. This effect is consistent with the hypothesis that the these new estimates are more precise in absolute terms. There is some possibility that the pseudo-gold labels to some extent represent "group think" of the subject filters, which all rely exclusively on the text of the message for classification, albeit using different algorithms and feature engineering. This concern is similar to that raised with respect to the pooling method for IR evaluation. This concern applies mainly to the magnitude of reported error rates, as differential comparison is insensitive to it.

Differential comparison, which requires some adjudication, may be used to demonstrate that the pseudo-gold standard has a lower error rate than other available labels. And, were a radical new filter to discover errors in the labels, this situation could be discoverable by differential comparison. It is important to note that differential comparison does not require that the adjudicator be more accurate than the labels, just that the adjudicator be more accurate than chance. Differential comparison may also be used to rank filters directly; the rankings achieved by tournament ranking, even with noisy adjudication, are quite powerful, although they yield no quantitative estimate of filter error rates.

We call into question Sculley and Cormack's claim [25] that spam filters perform more poorly with natural than with random training label noise. Our results are consistent with the hypothesis that the filters perform about as well on both, and that the reported results for the best-performing filter – SVM with a low C parameter – were substantially confounded by errors in the evaluation labels. Other approaches designed to mitigate label noise, including slowed learning rates and bagging, generally improved performance as expected. Our overall conclusion is that noise-tolerant spam filters perform not quite as well as the best filters with clean data, but not nearly as poorly as previously reported.

11. REFERENCES

[1] The CEAS 2008 live spam challenge. http://www.ceas.cc/2008/challenge/challenge.html, 2007.

[2] Assis, F. OSBF-Lua. http://osbf-lua.luaforge.net/.

[3] Bratko, A., Cormack, G. V., Filipič, B., Lynam, T. R., and Zupan, B. Spam filtering using statistical data compression models. *Journal of Machine Learning Research 7*, Dec (2006), 2673–2698.

[4] Brodley, C. E., and Friedl, M. A. Identifying mislabeled training data. *JAIR 11* (1999), 131–167.

[5] Buckley, C., and Voorhees, E. M. Retrieval system evaluation. In *TREC - Experiment and Evaluation in Information Retrieval*, E. M. Voorhees and D. K. Harman, Eds. MIT Press, Boston, 2005.

[6] Cormack, G. V. TREC 2006 Spam Track Overview. In *Fifteenth Text REtrieval Conference (TREC-2006)* (Gaithersburg, MD, 2006), NIST.

[7] Cormack, G. V. TREC 2007 Spam Track Overview. In *Sixteenth Text REtrieval Conference (TREC-2007)* (Gaithersburg, MD, 2007), NIST.

[8] Cormack, G. V. University of waterloo participation in the trec 2007 spam track. In *Sixteenth Text REtrieval Conference (TREC-2007)* (Gaithersburg, MD, 2007), NIST.

[9] Cormack, G. V., and Bratko, A. Batch and on-line spam filter evaluation. In *CEAS 2006: The Third Conference on Email and Anti-Spam* (Mountain View, CA, 2006).

[10] Cormack, G. V., and Lynam, T. R. Spam corpus creation for trec. In *CEAS* (2005).

[11] Cormack, G. V., and Lynam, T. R. TREC 2005 Spam Track overview. http://plg.uwaterloo.ca/~gvcormac/trecspamtrack05, 2005.

[12] Cormack, G. V., and Lynam, T. R. On-line supervised spam filter evaluation. *ACM Transactions on Information Systems 25*, 3 (2007).

	LAM(%) (95% c.l.)	
filter	lmns labels	trec labels
bogom	0.54 (n/a)	0.56 (n/a)
tft1-0.5	0.73 (0.66 - 0.80)	1.80 (1.72 - 1.89)
logbagfm	1.15 (1.07 - 1.24)	2.06 (1.95 - 2.17)
dmcm	1.18 (1.09 - 1.27)	2.02 (1.92 - 2.12)
lrslow	1.19 (1.11 - 1.28)	2.10 (2.00 - 2.21)
bogo	1.21 (1.05 - 1.39)	2.58 (2.43 - 2.74)
logbagm	1.22 (1.14 - 1.30)	2.04 (1.94 - 2.15)
wat1m	1.25 (1.18 - 1.32)	2.15 (2.04 - 2.26)
logbagf	1.29 (1.20 - 1.39)	2.11 (2.02 - 2.22)
osbfm	1.44 (1.36 - 1.54)	2.21 (2.11 - 2.33)
dmc	1.76 (1.67 - 1.85)	2.70 (2.59 - 2.82)
nobs	3.24 (3.11 - 3.38)	4.02 (3.88 - 4.17)
bayes	3.38 (3.25 - 3.51)	4.27 (4.11 - 4.44)
bayesm	3.76 (3.62 - 3.90)	4.63 (4.48 - 4.78)
wat1	3.80 (3.65 - 3.95)	4.43 (4.27 - 4.59)
nobsm	3.85 (3.71 - 3.99)	4.66 (4.53 - 4.80)
tft1	5.55 (5.38 - 5.72)	6.11 (5.92 - 6.30)

Table 13: Logistic average misclassification estimates using lmns and trec labels as ground truth.

	(1-AUC)(%) (95% c.l.)	
filter	natural labels	lmns labels
logbagf	5.47 (5.37 - 5.57)	1.50 (1.46 - 1.55)
wat1m	5.51 (5.40 - 5.63)	1.47 (1.43 - 1.51)
tft1	5.54 (5.44 - 5.64)	1.55 (1.50 - 1.59)
wat1	5.75 (5.65 - 5.85)	2.93 (2.87 - 3.00)
bogo	5.87 (5.76 - 5.99)	1.59 (1.55 - 1.64)
bayes	5.97 (5.87 - 6.08)	2.24 (2.19 - 2.29)
logbagfm	6.51 (6.39 - 6.63)	2.83 (2.77 - 2.89)
tft1m	6.52 (6.42 - 6.63)	2.88 (2.82 - 2.94)
bayesm	6.64 (6.52 - 6.75)	2.88 (2.82 - 2.94)
nobs	6.72 (6.60 - 6.84)	3.30 (3.23 - 3.368)
nobsm	7.59 (7.47 - 7.70)	4.20 (4.12 - 4.28)

Table 14: ROC area estimates using natural and pseudo-gold labels as ground truth, CEAS dataset.

	LAM(%) (95% c.l.)	
filter	natural labels	lmns labels
bogo	4.49 (4.31 - 4.68)	1.07 (0.88 - 1.29)
wat1m	6.02 (5.90 - 6.15)	2.57 (2.46 - 2.68)
tft1	6.21 (6.09 - 6.33)	2.69 (2.59 - 2.79)
logbagf	6.81 (6.69 - 6.94)	3.13 (3.03 - 3.24)
bayesm	7.69 (7.55 - 7.84)	4.58 (4.44 - 4.71)
bayes	7.84 (7.72 - 7.97)	4.68 (4.57 - 4.79)
tft1m	8.01 (7.89 - 8.13)	5.12 (5.00 - 5.24)
logbagfm	8.18 (8.04 - 8.31)	5.23 (5.10 - 5.36)
nobs	8.36 (8.21 - 8.53)	6.36 (6.20 - 6.53)
nobsm	8.86 (8.70 - 9.02)	7.05 (6.91 - 7.19)
wat1	10.47 (10.33 - 10.61)	7.94 (7.81 - 8.07)

Table 15: Logistic average misclassification rate using natural and pseudo-gold labels as ground truth, CEAS dataset.

evaluation took kit. http://plg.uwaterloo.ca/~trlynam/spamjig.

[13] Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., and Bossuyt, P. M. M. The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology 56*, 11 (2003), 1129–1135.

[14] Goodman, J., and tau Yih, W. Online discriminative spam filter training. In *The Third Conference on Email and Anti-Spam* (Mountain View, CA, 2006).

[15] Graham, P. A plan for spam. http://www.paulgraham.com/spam.html, 2002.

[16] Graham-Cumming, J. SpamOrHam. *Virus Bulletin* (2006-06-01).

[17] Kocz, A., and Alspector, J. SVM-based filtering of E-mail spam with content-specific misclassification costs. *TextDM 2001 (IEEE ICDM-2001 Workshop on Text Mining)* (2001).

[18] Lam, C. P., and Stork, D. G. Evaluating classifiers by means of test data with noisy labels. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03)* (Acapulco, Mexico, 2003).

[19] Lynam, T., and Cormack, G. Trec spam filter

[20] Lynam, T. R., and Cormack, G. V. On-line spam filter fusion. In *29th ACM SIGIR Conference on Research and Development on Information Retrieval* (Seattle, 2006).

[21] Mojdeh, M., and Cormack, G. V. A mail client plugin for privacy-preserving spam filter evaluation. In *Proceedings of the 5th Conference on Email and Anti-Spam (CEAS 2008)* (2008).

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

[23] Rebbapragada, U., and Brodley, C. E. Class noise mitigation through instance weighting. In *Proceedings of the 18th European Conference on Machine Learning* (2007).

[24] Robinson, G. A statistical approach to the spam problem. http://www.linuxjournal.com/article.php?sid=6467, 2003.

[25] Sculley, D., and Cormack, G. V. Filtering spam in the presence of noisy user feedback. In *Proceedings of the 5th Conference on Email and Anti-Spam (CEAS 2008)* (2008).

[26] Sculley, D., and Wachman, G. M. Relaxed online support vector machines for spam filtering. In *30th ACM SIGIR Conference on Research and Development on Information Retrieval* (Amsterdam, 2007).

[27] Sculley, D., and Wachman, G. M. Relaxed online SVMs in the TREC Spam Filtering Track. In *Sixteenth Text REtrieval Conference (TREC-2007)* (Gaithersburg, MD, 2007), NIST.

[28] Sebastiani, F. Machine learning in automated text categorization. *ACM Computing Surveys 34*, 1 (2002), 1–47.

[29] Swets, J. A. Effectiveness of information retrieval systems. *American Documentation 20* (1969), 72–89.

[30] tau Yih, W., McCann, R., and Kołcz, A. Improving spam filtering by detecting gray mail. In *Proc. CEAS 2007 – Fourth Conference on Email and Anti-Spam* (Mountain View, CA, 2007).