Scalability of Continuous Active Learning for Reliable High-Recall Text Classification

Gordon V. Cormack
University of Waterloo
gvcormac@uwaterloo.ca

Maura R. Grossman
Univerity of Waterloo
maura.grossman@uwaterloo.ca

ABSTRACT

For finite document collections, continuous active learning ("CAL") has been observed to achieve high recall with high probability, at a labeling cost asymptotically proportional to the number of relevant documents. As the size of the collection increases, the number of relevant documents typically increases as well, thereby limiting the applicability of CAL to low-prevalence high-stakes classes, such as evidence in legal proceedings, or security threats, where human effort proportional to the number of relevant documents is justified. We present a scalable version of CAL ("S-CAL") that requires $\mathcal{O}(\log N)$ labeling effort and $\mathcal{O}(N \log N)$ computational effort—where N is the number of unlabeled training examples—to construct a classifier whose effectiveness for a given labeling cost compares favorably with previously reported methods. At the same time, S-CAL offers calibrated estimates of class prevalence, recall, and precision, facilitating both threshold setting and determination of the adequacy of the classifier.

Keywords: Technology-assisted review; TAR; predictive coding; electronic discovery; eDiscovery; test collections; relevance feedback; continuous active learning; CAL; text categorization; volume estimation.

1. INTRODUCTION

Continuous active learning ("CAL") is a method designed to address the technology-assisted review ("TAR") problem, which has as its objective, to find and review, in a *finite* collection, as nearly all of the relevant documents as practicable, with the least possible effort [9]. The TREC 2015 Total Recall Track addressed the TAR problem, providing as a "baseline model implementation" ("BMI"), a particular implementation of CAL [13, 29]. No system evaluated by the Track substantially or reliably outperformed BMI, which consistently achieved over 90% recall, across six collections, with a labeling and review budget for each topic equal to $2R + 1000$, where R is number of documents in the collection that are relevant to the topic.

CIKM'16 October 24-28, 2016, Indianapolis, IN, USA

ACM ISBN 978-1-4503-4073-1/16/10.

DOI: http://dx.doi.org/10.1145/2983323.2983776

In this study, we consider, for the first time, the scalability of CAL to collections with $D \to \infty$ documents, of which $R = \rho D$ are relevant, where ρ is the prevalence of relevant documents. We are particularly concerned with topics having low prevalence ($\rho \ll 1$), for which it is impractical to label a large enough random sample to be useful for supervised learning, tuning, or validation. It may also be impractical to label $\Theta(R)$ documents, as required for CAL.

Our solution, scalable continuous active learning ("S-CAL"), uses a large, initially unlabeled training set, drawn at random from a potentially infinite collection, and a synthetic relevant document constructed from a query. Batches of documents of exponentially increasing size are identified using relevance feedback, and labels are requested for a finite random sub-sample of each batch. The labeled examples comprise a stratified statistical sample of the entire collection, which is used for training and estimation.

In the following sections, we detail S-CAL, and show its running time to be $\mathcal{O}(N \log N)$, and its labeling cost to be $\mathcal{O}(\log N)$, where N is the size of the unlabeled training set. Using six large datasets, we evaluate the effectiveness of the classifier and the accuracy of the estimates produced by S-CAL, and we find that they comparable favorably to the best available baselines.

2. MOTIVATING APPLICATION

The challenge of reliably and efficiently achieving high recall for large datasets is of critical importance, but has not been well addressed in the literature. Within the context of electronic discovery ("eDiscovery") in legal matters, this need has been particularly acute, as voiced by parties and their counsel, technology providers, and the courts. Yet solutions have remained elusive. In the absence of a viable solution, parties have agreed—or been required—to undertake burdensome protocols that offer little assurance of success.

In a case of first impression concerning the use of TAR for document production in a legal matter [1], U.S. Magistrate Judge Andrew J. Peck repeated his previously published view [27] that:

> [I]f the use of [TAR] is challenged in a case before me, I will want to know what was done and why that produced defensible results. I may be less interested in the science behind the "black box" of the vendor's software than in whether it produced responsive documents with reasonably high recall and high precision.
>
> That may mean allowing the requesting party to see the documents that were used to train

the computer-assisted coding system. (Counsel would not be required to explain why they coded documents as responsive or non-responsive, just what the coding was.) Proof of a valid "process," including quality control testing, also will be important.

The protocol that was proposed by the parties and ordered by the court in that case is prototypical:

1. An initial random sample of 2,399 documents[1] (the "seed set") was used to train a classifier.

2. 4,000 additional documents identified through ad-hoc searches ("judgmental sampling") were added to the seed set.

3. At least seven rounds of training were to be performed, in which "senior attorneys (not paralegals, staff attorneys, or junior associates)" were required to review and label "at least 500 documents from different concept clusters to see if the computer is returning new relevant documents."

4. The rounds were to continue until "the computer" was deemed, according to unspecified criteria, to be "well trained and stable."

5. A final validation sample (also of 2,399 documents) would be taken from the "discards (*i.e.*, documents [classified] as non-relevant)," to estimate the number of relevant documents that were missed.

Assuming at least seven rounds of training, the protocol would entail the review of at least 12,298 documents, yet offer no assurance of quality. For many of the topics evaluated in the present study, a sample of size 2,399 would contain *no* relevant documents, and hence offer negligible information regarding the effectiveness of the classifier. Accordingly, one would be left to rely on the "science behind the 'black box.'"

Our study furthers such science by offering the following protocol that, we demonstrate, reliably achieves high recall, regardless of prevalence:

1. Using S-CAL to induce a scoring function S from a random sample of N documents with sub-sample size n, incurring a labeling effort of l documents. Appropriate choices might be $N = 350,000$ and $n = 30$, resulting in $l = 2,332$.

2. Using the estimate of prevalence $\hat{\rho}$ provided by S-CAL to set the threshold t so that the classifier,

$$C(d) = \begin{cases} relevant & [S(d) \geq t] \\ nonrelevant & [S(d) < t] \end{cases},$$

achieves a high recall target; *e.g.*, $\widehat{recall} = 0.9$.

3. Using $\hat{\rho}$ and $\widehat{recall}$, estimate precision $\widehat{prec}$.

4. If $\widehat{prec}$ is inadequate, one or more of: repeating steps 1 through 3 with a larger sub-sample size n; reducing the recall target; revising the definition of relevance; or discontinuing the review, on the grounds that the value of the information sought is not proportionate to the effort that would be required to find it.

[1] Although a sample of size 2,399 yields an estimate of *prevalence* with a "95% confidence interval (plus or minus two percent)," it does not, as widely misconstrued, yield such an estimate of *recall*, or offer any insight into the effectiveness of the classifier trained using such a sample.

3. RELATED WORK

The research literature on TAR is limited, due to its relatively recent introduction as a method for eDiscovery (*see* [3, 4, 9, 10, 12, 13, 18, 25, 28, 30]). Some aspects of TAR have been previously addressed within the context of information retrieval ("IR") evaluation, where the widely followed Cranfield paradigm (*see* [38]) requires substantially complete labeling of an evaluation dataset (*see* [15, 31, 35, 37, 45]). The same problem has been observed within the context of systematic review in evidence-based medicine [22], but the literature has limited discussion of technological solutions to this problem [6, 40]. Spam filtering, threat detection, privacy protection, audit, and investigative research reflect other potential applications of TAR.

Most recently, the TREC 2015 Total Recall Track [29] highlighted the TAR problem as a problem of interest in its own right, inviting participants to find as many relevant documents as possible—from datasets representing diverse applications—with the minimum possible labeling effort.

Solutions to the TAR problem draw from research in IR and machine learning, but at the same time, they have the potential to contribute to IR and machine-learning research by addressing a broader range of problems than just the TAR problem *per se*. Our results not only advance the state of the art in TAR, but also offer a novel approach to a more general problem that has not been well studied: reliably inducing a high-recall text classifier in a vast dataset, with low class prevalence.

3.1 The TAR Problem

TAR lies at the cusp of IR and machine learning for text categorization. TAR is similar to ad-hoc retrieval in that the objective is to find documents to satisfy an information need, given a query; however, the information need is met only when *substantially all* of the relevant documents have been retrieved. Accordingly, the TAR problem is one of active transductive learning for classification over a finite population, with an initially unlabeled training set consisting of the entire population. While TAR methods typically construct a sequence of classifiers, their ultimate objective is to produce a finite list containing substantially all relevant documents, not to induce a general classifier.

The most effective TAR method of which we are aware is "AutoTAR" [13], a version of CAL that is fully *autonomous*, in that it requires no topic- or dataset-specific tuning or adjustment of meta parameters. BMI[2] implements AutoTAR, as shown in Algorithm 1, except for its use of Sofia ML instead of SVM^{light} as the base classifier. For the purpose of this study, we found two aspects of BMI to be more suitable than AutoTAR: (i) BMI is distributed under an open-source license; and (ii) BMI has $\mathcal{O}(N \log N)$ running time. To ensure that using BMI would not result in substantially reduced effectiveness, we applied it to the same datasets for which AutoTAR results have previously been reported [13].

[2] `plg.uwaterloo.ca/~gvcormac/trecvm/`.

Algorithm 1 AutoTAR.

1: Find a relevant seed document using ad-hoc search, or construct a synthetic relevant document from the topic description.
2: The initial training set consists of the seed document identified in step 1, labeled "relevant."
3: Set the initial batch size B to 1.
4: Temporarily augment the training set by adding 100 random documents from the collection, temporarily labeled "not relevant."
5: Construct a classifier from the training set.
6: Remove the random documents added in step 4.
7: Select the highest-scoring B that have not yet been reviewed.
8: Review the documents, labeling each as "relevant" or "not relevant."
9: Add the documents to the training set.
10: Increase B by $\lceil \frac{B}{10} \rceil$.
11: Repeat steps 4 through 10 until a sufficient number of relevant documents have been reviewed.

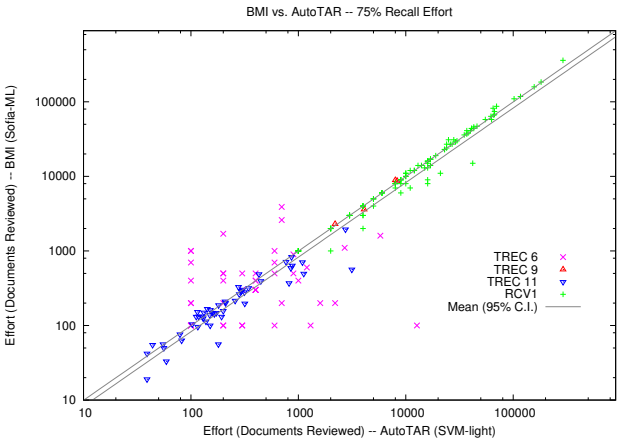

Figure 1: Sofia ML requires 90.6% as much effort to achieve 75% recall as SVM^{light}. 95% confidence limits (82.5% – 99.5%) shown in grey.

We found that BMI's Sofia ML yielded a small but significant improvement over AutoTAR's SVM^{light} (Figure 1).

The TREC 2015 Total Recall Track [29] evaluated participating systems primarily in terms of the recall they achieved as a function of review effort, which, for that task, was equivalent to labeling effort. The principal evaluation measure was recall for review effort $aR + b$, where a represents effort proportional to the number of relevant documents, and b represents fixed overhead. The TREC 2015 proceedings report all combinations of $a \in \{1, 2, 4\}$ and $b \in \{0, 100, 1000\}$. In Table 1, we reproduce the results for BMI and the best non-BMI run, where $a = 2$ and $b = 1000$, which might represent reasonable effort to find substantially all instances of a high-value class.

Recall at effort $aR + b$, even if known to be high for reasonable values of a and b, offers little guidance as to when to terminate a particular review effort. If R were known, and the TAR method were known to achieve suitably high recall for certain a and b, it would be a simple matter to continue the review until $aR + b$ documents had been retrieved and

Collection	BMI	Best Non-BMI Run	
athome1	0.956	0.952	TUW (automatic)
athome2	0.940	0.959	WaterlooClarke (automatic)
athome3	0.943	0.963	eDiscoveryTeam (manual)
MIMIC	0.969	0.973	WaterlooClarke (automatic)
Kaine	0.913	0.913	WaterlooClarke (automatic)

Table 1: TREC 2015 Total Recall Results: Recall for effort $2R + 1000$, where effort is the number of documents reviewed, and R is the number of relevant documents in the collection.

Collection	R	WC Method	Best Non-WC
athome1	4,398	0.948 (7,905)	0.927 (6,229)†
athome2	2,001	0.966 (10,473)	0.923 (3,665+683)*
athome3	643	0.953 (3,305)	0.970 (9,124)†
MIMIC	7,794	0.800 (10,624)	0.489 (4,905)†
Kaine	83,060	0.808 (83,053)	0.921 (114,895)†

Table 2: TREC 2015 Total Recall Results: Number of relevant documents (R), average recall, and average review effort (in parentheses) for different collections and review termination strategies. "WC" is the authors' WaterlooCormack system; (†) indicates the UVA-ILPS automatic method; (*) indicates the eDiscoveryTeam manual method.

reviewed. But R is generally unknown, unless additional effort is incurred to label a statistical sample from which R, recall, and other effectiveness measures, may be estimated (*see* [4]).

To test various strategies for determining when to terminate the review, TREC 2015 participants were invited to "call their shot"; that is, to indicate when they would have stopped their review to optimize various criteria, without actually stopping. Three target criteria were identified: 70% recall, 80% recall, and "reasonable and proportionate," which participants were free to interpret as they chose, but was intended to reflect a judgement call about when further improvement in recall would no longer be justified by the amount of effort necessary to achieve it. To our knowledge, no participating team made a bona fide effort to quantify recall; in our TREC 2015 submission, we simply treated the three target criteria as three interpretations of "reasonable and proportionate," with increasing emphasis on high recall at the expense of greater review effort. Table 2 shows the results for a non-statistical method we employed for stopping [11]: The review terminates when $1.2\hat{R} + 2399$ documents have been retrieved, where $\hat{R}$ is the number relevant among those retrieved. Table 2 compares this stopping criterion to the best competing result for each TREC 2015 dataset. Since TREC 2015, we have reported additional statistical and non-statistical stopping methods [8].

3.2 Text Categorization

The application of supervised machine learning to text categorization has been well studied [33], typically with small datasets that are split into training and holdout test sets (*e.g.*, Reuters 21578,[3] 20 Newsgroups[4]), often excluding classes with low prevalence. Empirical work with large

[3] `www.daviddlewis.com/resources/testcollections/.`
[4] `qwone.com/~jason/20Newsgroups/.`

datasets has primarily focused on multi-label categorization, where the objective is to label a dataset with respect to a large number of generic categories, not to achieve high recall for specific information needs (*see* [17, 26, 44]).

Perhaps the most suitable dataset for which a state-of-the-art baseline result is available is the Reuters RCV1-v2 Dataset [24], containing 803,414 news articles. Although the 103 subject categories are generic, they are sufficiently well defined and accurately labeled to serve as information needs. Lewis et al. [24] created the "LYRL2004" split, consisting of a training set of 23,149 documents, and an evaluation set of 781,265 documents, where the documents in the training set chronologically precede the documents in the evaluation set. Only 101 of the 103 subject categories are represented in the LYRL2004 training set. In the overall dataset, the prevalence of these 101 categories ranges from $6 \cdot 10^{-5} \leq \rho \leq 5 \cdot 10^{-1}$.

Lewis et al. reported that SVM^{light} achieved a macro-averaged F_1 score of 0.619, using the LYRL2004 split. We reproduced this result, achieving $F_1 = 0.620$, using the same version and configuration of SVM^{light}, the feature set computed by BMI, and a threshold setting that retrieved $\hat{\rho}N$ documents from the evaluation set, where $\hat{\rho}$ was the prevalence of relevant documents in the training set, and N was the size of the evaluation set. We have conducted a thorough literature search and found no superior result for this dataset.

Neither did our literature search find empirical results addressing the problem of reliably inducing a high-recall text classifier in a vast dataset, with low class prevalence. Bottou and Bousquet [5] consider the computational efficiency—but not the labeling efficiency—of inducing text classifiers with large training sets, using only the single highest-prevalence topic from RCV1 (CCAT, $\rho \approx 0.5$).

3.3 Active Learning

The property that distinguishes active learning from supervised learning is that with active learning, the algorithm is able to choose the documents from which it learns (*see* [34]). In a pool-based setting, which is the subject of our interest, the algorithm has access to a large pool of unlabeled examples, and requests labels for some of them. The size of the pool is limited by the computational effort necessary to process it, while the number of documents for which labels are requested is limited by the human effort required to label them. In their seminal work, Lewis and Gale [23] compared three strategies for requesting labels: random sampling, relevance sampling, and uncertainty sampling, concluding that, for a fixed labeling budget, uncertainty sampling generally yields a superior classifier. At the same time, however, uncertainty sampling offers no guarantee of effectiveness, and may converge to a sub-optimal classifier. Subsequent research in pool-based active learning has largely focused on methods inspired by uncertainty sampling, which seek to minimize classification error by requesting labels for the most informative examples (*see* [20, 34]). Over and above the problem of determining the most informative examples, the computational cost of selecting examples and re-training the classifier is of concern, motivating research into more efficient algorithms and batch learning methods [5, 7, 16, 21, 36, 43].

The Active Learning Challenge [20] employed 12 datasets, which continue to be available for on-line testing. Two of these datasets, "NOVA" and "D," represent text datasets with 19,466 and 20,000 documents, respectively, reduced to binary feature vectors. Labels are available via an on-line server for half of the documents; labels are withheld for the other half, so as to facilitate evaluation of submitted classification results, reported as area under the receiver operating characteristic curve ("AUC"), and area under the learning curve ("ALC"). ALC summarizes improvement in AUC as a function of labeling effort. AUC is a measure of the overall quality of the ranking effected by the classifier; like the 2015 Total Recall Track's $aR + b$ measure, it does not shed light on where the most appropriate cut should be made to discriminate between relevant and non-relevant documents. As a consequence, the results do not show whether or not any competing method reliably achieves high recall with limited labeling effort.

Our literature search also failed to identify empirical results applying active learning to the problem of reliably inducing a high-recall text classifier in a vast dataset. Vlachos, for example [36], considers the problem of when to stop active learning so as to optimize F_1, using only the highest-prevalence CCAT topic of RCV1. Vlachos does not consider the problem of achieving high recall, or of calculating calibrated estimates of recall, precision, or F_1. Yang et al. [44] present the results of applying active learning for multi-label classification to only a small sample (3,000 documents) of the RCV1-v2 dataset. All of the results revealed by our search used a small collection and/or a small number of high-prevalence topics, and did not consider the problems of achieving high recall or a calibrated estimate of effectiveness.

Algorithm 2 Scalable Continuous Active Learning.

1: Find a relevant seed document using ad-hoc search, or construct a synthetic relevant document from the topic description.
2: The initial training set consists of the seed docrecaument identified in step 1, labeled "relevant."
3: Draw a large uniform random sample U of size N from the document population.
4: Select a sub-sample size n.
5: Set the initial batch size B to 1.
6: Set $\hat{R}$ to 0.
7: Temporarily augment the training set by adding 100 random documents from the U, temporarily labeled "not relevant."
8: Construct a classifier from the training set.
9: Remove the random documents added in step 7.
10: Select the highest-scoring B documents from U.
11: If $\hat{R} = 1$ or $B \leq n$, let $b = B$; otherwise let $b = n$.
12: Draw a random sub-sample of size b from the B documents.
13: Review the sub-sample, labeling each as "relevant" or "not relevant."
14: Add the labeled sub-sample to the training set.
15: Remove the B documents from U.
16: Add $\frac{r \cdot B}{b}$ to $\hat{R}$, where r is the number of relevant documents in the sub-sample.
17: Increase B by $\lceil \frac{B}{10} \rceil$.
18: Repeat steps 7 through 16 until U is exhausted.
19: Train the final classifier on all labeled documents.
20: Estimate $\hat{\rho} = \frac{1.05\hat{R}}{N}$.

Dataset	Source	Description	# Docs.	# Train	# Test	Topics	ρ
RCV1-v2	Reuters	News articles	804,414	758,116	23,149	101	$6 \cdot 10^{-6} \sim 5 \cdot 10^{-1}$
AQUAINT	TREC 2005 Robust	News articles	1,033,461	750,000	283,461	50	$2 \cdot 10^{-5} \sim 4 \cdot 10^{-4}$
athome1	TREC 2015 Total Recall	Jeb Bush email	290,000	*not split*		10	$8 \cdot 10^{-4} \sim 6 \cdot 10^{-2}$
athome2		Hacker forums	465,147	*not split*		10	$4 \cdot 10^{-4} \sim 2 \cdot 10^{-2}$
athome3		Local news	902,434	*not split*		10	$3 \cdot 10^{-5} \sim 2 \cdot 10^{-3}$
MIMIC		Clinical records	31,538	*not split*		19	$6 \cdot 10^{-3} \sim 6 \cdot 10^{-1}$

Table 3: The six evaluation collections used in this study.

3.4 Electronic Discovery

In eDiscovery, statistical estimation is a subject of great interest, both for determining when the classifier is "well trained and stable," and for determining whether or not the final classifier has achieved sufficiently high recall. Ravid [28] describes the use of a random holdout set (which has come to be known in the legal community as a "control set"), that is used to estimate F_1 for successive classifiers, deeming "stabilization" to have occurred when the estimate improves insubstantially from one round to the next. The control set is drawn and labeled incrementally, until it contains at least k relevant documents. These k documents form an unbiased random sample of the population of relevant documents, from which recall may be estimated; however, in order to draw a control set with k relevant documents, it is necessary to draw and label about $\frac{k}{\rho}$ documents from the population at large. Assuming $k = 20$ (the "basic" level of quality control proposed by Ravid), drawing a control set would require an average labeling effort of 9,949 documents for each of the 101 RCV1-v2 subject categories, and 298,588 documents for the 50 TREC 2005 Robust topics used in our experiments (*see* Table 3 and Section 5). The superior, "statistical" level of quality control, for which $k = 70$, would require 3.5 times as much effort (subject to an upper bound of the size of the dataset, if it contains fewer than k relevant documents). An even higher level of $k = 385$ has been proposed as a general requirement, and adopted in at least one legal proceeding [2], because it permits the estimation of recall with a margin of error of ± 0.05, with 95% confidence.

Webber et al. [42] show that the use of a control set amounts to an invalid sequential-testing protocol that results in premature termination, and suggest empirical methods to compensate. Bagdouri et al. [4] consider how to divide a fixed labeling budget between training, and sampling for the purpose of a one-shot "certification" test.

Some have argued that the limitations of sampling preclude the application of TAR when prevalence is low, and that one must ensure high prevalence by carefully targeting the collection, or by culling the collection using ad-hoc strategies [32] (*but see* [19]). Webber [41] discusses approximate recall estimators that, while more label-efficient than the methods described above, require extensive labeling effort for small ρ.

In a recent article [8], we describe methods to reliably determine when BMI has achieved high recall. One method offers a statistical guarantee of reliably high recall based on a sample of $k = 10$ relevant documents, therefore (like the methods described above) entailing $\Omega(\rho^{-1})$ labeling effort. Another approach offers an empirically reliable method based on finding a "knee" in the gain curve representing recall as a function of cumulative effort. Both methods are specific to the TAR problem, in that they involve human review of every document that the classifier deems relevant, incurring $\Omega(R)$ labeling effort, where R is the number of relevant documents.

The work presented here is distinguished from our previous work in that, while providing a statistical estimate, it uses a different approach that requires $o(\rho^{-1})$, and $o(R)$, labeling effort. That is, it requires the review of asymptotically fewer documents than estimation using random sampling, and asymptotically fewer documents than CAL, which requires the review of every retrieved relevant document.

4. SCALABLE CAL

The essential difference between S-CAL (Algorithm 2) and CAL (Algorithm 1) is that for S-CAL, only a finite sample of documents from each successive batch is selected for labeling, and the process continues until the collection—or a large random sample of the collection—is exhausted. Together, the finite samples form a stratified sample of the document population, from which a statistical estimate of ρ may be derived. While an estimate of proportion using the Horvitz-Thompson estimator would be unbiased, in that it would yield the true proportion on average, it would almost always yield an underestimate due to the sparsity of some of the sub-samples. For practical purposes, we find that a small positive bias, as in line 20 of Algorithm 2, nearly always yields a more accurate estimate.

The estimate of prevalence $\hat{\rho}$ is used to determine the threshold setting for a given target measure. We illustrate the process for two targets: a 90% recall floor, and maximal F_1. To achieve at least 90% recall in a sample of N documents, it is necessary to retrieve $0.9\rho N$ relevant documents. To accomplish this, it is necessary to estimate the minimal m such that $0.9\rho N$ of the m top-ranked documents are relevant. Given m, we estimate the threshold setting t such that m of the N documents achieve a classifier score of t or greater. Assuming the various estimates to be accurate, any relevant document in the population will, with 90% probability, achieve a classifier score of t or greater. To maximize F_1, we observe that the maximum generally occurs near the estimated recall-precision break-even point, where $m = \hat{\rho}N$. Given m, the threshold setting t is determined as described above.

Let $U_0, U_1, \ldots U_k$ be the values of U for successive steps of Algorithm 2. Let $\hat{R}_0, \hat{R}_1, \ldots \hat{R}_k$ be the corresponding values of $\hat{R}$, and $S_0, S_1, \ldots S_k$ be the scoring functions of the corresponding classifiers. Let S_{k+1} be the scoring function of the final classifier, and let $\hat{\rho}$ be the final prevalence estimate. To target 90% recall, consider the smallest j such that $\hat{R}_j \geq 0.9\hat{\rho}N$. To target maximal F_1, consider the smallest j such that $N - |U_j| \geq \hat{\rho}N$. In either case, $m = N - |U_j|$. Let

N	n	$\overline{\left(\frac{\hat{\rho}-\rho}{\rho}\right)}$	$\overline{Effort}$	$\overline{Recall}$	$\overline{Prec.}$	$\overline{F_1}$
23,149	10	2308. (S.D. 8243.)	809	0.88	0.31	0.41
23,149	30	2111. (S.D. 6290.)	1,766	0.90	0.34	0.44
23,149	100	2208. (S.D. 6999.)	4,374	0.91	0.37	0.47
23,149	300	2110. (S.D. 6413.)	9,577	0.91	0.38	0.47
129,151	10	-0.033 (S.D. 0.201)	766	0.87	0.35	0.45
129,151	30	-0.010 (S.D. 0.117)	2,092	0.91	0.37	0.47
129,151	100	-0.017 (S.D. 0.058)	5,986	0.92	0.40	0.50
129,151	300	-0.014 (S.D. 0.053)	14,841	0.92	0.43	0.53
758,116	10	-0.005 (S.D. 0.146)	945	0.88	0.32	0.42
758,116	30	-0.010 (S.D. 0.102)	2,632	0.91	0.36	0.46
758,116	100	-0.014 (S.D. 0.044)	7,786	0.91	0.41	0.51
758,116	300	-0.010 (S.D. 0.018)	20,241	0.92	0.45	0.55

Table 4: Results on RCV1-v2 collection. Average effort, recall, precision, F_1, and prevalence error ($\frac{\hat{\rho}-\rho}{\rho}$); 101 RCV1-v2 subjects; 90% recall target.

N	n	$\overline{\left(\frac{\hat{\rho}-\rho}{\rho}\right)}$	$\overline{Effort}$	$\overline{Recall}$	$\overline{Prec.}$	$\overline{F_1}$	$\overline{\Delta F_1}$ vs. Baseline
23,149	10	2308. (S.D. 8243.)	809	0.62	0.59	0.59	$-.08$ ($p < 0.00001$)
23,149	30	2111. (S.D. 6290.)	1,766	0.62	0.65	0.61	-0.6 ($p < 0.00001$)
23,149	100	2208. (S.D. 6999.)	4,374	0.55	0.71	0.59	-0.8 ($p < 0.00001$)
23,149	300	2110. (S.D. 6413.)	9,577	0.52	0.72	0.58	-0.9 ($p < 0.00001$)
129,151	10	-0.033 (S.D. 0.201)	766	0.65	0.63	0.63	$-.04$ ($p < 0.01$)
129,151	30	-0.010 (S.D. 0.117)	2,092	0.69	0.67	0.68	$+.01$
129,151	100	-0.017 (S.D. 0.058)	5,986	0.68	0.71	0.69	$+.02$
129,151	300	-0.014 (S.D. 0.053)	14,841	0.67	0.73	0.69	$+.02$ ($p < 0.01$)
758,116	10	-0.005 (S.D. 0.146)	945	0.66	0.63	0.64	$-.03$ ($p < 0.01$)
758,116	30	-0.010 (S.D. 0.102)	2,632	0.69	0.67	0.68	$+.01$
758,116	100	-0.014 (S.D. 0.044)	7,786	0.72	0.71	0.71	$+.04$ ($p < 0.00001$)
758,116	300	-0.010 (S.D. 0.018)	20,241	0.72	0.71	0.71	$+.04$ ($p < 0.00001$)

Table 5: Results on RCV1-v2 collection. Average effort, recall, precision, F_1, and prevalence error ($\frac{\hat{\rho}-\rho}{\rho}$); 101 RCV1-v2 subjects; maximal F_1 target. ΔF_1 indicates the difference from the SVM^{light} result with 23,149 random training examples; p-values computed using paired t-test, shown where ($p < 0.05$).

$t = \max_{d \in U_0 \setminus U_j} S_j(d)$. The final binary classifier is:

$$C(d) = \begin{cases} relevant & [S_{k+1}(d) \geq t] \\ nonrelevant & [S_{k+1}(d) < t] \end{cases}.$$

Two aspects of this calculation are conservative in that they likely yield slightly higher recall, and hence slightly lower F_1, than intended. First, the only threshold values that are considered coincide with the labeling of a complete batch, although the optimal cutoff might be in the middle of a batch. Second, a sequence of intermediate classifiers, rather than the final classifier, is used to determine the ranking for which the threshold is derived. We would expect the intermediate classifiers—whose use is occasioned by the need to avoid using documents for both training and estimation—would yield a slightly inferior ranking, and therefore a higher threshold t than optimal for the final classifier.

5. EXPERIMENT

We applied S-CAL to 101 RCV1-v2 subject categories, configured to target 90% recall, and also configured to target maximal F_1. To test the transferability of our results, we also applied S-CAL to the 50 topics of the TREC 2005 Robust Track [39]. To test the applicability of our results to the TAR problem, we also applied S-CAL, configured to target 90% recall, to four of the five collections used in the TREC 2015 Total Recall Track [29]; the fifth collection (Kaine) was not used because it is non-public, and therefore was not available to us.

Statistics for the datasets are shown in Table 3. For RCV1-v2, we used only the 781,265 LYRL2004 "test documents" as our universe, which we split randomly into a training set of 758,116 documents, and a hold-out test set of 23,149 documents. We excluded the LYRL2004 "training documents" in order to control for the effect of their non-random selection when comparing our results to the baseline of Lewis et al. [24]. We used the AQUAINT dataset from the TREC 2005 Robust Track—a large labeled collection with low-prevalence topics—with a random split. We did not split the Total Recall datasets into training and test sets, so as to model the TAR problem, comparing our results to those achieved at TREC 2015.

We did no tuning based on these datasets. The only parameters of our method (beyond those embodied in BMI, which we used without modification) are: the positive bias factor of 1.05 that we used in our calculation of $\hat{\rho}$; the size N of the unlabeled training sample U; and, the sub-sample size limit n. The choice of bias factor was informed by our prior research [14] into the effectiveness of statistical esti-

N	n	$\overline{\left(\frac{\hat{\rho}-\rho}{\rho}\right)}$	$\overline{Effort}$	Target	$\overline{Recall}$	$\overline{Prec.}$	$\overline{F_1}$
750,000	30	0.028 (S.D. 0.169)	2,642	90% Recall	0.92	0.17	0.26
				Max. F_1	0.57	0.70	0.61

Table 6: Results on TREC 2005 Robust Track collection. Average effort, recall, precision, F_1, and prevalence error ($\frac{\hat{\rho}-\rho}{\rho}$); 50 topics; 90% recall and maximal F_1 targets.

			S-CAL			WaterlooCormack (TREC 2015)	
Dataset	N	n	$\overline{\left(\frac{\hat{\rho}-\rho}{\rho}\right)}$	$\overline{Effort_{train}}$	$\overline{Effort_{overall}}$	$\overline{Recall}$	$\overline{Effort_{overall}}$	$\overline{Recall}$
athome1	290,099	30	0.025 (S.D. 0.042)	2,332	9,504	0.90	7,905	0.95
athome2	460,881	30	0.033 (S.D. 0.050)	2,482	7,128	0.93	10,473	0.97
athome3	902,434	30	0.045 (S.D. 0.023)	2,662	3,160	0.94	3,305	0.95
MIMIC	31,174	30	0.042 (S.D. 0.040)	1,642	15,786	0.90	10,624	0.80

Table 7: Results on four TREC 2015 Total Recall collections. Transductive average recall, training effort, overall effort, and prevalence error ($\frac{\hat{\rho}-\rho}{\rho}$); 49 topics over 4 datasets; 90% recall target. These results may be compared to the WaterlooCormack 2015 TREC results reproduced from Table 2.

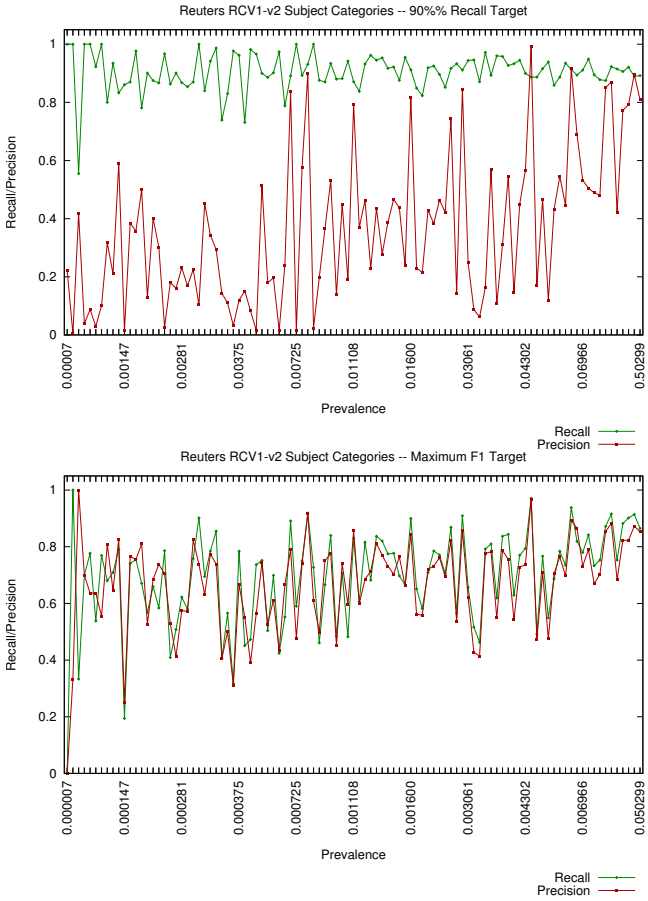

Figure 2: Plots showing per-topic recall and precision for the RCV1-v2 collection, $N = 758116$, $n = 30$.

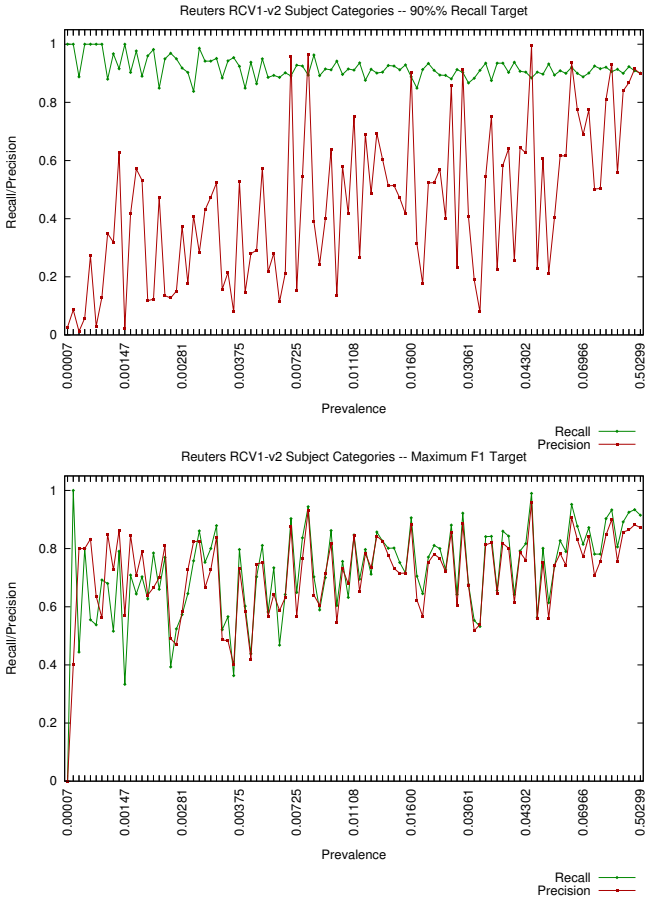

Figure 3: Plots showing per-topic recall and precision for the RCV1-v2 collection, $N = 758116$, $n = 300$.

mates for IR evaluation. For RCV1-v2, we tested three values of N in geometric progression, from 23,149 (the size of the LYRL2004 training set) through 758,116 (the maximum possible). For RCV1-v2, we used four values of n in approximate geometric progression, from $n = 10$, which entails labeling effort on the order of 1,000 documents, through $n = 300$, which entails labeling effort on the order of 20,000 documents—similar to the size of the LYRL2004 training set. For the other datasets, we conducted only one experiment, using the maximum possible N, and $n = 30$.

The results for 101 RCV1-v2 topics are summarized in Tables 4 and 5, and Figures 2 and 3. Table 4 shows, for the

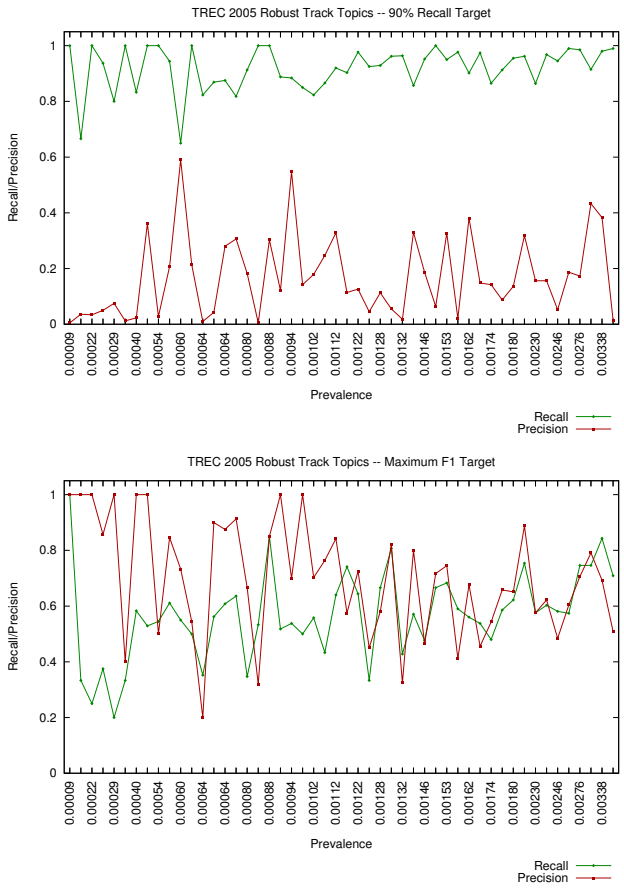

Figure 4: Plots showing per-topic recall and precision for the TREC 2005 Robust Track collection, $N = 750000$, $n = 30$.

parameter combinations described above, the average effort, $\hat{\rho}$ estimation error, recall, precision, and F_1 achieved when the threshold t is chosen to target 90% recall. Table 5 shows the same measures when t is chosen to optimize F_1. All combinations of N and n achieve what appear to be reasonable results for recall and F_1, although combinations involving the smallest value of N or n are inferior. $\hat{\rho}$ estimation error appears reasonable for $N = 129,151$ and $N = 758,116$, but unreasonably large for $N = 23,149$, due to a handful of extreme outliers.

Figures 2 and 3 show the recall and precision achieved for each individual topic for the largest N, ordered by prevalence. Figure 2 reflects $n = 30$, while Figure 3 reflects $n = 300$. The results in the top panels target 90% recall; the results in the bottom panels target maximal F_1. Both configurations appear to yield generally consistent results for both targets; the larger value of n achieves higher consistency, and higher precision, at the expense of substantially higher labeling effort.

Overall, the RCV1-v2 results suggest that it is beneficial to make N as large as possible, and to choose n to balance classifier effectiveness with labeling effort. When the 90% recall target is chosen, that target is generally met, on average, so improved effectiveness is reflected in lower variance and improved precision; when the maximal F_1 target is chosen, improved effectiveness is reflected in lower variance, improved precision, and improved recall.

While most of the F_1 scores reported in Table 2 are numerically greater than the $F_1 = 0.619$ reported by Lewis et al. [24], the scores are incomparable because the latter used the LYRL2004 training set, which is not a sample of the same population as the test set. Moreover, our test set was a much smaller sample of the LYRL2004 test set, which could result in chance differences. To adjust for these differences, we first ran SVM^{light} on the LYRL2004 split, and observed $F_1 = 0.620$, thus confirming that our implementation was comparable to Lewis et al.'s. We then trained the same implementation using our $N = 23,149$ training sample, and evaluated it using our test set, both of which contained 23,149 documents drawn from the LYRL2004 test set. The resulting F_1 score was 0.669, which we used as our baseline. Table 5 shows the difference between our F_1 results and 0.669, and the p-value resulting from a paired t-test.

The 2005 Robust Track and 2015 Total Recall experiments were configured with knowledge of the RCV1-v2 results. For both experiments, we predicted that the maximum possible N, and $n = 30$ would yield a reasonable compromise between effectiveness and labeling effort.

Summary results for the 50 TREC 2005 Robust Track topics and AQUAINT dataset are shown in Table 6. $\hat{\rho}$ error is positive and somewhat larger in magnitude, with larger variance than for the RCV1-v2 dataset, perhaps due to the preponderance of low-prevalence topics. The 90% recall target was exceeded, on average, and a maximum F_1 score of 0.61 appears adequate, notwithstanding the lack of an available baseline against which to compare. Per-topic results are shown in Figure 4.

Our final experiment departs from inductive classification to test the use of S-CAL for TAR. In the first phase, S-CAL is used to induce a classifier, using the entire collection as a training set. The classifier is then applied to the same collection, and any documents classified relevant but not labeled during the training process are retrieved. Gain is measured by recall; cost is the total number of documents retrieved, whether by the training process or by the final classifier. Table 7 shows the results on four of the TREC 2015 Total Recall datasets. The recall target of 90% is achieved, on average, for all datasets, but surpassed by the WaterlooCormack 2015 TREC submission for three of the datasets. For the fourth dataset, S-CAL achieves 90% recall compared to WaterlooCormack's 80%, at the cost of 50% more effort.

All experiments except MIMIC were run on a shared server with AMD Opteron 6274 processors having a total of 64 cores. The BMI feature extraction step was done once, and all topics were run in parallel. Only the TREC 2005 Robust Track experiments were conducted at a time when the server was lightly loaded. Featurization of the AQUAINT dataset took 1 hour, 9 minutes; the 50 topics were run concurrently, with a total elapsed time of 5 hours, 1 minute. Running times for RCV1-v2 were comparable: feature extraction took about one hour, and the 101 topics were run concurrently in about seven hours. We were unable to recover the feature-extraction times for the Total Recall runs. The 49 topics were started concurrently: The athome1 topics were completed in 1 hour, 8 minutes; the athome2 topics were completed in 1 hour 46 minutes; the athome3 topics were completed in 5 hours, 23 minutes. The MIMIC topics,

which were run on an AMD FX-8320 eight-core processor, were completed in 35 minutes.

6. DISCUSSION

The rationale for S-CAL is outlined in the following steps:

1. A naïve approach to constructing an amenable training set for supervised learning would be to:
   (a) draw a random sample of N documents,
   (b) use CAL to find and label nearly all relevant documents in the sample,
   (c) presumptively label the remaining documents as "non-relevant."
2. Labeling effort for the naïve approach is proportional to ρN, but ρ is unknown.
3. If $\rho \approx 0$,
   (a) we can afford to label ρN documents for very large N,
   (b) we require very large N to have sufficient positive training examples to yield an effective classifier,
   (c) the initial batches with size $B \leq n$ will likely contain sufficient positive training examples,
   (d) samples of later batches with size $B > n$ will contain representative negative training examples.
4. If $\rho \gg 0$,
   (a) we cannot afford to label ρN documents,
   (b) a sample of the ρN relevant documents is sufficient to yield an effective classifier,
   (c) the initial batches with size $B \leq n$ will contain only the most likely relevant positive training examples,
   (d) samples of later batches with size $B > n$ will contain representative less-likely relevant positive examples, as well as representative negative examples.
5. In either case, the labeled documents form a stratified sample, from which ρ can be estimated.
6. The sequence of classifiers constructed by S-CAL can be used as a surrogate to estimate the effectiveness of the final classifier by employing the classifier constructed in round k to predict the labels of the documents sampled in round $k + 1$.
7. Given an estimate of ρ, and an estimate of classifier effectiveness, the threshold t may be chosen to optimize a target measure.

While our empirical results indicate that S-CAL solves a vexing problem that has eluded many—including the authors—for years, it is not magic. If we demand 90% recall for a topic that cannot be classified, extremely low precision will result. If we demand optimal F_1, extremely low (but near-optimal) F_1 will ensue. Our results, combined with the existing body of results for CAL [9, 10, 13, 29], suggest that, for well-defined topics, CAL—and hence S-CAL—can achieve superior results. Further characterizing what is meant by "well-defined topics" remains an avenue for further investigation.

An important benefit of S-CAL is predictability. For a fixed labeling budget, S-CAL offers not only an effective classifier, but also calibrated estimates of prevalence, recall, and precision. The estimate of precision can in turn be used to estimate the overall labeling effort if, in addressing the TAR problem, it will be necessary to label all positively classified documents. This—or other cost/benefit measures—may be calculated in support of scheduling, resource allocation, and the decision of whether to proceed, to revise the problem, or to seek a better classifier.

A number of our design choices are worthy of further investigation. The limit n on sub-sample size b need not be a constant, and indeed is not a constant in our implementation: When $\hat{R} = 0$, we allow b to grow beyond n to handle the pathological case in which the classifier fails to find any relevant documents in the first several batches. b could be a more complex function of $\hat{R}$, or n could shrink or grow in each iteration, rather than remaining constant. Any exponential growth rate will be efficient, so long as it is smaller than the 10% growth rate of B. More complex distributions might be used to determine the sampling rate.

We chose to use relevance sampling because it worked well for CAL. Algorithm 2 could be adapted to select documents in a different order, perhaps to prefer the most informative examples, as in uncertainty sampling. However, it is not immediately obvious how to set the threshold based on the resulting sequence of classifiers.

Instead of establishing a fixed labeling budget and proceeding through the documents in order of likely relevance, an incremental version of S-CAL might be designed to label as many documents as necessary to achieve an effectiveness target. Also of interest would be an on-line version of S-CAL, in which a stream of examples is presented to—rather than selected by—the algorithm, which must choose whether to label them or not.

There is a paucity of published baseline results for large-scale high-recall text classification, and for active learning in particular. We suggest that the test collections we have used here—Reuters RCV1-v2, TREC 2005 Robust, and TREC 2015 Total Recall—should be adopted as standard benchmarks for future investigations, with our BMI implementation of S-CAL as a baseline.

7. REFERENCES

[1] *Da Silva Moore v. Publicis Groupe*. 287 F.R.D. 182, S.D.N.Y., 2012.

[2] Case Management Order: Protocol Relating to the Production of Electronically Stored Information ("ESI"). In *In Re: Actos (Pioglitazone) Products Liability Litigation*. MDL No. 6:11-md-2299, W.D. La., July 27, 2012.

[3] M. Bagdouri, D. D. Lewis, and D. W. Oard. Sequential testing in classifier evaluation yields biased estimates of effectiveness. In *SIGIR 2013*.

[4] M. Bagdouri, W. Webber, D. D. Lewis, and D. W. Oard. Towards minimizing the annotation cost of certified text classification. In *SIGIR 2013*.

[5] L. Bottou and O. Bousquet. Learning using large datasets. *Mining Massive DataSets for Security*, 2008.

[6] A. M. Cohen, W. R. Hersh, K. Peterson, and P.-Y. Yen. Reducing workload in systematic review preparation using automated citation classification. *J. Am. Med. Inform. Assoc.*, 13(2), 2006.

[7] R. Collobert, F. Sinz, J. Weston, and L. Bottou. Large scale transductive svms. *J. Mach. Learn. Res.*, 7, 2006.

[8] G. V. Cormack and M. R. Grossman. Engineering quality and reliability in technology-assisted review. In *SIGIR 2016*.

[9] G. V. Cormack and M. R. Grossman. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In *SIGIR 2014*.

[10] G. V. Cormack and M. R. Grossman. Multi-faceted recall of continuous active learning for technology-assisted review. In *SIGIR 2015*.

[11] G. V. Cormack and M. R. Grossman. Waterloo (Cormack) participation in the TREC 2015 Total Recall Track. In *TREC 2015*.

[12] G. V. Cormack and M. R. Grossman. *Systems and methods for classifying electronic information using advanced active learning techniques*. United States Patent 9122681, 2013.

[13] G. V. Cormack and M. R. Grossman. Autonomy and reliability of continuous active learning for technology-assisted review. *arXiv:1504.06868*, 2015.

[14] G. V. Cormack and E. Lee. Information retrieval effectiveness measurement using very sparse relevance assessments. Technical report, University of Waterloo, 2011.

[15] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient construction of large test collections. In *SIGIR 1998*.

[16] T.-N. Do and J.-D. Fekete. Large scale classification with support vector machine algorithms. In *ICMLA 2007*.

[17] A. Esuli and F. Sebastiani. Active learning strategies for multi-label text classification. In *Advances in Information Retrieval*. Springer, 2009.

[18] M. R. Grossman and G. V. Cormack. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond J. L. & Tech.*, 17(3), 2011.

[19] M. R. Grossman and G. V. Cormack. Comments on "The implications of Rule 26(g) on the use of technology-assisted review". *Fed. Cts. L. Rev.*, 7, 2014.

[20] I. Guyon, G. C. Cawley, G. Dror, and V. Lemaire. Results of the Active Learning Challenge. *Workshop on Active Learning and Experimental Design*, JMLR Workshop and Conference Proceedings 16, 2011.

[21] S. C. Hoi, R. Jin, and M. R. Lyu. Large-scale text categorization by batch mode active learning. In *WWW 2006*.

[22] C. Lefebvre, E. Manheimer, and J. Glanville. Searching for studies. *Cochrane handbook for systematic reviews of interventions*, 2008.

[23] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *SIGIR 1994*.

[24] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397, 2004.

[25] P. Oot, A. Kershaw, and H. L. Roitblat. Mandating reasonableness in a reasonable inquiry. *Denver L. Rev.*, 87:533, 2010.

[26] I. Partalas, A. Kosmopoulos, N. Baskiotis, T. Artieres, G. Paliouras, E. Gaussier, I. Androutsopoulos, M.-R. Amini, and P. Galinari. Lshtc: A benchmark for large-scale text classification. *arXiv:1503.08581*, 2015.

[27] A. Peck. Search, forward: Will manual document review and keyword searches be replaced by computer-assisted coding? *Law Tech. News*, Oct. 1, 2011.

[28] Y. Ravid. *System for Enhancing Expert-Based Computerized Analysis of a Set of Digital Documents and Methods Useful in Conjunction Therewith*. United States Patent 8527523, 2013.

[29] A. Roegiest, G. V. Cormack, M. R. Grossman, and C. L. A. Clarke. TREC 2015 Total Recall Track Overview. In *TREC 2015*.

[30] H. Roitblat, A. Kershaw, and P. Oot. Document categorization in legal electronic discovery: Computer classification vs. manual review. *J. Assoc. Inf. Sci. Technol.*, 61(1), 2010.

[31] M. Sanderson and H. Joho. Forming test collections with no system pooling. In *SIGIR 2004*.

[32] K. Schieneman and T. Gricks. The implications of Rule 26(g) on the use of technology-assisted review. *Fed. Cts. L. Rev.*, 7, 2013.

[33] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1), 2002.

[34] B. Settles. *Active learning literature survey*. University of Wisconsin, Madison, 2010.

[35] I. Soboroff and S. Robertson. Building a filtering test collection for TREC 2002. In *SIGIR 2003*.

[36] A. Vlachos. A stopping criterion for active learning. *Comput. Speech Lang.*, 22(3), 2008.

[37] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Inf. Process. Manag.*, 36(5), 2000.

[38] E. M. Voorhees. The philosophy of information retrieval evaluation. In *Evaluation of cross-language information retrieval systems*, 2002.

[39] E. M. Voorhees. The TREC 2005 Robust Track. In *ACM SIGIR Forum*, volume 40. ACM, 2006.

[40] B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos. Active learning for biomedical citation screening. In *KDD 2010*.

[41] W. Webber. Approximate recall confidence intervals. *ACM Trans. Inf. Syst.*, 31(1), 2013.

[42] W. Webber, M. Bagdouri, D. D. Lewis, and D. W. Oard. Sequential testing in classifier evaluation yields biased estimates of effectiveness. In *SIGIR 2013*.

[43] Z. Xu, C. Hogan, and R. Bauer. Greedy is not enough: An efficient batch mode active learning algorithm. In *ICDMW 2009*.

[44] B. Yang, J.-T. Sun, T. Wang, and Z. Chen. Effective multi-label active learning for text classification. In *KDD 2009*.

[45] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *SIGIR 1998*.