Comparison of Tools and Methods for Technology-Assisted Review

Tom O'Halloran*
Information Retrieval Services
Grant Thornton Ireland
Dublin, Ireland
ohalloranthomas1@gmail.com

Bronagh McManus
Information Retrieval Services
Grant Thornton Ireland
Dublin, Ireland
bronagh.mcmanus@ie.gt.com

Andrew Harbison
Information Retrieval Services
Grant Thornton Ireland
Dublin, Ireland
aharbson@hotmail.com

Maura R. Grossman
University of Waterloo
Waterloo, Ontario, Canada
maura.grossman@uwaterloo.ca

Gordon V. Cormack
University of Waterloo
Waterloo, Ontario, Canada
gvcormack@uwaterloo.ca

***Abstract*—In a large-scale eDiscovery effort in Irish litigation, human assessors participated in two technology-assisted reviews ("TAR") employing continuous active learning ("CAL") processes, one using Grossman and Cormack's logistic regression CAL tool and the other using a leading eDiscovery provider's support-vector-machine-based ("SVM") tool. In this work, we investigate the extent to which the different tools and associated methods impacted the effectiveness and efficiency of the competing TAR reviews across the same document population, measured by recall, precision, review effort, and the average cost incurred per relevant document found. Our results show that the tool and method underlying the TAR model matters – the CAL process outperformed the provider's process on all measures.**

Keywords—Continuous active learning, CAL, logistic regression, support vector machine, SVM, LIBLINEAR, Technology-Assisted Review, TAR, information retrieval, Big data, knowledge management

I. INTRODUCTION

The objective of high-recall information retrieval ("HRIR") is to find all or nearly all relevant documents in a collection. Applications include electronic discovery (eDiscovery) in legal matters [1], systematic review in evidence-based medicine [2], and the creation of test collections for information-retrievals tasks [3, 4]. In the litigation context, it is typical for parties to review documents in response to requests for production ("RFPs") received from their adversaries, which specify categories of information sought concerning the dispute, either agreed on by the parties, or mandated by the Court or Arbitrator. The onus is on the producing party to identify as much responsive material relating to these RFPs as possible from their own datasets to hand over to the opposing side [2].

In the last 20 years, the discovery landscape has changed dramatically with the proliferation of electronically stored information ("ESI"). The total data created, captured, copied, and consumed globally has risen from 2 to 64 zettabytes ("ZB") from 2010 to 2020, with a forecasted increase to 181 ZB by 2025 [5]. The digitization of information and the continuous development of new forms of communication and unstructured data types subject to the eDiscovery process means that data volumes to be collected and searched for relevant material have grown year over year. The challenge imposed on the legal sector is, therefore, how to identify all or nearly all the relevant material in ever-expanding document collections in order for a party to discharge its discovery obligations in a manner that is proportionate and cost effective.

Traditionally, exhaustive manual review has involved the inspection of every document in a dataset to identify the responsive ones, either with or without pre-culling using Boolean search terms. This meant that the length of time required and the cost associated with completing a document review was directly proportional to the number of documents in the corpus. With continually growing corpora, this has raised significant issues for legal systems in the digital era, from both a practical and cost perspective. If the average reviewer can review approximately 300 documents per day, it would take them 9 years, with no days off, to review a collection containing one million documents, and the cost incurred in doing so might far outweigh the amount of the claim. One solution to this problem was introduced with the appearance of technology-assisted review ("TAR"). Studies have shown that TAR can achieve superior results in terms of recall, precision, and review effort compared to Boolean search methods and exhaustive manual review [6, 7].

The landmark U.S. decision in *Da Silva Moore v. Publicis Groupe*, which sanctioned the use of TAR in eDiscovery, means that the time and cost incurred in completing a legal review is now proportionate to the number and importance of relevant documents in a dataset, not the total number of documents to be searched. In his opinion, U.S. Magistrate Judge Andrew J. Peck opined [8] that:

> If 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.

Unfortunately, as predicted by Goodhart's law [9], the focus on achieving "high recall and high precision" has encouraged gaming of metrics performed to validate a production effort.

The first objective of our study is to assess to what extent the "science behind the black box of the vendor's software" (or learning algorithm), impacts the user's ability to achieve high recall and high precision in real case scenarios with live case data. Our second objective is to address the notion of defensibility and defensible results through an analysis of the reliability of current validation methods used in practice to evaluate the quality of a review, as measured, in particular, by recall. Recall is defined as the fraction of all relevant documents found by the user [10].

We achieve our objectives by conducting a comparative analysis of the performance of CAL® against a leading eDiscovery provider's TAR tool, which uses a modified instance of LIBLINEAR, an open source algorithm. Effectiveness is measured using the metrics of recall, precision, review effort, and cost incurred per relevant document found. The results show the provider's tool to be less effective than CAL® on all metrics, suggesting that the underlying algorithm employed in a continuous active learning TAR review can profoundly impact the success of a review effort. The results also show that the validation protocol recommended by the provider, commonly referred to as the Elusion test [10] - produces inconsistent and improbably high recall estimates, calling into question its continued use in eDiscovery.

This work is novel because the study was conducted in the context of a real litigation with actual "live" data and trained barristers conducting the review, rather than in the lab with a non-representative, "clean" dataset and volunteer reviewers.

II. BACKGROUND AND RELATED WORK

HRIR has been a long-standing area of interest in the field of IR. As previously mentioned, the objective of TAR is to identify substantially all the relevant documents in a collection with a reasonable review effort. The TAR problem differs from other types of IR, such as text categorisation, pooling, routing, and others, insofar as it requires identifying all or most of the relevant documents in a collection with no prior knowledge of the dataset [1, 11]. This section highlights the principal related work and provides background on the HRIR methods we studied.

In their seminal 1985 study, Blair and Maron evaluated the effectiveness of using Boolean search queries to identify relevant documents with the goal of achieving a target of 75% recall. Experienced searchers stopped the review when they believed they had achieved that goal, but in fact, they had attained only 20% recall [12].

Additional forms of HRIR include the pooling method. This method is used to evaluate IR effectiveness by forming a judging pool from highly ranked representative documents relating to an information need, assessed for relevance and serving as the "gold standard" (or "ground truth") to evaluate the retrieval efforts used to form the pool [13]. Using their interactive searching and judging ("ISJ") method involving the review of highly ranked documents, Cormack et al. were able to yield a judging pool with equivalent IR evaluation effectiveness to a pooling method-derived pool that was five times larger than the ISJ pool, on the TREC 6 evaluation set [14].

Studies by Lewis show that SVM classifiers using SVMlight, applied to the Reuters RCV1 dataset, performed well in routing and text categorisation tasks achieving F1 =0.619 [15]. Both routing and text categorisation differ from the problem faced in the TAR context because they require identifying relevant material with respect to a known subject matter, within a new collection, instead of a previously unknown subject matter, as is typically the case in litigation. Routing updates document ranks for subsequent review, while text categorisation automatically tags as relevant all documents above a predefined rank cut-off, thereby not requiring further review [15].

Studying active learning, Grossman and Cormack show in their comparative study on three different TAR protocols that CAL® is superior to Simple Passive Learning ("SPL") and Simple Active Learning ("SAL") methods in terms of recall and review effort required to complete a TAR review [1]. SPL involves selecting a training set without using the learning algorithm, usually by selecting a random sample of documents. The training set is reviewed and used to generate a candidate review set. CAL® and SAL both use seed documents initially to train the algorithm. However, thereafter, SAL continues training using uncertainty sampling [16], whereas CAL® continues training using relevance feedback.

The current state of the art is an improved version of Grossman and Cormack's CAL® system, known as AutoTAR, but referred to simply as CAL® here for convenience. The difference lies in several improvements, including tf-idf features, a single relevant seed document, presumptively labelled 'not relevant' documents, and exponentially larger batch sizes, which together improve recall [11]. The Baseline Model Implementation ("BMI") is a free version of AutoTAR used by participants to automate the TREC 2015 Total Recall Track [17]. The primary difference between AutoTAR and BMI is the underlying algorithm, AutoTAR uses SVMlight and BMI uses logistic regression.

Additional noteworthy research [18, 19, 20, 21] addresses the commonly asked question associated with TAR reviews in eDiscovery: When can the review be stopped? Grossman and Cormack offer empirically and mathematically reliable methods in answering this question. One such method, which they have named "the Knee" Method, involves a geometric stopping procedure based on the shape of the gain curve, and achieves reliability for recall regardless of prevalence [18].

Volume estimation is a problem that relates to accurately predicting the number of relevant documents in a collection [22]. It is possible to achieve this result by combining the Knee Method (finding all high-scoring documents pre-knee) with stratified sampling techniques to estimate the likely number of relevant documents still unfound. Grossman and Cormack further improve on this stratified sampling method to create Scalable-Continuous Active Learning ("S-CAL"), which accurately and reliably predicts the number of documents requiring review to attain a given recall early in the TAR review [23]. This method importantly allows for cost and budget estimations and permits adjustments, reformulation, and refinement of RFPs, etc. where indicated.

We base our study on Grossman and Cormack's BMI of CAL® [17] and a leading eDiscovery provider's LIBLINEAR's linear kernel SVM classifier as employed in a widely used eDiscovery review platform [proprietary feature engineering and hyperparameter settings].

III. METHODS AND MATERIALS

In this section, we describe our experiment in detail. We describe the search topics and document collection, the experimental design, the active learning algorithms and their implementation, and other details of the experiment, including how we measured performance and validation of each TAR tool and protocol.

A. Search Topics and Documents

Our document collection arose from the data collected on behalf of the plaintiff in a large-scale Irish litigation involving the accounting practices preceding the failure of a major insurance company, with claims of approximately US$1 billion.

In total, over 300 million documents were collected from many data custodians and sources. This collection included a subset of data consisting of 169,000 documents relating to three key custodians in their role as company administrators. The defendant requested that these data be reviewed and produced separately; therefore, this subset of documents was used for testing purposes. The defendant in the proceedings identified 55 individual RFPs specifying different categories of information sought. These were the basis of the search topics used in our test programme. There was considerable conceptual similarity and overlap among the 55 RFPs. Therefore, we decided to group these 55 RFPs into ten broader Consolidated-RFPs ("C-RFPs") to increase the speed of the review and to decrease the burden on reviewers in having to keep many issues in their minds at one time when reviewing for relevance. Previous studies show that such an approach is acceptable because a CAL® system using relevance feedback will achieve high levels of recall for multi-faceted RFPs without excluding any single RFP by applying a depth-first-by-width approach to identifying potentially relevant documents for review [24].

Each reviewer was provided with a briefing pack for the C-RFP they were assigned to review, which included the C-RFP description, seed documents, an appendix containing a non-exhaustive list of illustrative examples of relevant documents, as well as the wording of the underlying RFPs themselves to be referred to, if needed. Due to the commercially and personally sensitive nature of the case data and the ongoing confidentiality clause in the case settlement agreement, we are unable to release this document corpus into the public domain at present. However, the redacted wording of the C-RFPs is provided below (see Table I), and the reviewer coding decisions will be made publicly available to allow for verification of our statistical results.

B. Performance Metrics

As discussed above, eDiscovery tasks are generally understood to require high-recall retrieval. In terms of data volumes, only a tiny number of legal matters worldwide see a larger document review than our current example of 300 million documents. Given the size of the dataset, the only logistical way to conclude such a review within a reasonable timeframe and budget was to use TAR.

It is not uncommon in eDiscovery reviews for there to be two levels of relevance judging. The first-level reviewers will review the documents for relevance and categorize them into RFPs. The second level review, typically conducted using more experienced subject matter experts, consists of a QA review of the relevant documents to determine their final status and to complete a privilege and redaction review. In our experiment, we performed one review pass, plus quality control on a 10% sample, and adjudged all documents marked as responsive to be final.

TABLE I. CONSOLIDATED REQUESTS FOR PRODUCTION

C-RFP No.	C-RFP Instructions
1	Tag as relevant any documents which record or relate to [The Company]'s technical provisions (also called "technical reserves"), including how technical provisions were estimated. Documents regarding wider market practice / benchmarks for technical reserves should also be considered relevant.
2	Tag as relevant all [The Company] management information / reports and all documents relating to the management and oversight of [The Company], including in respect of regulatory and financial issues.
3	Tag as relevant any documents which relate to [The Company]'s claims handling function. In this context, "claims handling" covers the processing and administration of claims, as well as the setting / adjustment of claims reserves and the settlement of claims.
4	Tag as relevant any documents which relate to [The Company]'s underwriting or pricing policies, procedures or practices. (Note that information of general application is being targeted - individual contracts of insurance should be tagged as not relevant).
5	Tag as relevant all documents relating to [The Defendant]'s audit work for [The Company].
6	Tag as relevant all documents relating to any assessments or analysis of the insurance market, [The Company]'s own business model / practices (and those of its competitors) and/or the geographic markets in which [The Company] operated (i.e., Ireland, the UK and Europe).
7	Tag as relevant (1) all documents relating to interaction between [Separate Entity] individuals[1] or entities[2] and [The Company] in relation to [The Company]'s technical provisions or claims reserving; (2) monetary transfers or reinsurance arrangements between [The Company] and [Separate Entity] entities (other than [The Company] subsidiaries) and (3) the decision of the directors of [Separate Entity] that it could prepare its accounts on a going concern basis from 2005 – 2009.
8	For each of the guarantor subsidiaries, tag as relevant documents related to [The Company]'s consideration of their management accounts / financial statements and any documents relating to [The Company]'s knowledge of the activities of the guarantor subsidiaries.
9	Tag as relevant (1) all documents relating to the original financing (and subsequent refinancing) of [Separate Entity] Limited by [The Bank]; and (2) all documents which relate to or demonstrate knowledge of the guarantees provided by [The Company] subsidiaries within [The Company] or the [Separate Entity].
10	Tag as relevant all documents relating to the management of [The Company] by the Joint Administrators.

Given the nature of the allegations and the amount claimed in this case, achieving a high-recall production was particularly important to both parties in the dispute. Therefore, the key performance measure used to evaluate the efficacy of both TAR systems was the recall they achieved.

The perceived success of any TAR review can also hinge on other factors, beyond recall. Precision [10] refers to the fraction of documents identified by a search or review effort that are relevant to the information request, and is often measured to ensure that an excessive amount of non-responsive information has not been included in the production [10]. Further, the time and effort it takes to complete a review are also considered important, given most reviews are carried out under a court-mandated deadline. Additionally, the amount of time spent looking at non-relevant documents, a wasted effort, is often important to the party responsible for paying for the cost of the review effort. Therefore, we included measures of the precision that reviewers achieved for each C-RFP review and the total review effort (i.e., total number of documents reviewed) as key performance indicators. Intrinsically linked to precision and review effort is the cost of identifying relevant documents. We determined the amount of time taken by each reviewer to complete their task, as determined from the review platform's audit logs, to calculate the average cost of identifying relevant documents per assessor.

It was apparent that, given the scale of the review and the likely proportion of relevant documents, a TAR-based approach was the only solution likely to be at all practical in reviewing this dataset. What we were concerned with was convincing the court that the use of CAL® for this review would not lead to an inferior production set as compared to the leading eDiscovery provider's TAR tool, which was what the adversary was using.

C. Experiment Design

The question of when to terminate a review is a well-known problem in TAR reviews. The answer must factor in proportionality considerations, such as those outlined in U.S. Federal Rules of Civil Procedure 26(b)(2)(C) and 26(g)(1)(B)(iii), which state that the burden of review must not outweigh its benefits, and that discovery must not be unduly expensive. A good predictor of high-recall, and therefore an appropriate point to stop a review, is when marginal precision falls below a certain relevance rate for consecutive batches [24]. We denoted this as the "stopping point" for our reviews.

Our design specified that assessors – ten qualified and experienced barristers, paid commercial rates for their work – would review documents for relevance to one C-RFP at a time until marginal precision fell below 5% for three consecutive batches of 200 documents, our defined "stopping point." Assessors were grouped in five review pair teams; each assigned one odd and even numbered C-RFP to review on CAL® and the provider's platform. In two distinct review sessions, five review team pairs first applied the CAL® tool to C-RFPs 1, 3, 5, 7 and 9, and then the provider's tool to C-RFPs 2, 4, 6, 8 and 10. In a further two review sessions, the same C-RFPs were re-reviewed using the alternative tool, i.e., applying the provider's tool to C-RFPs 1, 3, 5, 7 and 9, and then the CAL® tool to C-RFPs 2, 4, 6, 8 and 10. The C-RFP review order is shown in Table II below.

TABLE II. C-RFP REVIEW ORDER ON CAL® AND THE PROVIDER'S TOOL

Review Order	TAR System	C-RFP
1	CAL	1,3,5,7,9
2	The Provider's Tool	2,4,6,8,10
3	The Provider's Tool	1,3,5,7,9
4	CAL	2,4,6,8,10

This review order ensured there was a suitable lapse of time (i.e., washout period) between re-reviewing the same C-RFP. Having the review teams perform the review of a C-RFP for the first time on CAL® and the provider's tool reduced the risk of memory bias unduly advantaging one system over the other when the same C-RFP was re-reviewed.

As a seed set, a minimum of five responsive and five non-responsive documents were selected for each C-RFP by subject matter experts in the law firm responsible for case management. This was done because the provider's TAR tool recommends a minimum of five responsive and five non-responsive seed documents to "kick-start" their learning model.

Upon completion of each review, we computed recall for each C-RFP. In each case, an independent reviewer performed a blind review of a Confusion matrix sample to estimate recall for the CAL® and provider's tools. In the case of the tests of the provider's tool, we further evaluated recall for each C-RFP using an Elusion test sample, consistent with the evaluation procedure recommended by the provider. An Elusion test (a common form of evaluating recall in modern commercial litigation reviews) involves a review of a random sample of documents from the unreviewed population with a given confidence level and margin of error. Relevant documents identified can be used to calculate the estimated number of documents found, and, therefore, recall. The Confusion matrix is a more statistically robust validation protocol because it measures the performance of a classification model by breaking the review into actual and predicted results, and unlike the Elusion test, considers reviewer false positives and false negatives in its recall calculations. Table III provides a breakdown of the sub-collections and document count comprising the Confusion matrix sample, and the formula used to compute recall for both TAR systems. The recall formula used to calculate recall for each TAR system was the straightforward approach originally proposed in the legal case in *re Broiler Chicken Antitrust Litigation* [25].

Review effort is the total number of documents reviewed per C-RFP. Precision is measured as the proportion of relevant documents identified in each C-RFP.

We also measured the total time spent on the review platform by each assessor in reviewing each C-RFP. We then calculated the average cost incurred to identify each relevant document by taking the overall cost to review every document based on hourly rates / number of relevant documents found. Lapses of time >5 mins between coding documents were ignored as possible breaks in review.

TABLE III. CONFUSION MATRIX SUB-COLLECTION PARTITIONS, SUB-SAMPLE SIZES, AND FORMULA USED TO ESTIMATE RECALL FOR THE PROVIDER'S TOOL AND CAL®.

Subcollection Partitions	**#Subsample Docs**
Docs identified by both, coded responsive	400
Docs identified by the Provider's Tool NOT CAL, coded responsive	400
Docs identified by CAL NOT the Provider's Tool, coded responsive	400
Docs identified by both, coded unresponsive	400
Docs identified by the Provider's Tool NOT CAL, coded unresponsive	400
Docs identified by CAL NOT the Provider's Tool, coded unresponsive	400
Docs identified by both, coded responsive on the Provider's Tool, unresponsive on CAL	400
Docs identified by both, coded unresponsive on the Provider's Tool, responsive on CAL	400
Docs identified by CAL NOT the Provider's Tool, Rel judgment used for training (redundant Not Rel judgment)	400
Docs identified by CAL NOT the Provider's Tool, Not Rel judgment used for training (redundant Rel judgment)	400
Docs identified by neither the Provider's Tool NOR CAL	1600
Recall formula for the Provider's Tool	
Estimated Recall is calculated by the number of relevant documents in the sample for The Provider's Tool / the total number of relevant documents in the sample for both systems + the number of mislabelled relevant documents in the sample + the number relevant documents in the unreviewed population	
Recall formula for CAL®	
Estimated Recall is calculated by the number of relevant documents in the sample for CAL / the total number of relevant documents in the sample for both systems + the number of mislabelled relevant documents in the sample + the number relevant documents in the unreviewed population	

D. Validation Protocols

A major source of contention in the eDiscovery industry is the best method for demonstrating the effectiveness of a given review effort. The most-frequently employed metric used to assess the "completeness" of a review is by calculating an estimate of its recall. How this is done can vary widely, with some approaches yielding more accurate and reliable estimates than others. In practice, as opposed to in the lab, legal practitioners do not have the luxury of comparing the accuracy of their results to a "gold standard" evaluation set using the Cranfield Method [13]. In litigation, this is impossible for many reasons, not the least of which is that the very notion of relevance or responsiveness is subjective. It has been shown that reviewers will disagree with one another, and even with themselves when reviewing the same document set at different times, regardless of their level of expertise [26, 27, 13, 28, 7, 29, 30, 31, 32]. In order to be defensible, therefore, a validation method used to calculate recall should be objective, unbiased, and independent.

The Elusion test is one of the most widely-utilised validation methods for estimating recall in eDiscovery, although it does not meet the criteria set forth above. As previously mentioned, this validation method typically occurs at the end of a review when the "stopping point" has been reached. A random sample of low-scoring documents targeting a particular confidence level and margin of error are taken from the unreviewed document corpus and assessed for relevance (often by the original reviewers who are aware the review task is almost complete and that the sample—referred to as a "null set" is not supposed to contain relevant documents). A typical sample size is 2,395, targeting a margin of error of 2.5% at a confidence level of 95%. The number of relevant documents found in this Elusion set allows for extrapolation across the unreviewed dataset to estimate the likely number of missing relevant documents still not found – known as "Eluded" documents. Therefore, Elusion test recall can easily be calculated by the total number of relevant documents found / the Eluded documents + the total number of relevant documents found.

Many legal practitioners consider the Elusion test the industry standard validation method. The accuracy and reliability of the results it produces are often accepted without question and relied upon by most as the benchmark of completeness [33]. We contend that this test is flawed because it is subject to reviewer bias and easily manipulated to achieve a desirably (but inaccurately) high recall estimate, and because the method lacks objectivity and independence. We advise against its continued use in the eDiscovery sector as a validation method and propose adopting a more statistically robust alternative: the Confusion test.

A Confusion test is a performance measurement for text classification exercises where output can be split into two or more classes (see Table III). The critical difference which renders it more reliable and accurate than an Elusion test is that it looks at predicted and actual results, which incorporate false negatives and false positives into recall calculations, whereas the Elusion test presupposes all documents marked as relevant on the first pass are accurately coded, which is pure fiction. Further, having an independent reviewer perform a blind review of each sub-collection sample of documents without knowing their previous coding decisions greatly reduces reviewer bias and makes the process more objective. Table IV below depicts a standard Confusion matrix table.

E. CAL® and the Provider's Algorithms

LIBLINEAR is an open-source library that uses SVM classifiers in natural language processing tasks such as binary text classification. The aim of SVM classifiers in binary text classification exercises is to construct an optimal hyperplane margin in a high dimensional space that effectively splits the data into two classes, e.g. responsive or non-responsive in eDiscovery tasks [34, 35]. The provider's tool uses a linear kernel and modified libraries [proprietary feature engineering and hyperparameters].

TABLE IV. CONVENTIONAL CONFUSION MATRIX TABLE

		Actual Values	
		positive (1)	**negative (0)**
Predicted Values	**positive (1)**	TP	FP
	negative (0)	FN	TN

The CAL® tool used for testing purposes is the BMI implementation of CAL® used by participants at the TREC 2015 Total Recall Track. All hyperparameters and feature engineering are unchanged. The logistic regression algorithm uses a sigmoid function on training data, mapping input features to a probability of 0 and 1 [36]. In binary classification exercises, this probability predicts the likelihood of an input belonging to a positive class, e.g., (responsive) in eDiscovery tasks.

F. Participants

Participants in the study were qualified barristers with many years of experience performing TAR and manual reviews in Ireland. They were aware that their coding decisions would ultimately be used in legal proceedings. Therefore, they were all properly incentivised to perform to the best of their capabilities. They were paid commercial hourly rates for their efforts.

IV. RESULTS

In this section, we discuss the outcome of our experiment in terms of the key performance metrics referenced above, i.e., recall, precision, review effort, and cost incurred per relevant document found.

A. Precision and Review Effort

CAL® achieved a higher precision for every C-RFP than the provider's tool (see Table V). In most test cases, CAL® also required less review effort than the provider's tool. In C-RFP 4 and 8 instances, however, CAL® achieved higher precision and found more relevant documents than the provider's tool and, as a result, required additional review effort to do so before reaching the "stopping point".

B. Recall

In every completed C-RFP, CAL® yielded higher recall than the provider's tool, substantially so for five out of six C-RFPs according to the Confusion test results, achieving recall >15% better in each instance. The CAL® reviews for C-RFP 2, 6 and 10 were terminated prematurely due to the resource limitations arising from the unanticipated case settlement. Consequently, no recall measure was calculated via a Confusion test for those C-RFPs. The review of C-RFP 8 was completed on both systems, however, the same issues also precluded the completion of a Confusion test. The C-RFP 3 review did not reach the defined "stopping point" for the provider's tool and was terminated for futility. However, because over 25% of the entire corpus was reviewed on both systems for this C-RFP, we calculated recall via a Confusion test for the given review effort.

There is a stark contrast between the recall estimations achieved using the Confusion test and the provider's Elusion test validation method. Table V shows the results of two separate Elusion tests of the provider's tool's results, in comparison to the Confusion test results for both the provider's tool and CAL®. For the C-RFPs 1, 3, 5, 7 and 9, the original assessor's Elusion test recall overestimates the provider's tool's recall by 29%, 20%, 16%, 25% and 36%, respectively, according to the corresponding Confusion tests. The Elusion test results appear entirely ungrounded in truth when compared to the more statistically reliable results achieved by the Confusion test and, in particular, when viewed in light of the fact that CAL® identified thousands of additional responsive documents than the provider's tool (see Table V precision rates) for these C-RFPs. The same logic can also be applied to the incompletely reviewed C-RFPs 2, 6 and 10, which estimate implausibly high recall.

TABLE V. RECALL FOR CAL® AND THE PROVIDER'S TOOL, ACCORDING TO A CONFUSION MATRIX SAMPLE REVIEWED BY AN INDEPENDENT ASSESSOR. RECALL FOR THE PROVIDER'S TOOL, ACCORDING TO AN ELUSION TEST SAMPLE REVIEWED BY THE ORIGINAL ASSESSORS, AND THE SAME ELUSION TEST SAMPLE REVIEWED BY AN INDEPENDENT ASSESSOR.

CAL Review Effort (K-docs)	**C-RFP**	**1**	**2***	**3***	**4**	**5**
		33.2	16	43.2	4.9	4
	C-RFP	**6***	**7**	**8**	**9**	**10***
		5.7	4.7	11	4.4	24
The Provider's Review Effort (K-docs)	**C-RFP**	**1**	**2***	**3***	**4**	**5**
		34.7	44.1	47.6	8.89	2.92
	C-RFP	**6***	**7**	**8**	**9**	**10***
		15.8	9.03	9.81	4.7	23.7
CAL Precision	**C-RFP**	**1**	**2***	**3***	**4**	**5**
		0.58	0.89	0.66	0.3	0.48
	C-RFP	**6***	**7**	**8**	**9**	**10***
		0.94	0.54	0.55	0.55	0.98
The Provider's Precision	**C-RFP**	**1**	**2***	**3***	**4**	**5**
		0.38	0.3	0.41	0.17	0.12
	C-RFP	**6***	**7**	**8**	**9**	**10***
		0.31	0.19	0.11	0.23	0.34
CAL Recall	**C-RFP**	**1**	**2***	**3***	**4**	**5**
		0.81	n/a	0.87	0.56	0.78
	C-RFP	**6***	**7**	**8**	**9**	**10***
		n/a	0.59	n/a	0.93	n/a
The Provider's Recall	**C-RFP**	**1**	**2***	**3***	**4**	**5**
		0.65	n/a	0.69	0.54	0.24
	C-RFP	**6***	**7**	**8**	**9**	**10***
		n/a	0.42	n/a	0.55	n/a
The Provider's Elusion Test Recall	**C-RFP**	**1**	**2***	**3***	**4**	**5**
		0.94	0.92	0.89	0.4	0.4
	C-RFP	**6***	**7**	**8**	**9**	**10***
		0.98	0.67	0.51	0.91	0.89
The Provider's Elusion Test Recall (Independent Review)	**C-RFP**	**1**	**2***	**3***	**4**	**5**
		0.94	0.96	0.98	0.54	0.32
	C-RFP	**6***	**7**	**8**	**9**	**10***
		1	0.93	0.53	0.4	0.15

C. Cost Incurred Per Relevant Document Found

Table VI shows the average cost incurred per relevant document reviewed on CAL® and the provider's tool. CAL® is far more cost-effective than the provider's tool, identifying relevant documents for half the price or less on six C-RFPs. The average cost per relevant document found for CAL® was 44% that of the provider's tool; this is offset somewhat by the fact that CAL® returned more relevant documents in each review.

TABLE VI. AVERAGE COST INCURRED, IN EUROS, TO IDENTIFY ALL RELEVANT DOCUMENTS ON CAL® AND THE PROVIDER'S TOOL, ACCORDING TO THE AVERAGE TIME SPENT REVIEWING ALL DOCUMENTS PER TOPIC.

C-RFP No.	CAL	The Provider
1	€3.37	€5.67
2	€2.75	€4.95
3	€2.93	€4.81
4	€8.15	€11.34
5	€5.49	€20.15
6	€2.97	€6.31
7	€3.91	€10.84
8	€2.78	€9.29
9	€3.81	€8.08
10	€2.25	€5.81

D. Incomplete Reviews

Resource limitations arising from the case settlement prevented the CAL® reviews for C-RFPs 2, 6, and 10 from reaching their designated "stopping point". However, for C-RFPs 2 and 6, CAL® surpassed the number of relevant documents identified by the provider's tool with a third of the review effort. For C-RFP 10, CAL® found three times as many relevant documents for the same review effort (see Table V). Consequently, the final recall for these C-RFPs was not calculated via a Confusion test. For the completed C-RFP 8, CAL® identified five times as many relevant documents than the provider's tool. Unfortunately, due to the case settlement, resources were unavailable to perform a Confusion test on this C-RFP. The provider's TAR review for C-RFP 3 also concluded slightly prematurely due to resource limitations once the matter settled. However, we calculated recall for the given review effort as both systems had comparable review efforts and had reviewed a considerable portion of the dataset. **Note: Reviews that did not reach the defined "stopping point" are marked with asterisks in the results tables.**

V. DISCUSSION

Our experimental results support our hypothesis that the effectiveness and efficiency with which assessors can complete a TAR review is directly impacted by the underlying tool and process it employs. Moreover, we found that the validation method adopted to measure a review's completeness can dramatically affect final recall estimates.

A. CAL® vs The Provider's Tool: Main Findings

Our findings underscore that the order in which documents are presented to reviewers for assessment is an important differentiator between TAR systems. Document review order appears to have a significant effect on a tool's performance.

CAL® aims to find as many relevant documents as possible as soon as possible. By serving the reviewers with the next most-likely-to-be-relevant documents in an unbroken sequence from the outset, CAL® allows the algorithm to learn and develop its model based on high-scoring examples from start to finish. It often and quickly achieves sustained batches of >90% precision until the next most-likely-to-be-relevant documents are exhausted. At this point, marginal precision drops below the predetermined level (i.e., typically 5 to 10%), and the review is concluded. This benefits the end user in numerous ways. The continual provision of high precision batches to reviewers maximises the trade-off between the effort spent reviewing relevant documents and the wasted effort spent looking at non-relevant ones. High precision and minimised review effort reduce the time it takes to complete a review and, therefore, the overall cost of completing a review and identifying relevant documents. This approach also tends to surface the most relevant and important documents first.

Whereas CAL® adopts a greedy approach to identifying relevant documents, maintaining high-precision batches throughout the review, the provider's tool, on the other hand, hedges its bets by implementing a hybrid relevance feedback and uncertainty sampling approach. 70% of the documents that appear in the provider's review queue are the highest ranked, 20% are low-scoring (in the 40-60 range), and 10% are randomly selected. The rationale for such diversity sampling is to improve recall and the classifier performance by allowing the model to learn from all definitions of relevance, thus acting as "index health" documents. We can see from the results in Table V that this does not appear to work. What this process certainly does, however, is decrease the provider's tool's precision and increases review effort considerably by requiring users to examine 30% "index health" documents when it is unlikely that many, if any, of these documents will be relevant. Therefore, precision at the outset of the provider's TAR review will usually be at most 60-65% per batch of 200 documents and will only decline further while still in the early stages of the review. Further, while CAL® and the provider's tool will be resilient to some inconsistent reviewer behaviour, constructing an effective classifier requires reviewers to make consistent coding decisions. Inconsistent coding decisions applied at the outset of a review to conceptually similar documents may impede the learning model's ability to identify likely-to-be-relevant documents, causing relevance rates to stagnate. Therefore, including 30% "index health" documents from the outset throughout the review may contribute to why the provider's tool underperformed regarding the recall it achieved for every C-RFP.

Additionally, the 10% of random documents inserted by the provider's tool further compounds the effort to complete the review because they appear in the review queue regardless of their score or the past assessments made on conceptually similar material. Therefore, should the document corpus consist of numerous similarly non-relevant documents, the assessor may consistently see this subset of documents irrespective of their continued non-relevant classifications. Naturally, this can give rise to substantial duplication of review effort and not only protract the length of time it takes to complete a review but incur additional expense on the user. Conversely, CAL® leverages every relevance assessment to continuously update its model and document ranks to ensure that documents marked as non-relevant will suppress similar documents from appearing in future batches.

Our findings with respect to the precision achieved by the CAL® tool support our abovementioned observations. The final precision for the concluded CAL® C-RFPs 1, 3, 5, 7, 8 and 9 was 0.48 or better, meaning that CAL® identified one relevant document out of every two reviewed (see Table V). C-RFP 4 was an outlier, achieving 0.3 precision, likely due to the ambiguous nature of the material or the poorly constructed

definition of relevance in the C-RFP, or both. A typical characteristic of a CAL® review is to quickly achieve batches containing nearly 100% relevant documents, with the gain curve sharply inclining in the early stages of learning, plateauing at sustained high relevance before sharply dropping off when all likely-to-be-relevant documents are exhausted. Evidence of this is found by looking at the incomplete CAL® reviews for C-RFPs 2, 6 and 10. 89% relevance or above was maintained for these C-RFPs still in the plateau learning phase after reviewing thousands of documents.

Our findings for precision with respect to the provider's tool are quite different from how a well-performing TAR review should look. The low precision scores are symptomatic of the abovementioned issues, with some C-RFPs, 5 and 8, starkly contrasting to their CAL® counterpart, by identifying only one relevant document out of every 10. This indicates a learning model that has overfitted and cannot generalize well to make successful predictions in unknown data. The effect is a review process that never achieves sustained high precision. This is evidenced by comparing Figures 1 and 2, which show the review trajectory for C-RFP 3.

In contrast to CAL®, the learning curve for the provider's tool appears to decline far more gradually, with no definable pattern, resurging and dropping several times (see Figure 2). Towards the later stages, the review queue contains mostly ambiguous documents and the continued 30% "index health" documents. This adversely impacts the learning curve's ability to flatten quickly and signify the end of the review, dramatically decreasing precision and increasing review effort and cost. The provider's tool's decision boundary has failed to reach stabilization, which also has negative implications for recall because it is nearly impossible to determine when the review is complete with any degree of certainty.

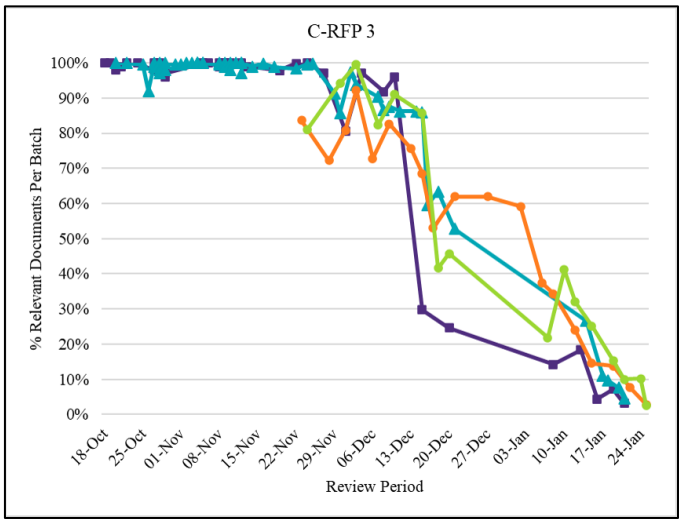

Fig. 1. Relevance rates for each reviewer per batch from the start to conclusion of the CAL® review for C-RFP 3.

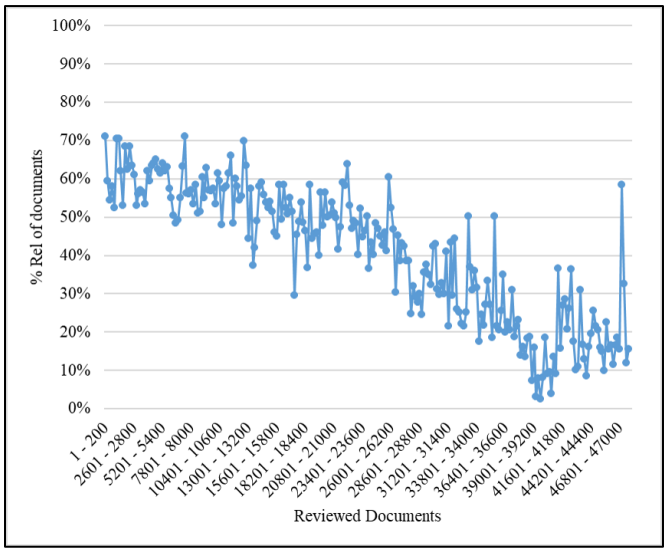

Fig. 2. Consolidated relevance rates for each reviewer per batch from the start to end of the provider's TAR review for C-RFP 3. Note: C-RFP 3 concluded slightly before the "stopping point" due to resource limitations.

The results in Table VI translate the difference in review performance into the costs incurred to complete each C-RFP review. While it is apparent that CAL® is better than the provider's TAR tool at identifying more relevant documents, with better precision, and for less review effort, what needs to be more readily apparent is how these findings also have a direct and significant monetary impact on the user; their chosen TAR system could potentially save or cost them millions of euros. In our current test dataset of 169,000 documents, the difference in price in completing the review is already high, but extrapolated across the entire dataset of 300 million documents, it would have been prohibitively so.

It is the authors' understanding that the leading eDiscovery provider has recently reconfigured its tool to allow for the removal of the 30% "index health" documents. This modification was implemented after our testing programme was complete. The extent to which this could improve the TAR system's performance is unknown, but it will likely decrease review effort and, therefore, result in some improvement in precision. However, it is difficult to predict the level of improvement, if any, that will be seen in the classifier's effectiveness at achieving high recall. It is possible that the provider's modifications could cause the platform's performance in that respect (or in others) to deteriorate to some degree. Only time will tell.

B. Review Validation: Elusion is an Illusion

The notion that there is a "ground truth" for the relevance for every document in a TAR review is illusory. Attaining 100% recall in a TAR review in a legal proceeding is impossible due to the subjective nature inherent in document reviews and the imprecision in the definition and assessment of relevance itself [26, 27, 13, 28, 7, 29, 30, 31, 32]. TAR users should be wary of service providers who report recall estimates closely approaching that number. However, what can be achieved using sound TAR tools and methods is reasonably high recall (i.e., 75%+) that is accurately reported through the proper application of a statistically robust validation protocol.

We demonstrate the influence of validation metrics by comparing the recall estimates provided through an Elusion test versus a Confusion test on the same production. Our results support our call for the immediate discontinuation of the Elusion test as a defensible validation protocol in eDiscovery.

Considering the precision and total number of relevant documents identified for C-RFPs 1, 2, 3, 6, 9 and 10 on the provider's tool versus CAL® (see Table V), the Elusion test recall estimates for these C-RFPs are dramatically overinflated (see Table V).

According to the original assessor's Elusion test reviews, the provider's tool achieves comparable recall for C-RFP 9 and superior recall than CAL® for C-RFPs 1, 3 and 7 (see Table V). However, the Confusion test clearly shows that the provider's tool achieves inferior recall and overestimates the provider's tool's recall by 36%, 29%, 20% and 25%, respectively. According to the Confusion test, the provider's tool did not achieve reasonably high recall (i.e. 75%) in any of the C-RFP reviews! This has potentially serious legal ramifications for users of this TAR system. Instead of achieving high recall as the Elusion test results would lead the producing party to believe, they are actually producing far fewer relevant documents than would likely be deemed adequate under the producing party's statutory obligation under, for example, (U.S.) Federal Rule of Civil Procedure 26(g)(1)(B), which requires attorneys to certify they have conducted a reasonable search to find all reasonably available relevant documents.

Why is it that the Elusion test is so fatally flawed? The answer is fourfold. First, in a voluminous dataset consisting of hundreds of thousands to millions of documents, with low prevalence, examining a relatively small random sample of documents with a confidence level of 95% and a margin of error of +/-21/2 to 3%, is a negligible representation of the entire collection that is unlikely to yield any positive examples, particularly when most of the relevant documents have already been removed. Therefore, the Elusion test predicts almost perfect recall (see C-RFP 6 in Table V), when that may not be the case. Our results show that the provider's TAR tool does not identify nearly as many relevant documents as CAL®. Yet, the Elusion test suggests an adequate result by calculating recall based on the number of relevant documents identified in a relatively small sample of the unreviewed collection and presupposing that all relevant judgments in the original review are infallible, which is an incorrect assumption. When combined, these make a recall estimate look better than it is. When a handful of relevant documents are identified in a low-prevalence Elusion test, this can drastically alter the final recall estimate [see C-RFP 9 in Table V for original and independent assessor divergence].

Second, for high-prevalence datasets, Elusion tests are rendered almost meaningless at accurately estimating recall because large numbers of relevant documents in an Elusion sample set have a negligible impact on reducing what appears as a high-recall estimate.

Third, it is often the case in practice that the same reviewers who conducted the original review also perform the Elusion test review and do so knowing that they are reviewing documents that are "supposed to be" non-responsive. This results in considerable reviewer bias and is subject to manipulation when review teams are acutely aware that the review is "nearly over," so they know they are not supposed to find more relevant documents, thereby achieving an acceptable Elusion recall estimate.

Fourth, and linked to the point above, is the notion of independence and objectivity that should be inherent in any reliable validation protocol. Studies have shown that a final recall estimate is positively impacted if the same assessor is used to train and validate the system performance [29]. One can easily receive full marks when assessing one's own work.

The Confusion test resolves all of these issues by incorporating both previously reviewed and unreviewed documents without providing the validation reviewer information about the prior coding (if any). This blind review of a sample of the entire dataset provides a more statistically robust overview of review performance, both of the TAR tool and the first-pass reviewers. Having an independent reviewer perform this exercise with no prior knowledge of the sub-collection from which the documents came or their previous coding decisions removes all bias and subjectivity. The outcome is a recall estimate that is defensible and accurate.

VI. Conclusion

We tasked our barristers with finding as many relevant documents as possible before marginal precision dropped below our predefined TAR "stopping point" on two systems: Grossman and Cormack's logistic regression CAL® tool and a leading eDiscovery service provider's SVM-based tool. Our results show that the "science behind the black box" may be underappreciated in practice and in the case law. The effectiveness of the learning algorithm used by a TAR system and the accompanying TAR process can substantially impact a system's ability to achieve high recall and high precision for reasonable effort. For all C-RFPs, our results were consistent: CAL® achieved significantly superior recall to the provider's tool according to the completed Confusion tests, while concluding the reviews with a much higher precision for less review effort. In every instance, the provider's tool achieved low to unacceptably low recall estimates by industry standards.

A serious consideration before undertaking any document review, whether by exhaustive manual review or TAR, is the question: How much will this cost? Our results conclusively show that CAL® achieves its superiority over the provider's tool while also doing so for less money. This is particularly important in jurisdictions that provide for cost allocation following proceedings; it becomes highly relevant for legal practitioners to know when the opposing side is using a TAR system that unnecessarily drives up costs by failing to achieve high levels of precision. This becomes increasingly important with larger datasets with low prevalence.

The Elusion test gives the user a misleading picture of the results of a review effort. Recall estimated from an Elusion test paints an overly rosy picture as compared to a proper recall estimate using a Confusion test. In truth, Elusion tests can easily mask the true quality of a production which may be inadequate. The statistical flaws, bias and subjectivity inherent in the Elusion Test make it unsuitable as a validation protocol for eDiscovery.

The authors urge that its use be discontinued in legal practice and offer a readily available alternative in its place, one that is easily implemented without additional effort and that ensures defensible and reliable recall estimates - the Confusion test. Its adoption is a must for any TAR user who wants to certify that they have properly discharged their statutory obligation to make a reasonable production.

ACKNOWLEDGMENT

The authors wish to express their gratitude to the Maples Group for their cooperation and support during this study. A special mention to Martin Elliott for his technical input.

REFERENCES

[1] G. V. Cormack and M. R. Grossman, “Evaluation of machine-learning protocols for technology-assisted review in electronic discovery,” In Proc of. SIGIR '14, Jul. 2014, doi: 10.1145/2600428.2609601.

[2] J. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, V. Welch (Editors), “Cochrane Handbook for Systematic Reviews of Interventions (2nd.ed.)," Chicester (UK): John Wiley & Sons, 2019.

[3] M. Sanderson and H. Joho, “Forming test collections with no system pooling,” In Proc of. SIGIR ’04, Jul. 2004, doi: 10.1145/1008992.1009001.

[4] G. V. Cormack and T. R. Lynam, “Spam Corpus Creation for TREC. In The Second Conference on Email and Anti-Spam," 2005, [Online]. Available at: https://www.ceas.cc/papers-2005/162.pdf

[5] P. Taylor. “Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025,” Available at: https://www.statista.com/statistics/871513/worldwide-data-created/.

[6] 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 Journal of Law and Technology, vol. 17, no. 3, p. 11, Jan. 2011, [Online]. Available at: https://scholarship.richmond.edu/cgi/viewcontent.cgi?article=1344&context=jolt

[7] G. V. Cormack and M. R. Grossman, “Navigating Imprecision in Relevance Assessments on the Road to Total Recall: Roger and Me,” In Proc of. SIGIR '17, Aug. 2017, doi: 10.1145/3077136.3080812.

[8] Da Silva Moore v. Publicis Groupe - 287 F.R.D. 182 (S.D.N.Y. 2012)

[9] C. Goodhart, “Problems of Monetary Management: The UK Experience. In: Monetary Theory and Practice,” 1984, doi 10.1007/978-1-349-17295-5_4.

[10] M. R. Grossman and G. V. Cormack, “The Grossman-Cormack Glossary of Technology-Assisted Review,” Fed. Courts. Law. Review, vol. 7, no. 1, 2013.

[11] G. V. Cormack and M. R. Grossman, “Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review,” Apr. 2015, [Online]. Available at: https://arxiv.org/pdf/1504.06868.pdf

[12] D. Blair and M. E. Maron, “An evaluation of retrieval effectiveness for a full-text document-retrieval system,” Commun. ACM vol. 28, no. 3, pp. 289-299, Mar. 1985, doi: 10.1145/3166.3197.

[13] E. M. Voorhees, “Variations in relevance judgments and the measurement of retrieval effectiveness,” Information Processing and Management, vol. 36, no. 5, pp. 697-716, Sep 2000, doi: 10.1016/s0306-4573(00)00010-8.

[14] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke, “Efficient construction of large test collections,” In Proc of. SIGIR '98, Aug. 1998, doi: 10.1145/290941.291009.

[15] D. Lewis, Y. Yang, T. Rose, and F. Li, “RCV1: a new benchmark collection for text categorization research,” Journal of Machine Learning Research, vol. 5, pp. 361–397, Dec. 2004, doi: 10.5555/1005332.1005345.

[16] D. Lewis and W. A. Gale, “A sequential algorithm for training text classifiers,” arXiv (Cornell University), pp. 3–12, Aug. 1994, doi: 10.5555/188490.188495.

[17] “Baseline Model Implementation for Automatic Participation in TREC 2015 Total Recall Track,” Available at: https://cormack.uwaterloo.ca/trecvm/. [Accessed 29 August 2023]

[18] G. V. Cormack and M. R. Grossman, “Engineering Quality and Reliability in Technology-Assisted Review,” In Proc of. SIGIR '16, Jul. 2016, doi: 10.1145/2911451.2911510.

[19] D. Li and E. Kanoulas, “When to Stop Reviewing in Technology-Assisted Reviews: Sampling from an Adaptive Distribution to Estimate Residual Relevant Documents,” ACM Trans. Inf. Syst. Vol. 38, no. 4, Article 41, pp. 1-36, Oct. 2020, doi: 10.1145/3411755.

[20] E. Yang, D. D. Lewis, and O. Frieder, “Heuristic stopping rules for technology-assisted review,” In Proc of. DocEng ’21, Aug. 2021, doi: 10.1145/3469096.3469873.

[21] A. Sneyd and M. Stevenson, “Stopping Criteria for Technology Assisted Reviews based on Counting Processes,” . In Proc of. SIGIR ’21, Jul. 2021, doi: 10.1145/3404835.3463013.

[22] H. Zhang, J. Lin, G. V. Cormack, and M. D. Smucker, “Sampling Strategies and Active Learning for Volume Estimation,” In Proc of. SIGIR ’16, Jul. 2016, doi: 10.1145/2911451.2914685

[23] G. V. Cormack and M. R. Grossman, “Scalability of Continuous Active Learning for Reliable High-Recall Text Classification,” In Proc of. CIKM ’16, Oct. 2016, doi: 10.1145/2983323.2983776.

[24] G. V. Cormack and M. R. Grossman, “Multi-Faceted Recall of Continuous Active Learning for Technology-Assisted Review,” In Proc of. SIGIR '15, Aug. 2015, doi: 10.1145/2766462.2767771.

[25] In re Broiler Chicken Antitrust Litigation, No. 1:2016cv08637

[26] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilma, “Relevance assessment: are judges exchangeable and does it matter,”. In Proc of. SIGIR '08, Jul 2008, doi: 10.1145/1390334.1390447.

[27] H. L. Roitblat, A. Kershaw, and P. Oot, “Document categorization in legal electronic discovery: computer classification vs. manual review,” Journal of the Association for Information Science and Technology, vol. 61, no. 1, pp. 70–80, Oct. 2009, doi: 10.1002/asi.21233.

[28] W. Webber, D. W. Oard, F. Scholer, and B. Hedin, “Assessor error in stratified evaluation,” In Proc of. CIKM ’10, Oct. 2010, doi: 10.1145/1871437.1871508.

[29] A. Roegiest, G. V. Cormack, C. L.A. Clarke, and M. R. Grossman, “Impact of Surrogate Assessments on High-Recall Retrieval,” In Proc of. SIGIR ’15, Aug. 2015, doi: 10.1145/2766462.2767754.

[30] T. Saracevic, “Relevance: A review of and a framework for the thinking on the notion in information science,” Journal of the American Society for Information Science, vol. 26, no. 6, pp. 321–343, Nov. 1975, doi: 10.1002/asi.4630260604.

[31] T. Saracevic, “Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance,” Journal of the Association for Information Science and Technology, vol. 58, no. 13, pp. 1915–1933, Jan. 2007, doi: 10.1002/asi.20682.

[32] T. Saracevic, “Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behaviour and effects of relevance,” Journal of the Association for Information Science and Technology, vol. 58, no. 13, pp. 2126–2144, Jan. 2007, doi: 10.1002/asi.20681.

[33] “There Is No One Size Fits All Sample Size Appropriate for TAR Validation (Part II),” The ACEDS eDiscovery voice community blog, Mar. 2020, Available at: https://www.jdsupra.com/legalnews/there-is-no-one-size-fits-all-sample-42659/.

[34] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” In Proc of. COLT ’92, Jul 1992, doi: 10.1145/130385.130401.

[35] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin, “LIBLINEAR: A Library for Large Linear Classification,” J. Mach. Learn. Res. Vol. 9, pp. 1871-187, 2008.

[36] D. Cox, “The Regression Analysis of Binary Sequences,” Journal of the Royal Statistical Society: Series B (Methodological), vol. 21, no. 1, p. 238, Jan 1959, doi: 10.1111/j.2517-6161.1959.tb00334.x.