# VoMBaT: A Tool for Visualising Evaluation Measure Behaviour in High-Recall Search Tasks

Wojciech Kusa
wojciech.kusa@tuwien.ac.at
TU Wien
Vienna, Austria

Aldo Lipani
aldo.lipani@ucl.ac.uk
University College London
London, United Kingdom

Petr Knoth
petr.knoth@open.ac.uk
The Open University
Milton Keynes, United Kingdom

Allan Hanbury
allan.hanbury@tuwien.ac.at
TU Wien
Vienna, Austria

## ABSTRACT

The objective of High-Recall Information Retrieval (HRIR) is to retrieve as many relevant documents as possible for a given search topic. One approach to HRIR is Technology-Assisted Review (TAR), which uses information retrieval and machine learning techniques to aid the review of large document collections. TAR systems are commonly used in legal eDiscovery and systematic literature reviews. Successful TAR systems are able to find the majority of relevant documents using the least number of assessments. Commonly used retrospective evaluation assumes that the system achieves a specific, fixed recall level first, and then measures the precision or work saved (e.g., precision at r% recall). This approach can cause problems related to understanding the behaviour of evaluation measures in a fixed recall setting. It is also problematic when estimating time and money savings during technology-assisted reviews.

This paper presents a new visual analytics tool to explore the dynamics of evaluation measures depending on recall level. We implemented 18 evaluation measures based on the confusion matrix terms, both from general IR tasks and specific to TAR. The tool allows for a comparison of the behaviour of these measures in a fixed recall evaluation setting. It can also simulate savings in time and money and a count of manual vs automatic assessments for different datasets depending on the model quality. The tool is open-source, and the demo is available under the following URL: https://vombat.streamlit.app.

## CCS CONCEPTS

• **Information systems** → **Evaluation of retrieval results**; *Retrieval effectiveness*; • **Human-centered computing** → Visual analytics.

## KEYWORDS

visual analytics, technology-assisted reviews, systematic reviews, citation screening, evaluation, evaluation measures, TNR



*SIGIR '23, July 23–27, 2023, Taipei, Taiwan*

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591802

**ACM Reference Format:**
Wojciech Kusa, Aldo Lipani, Petr Knoth, and Allan Hanbury. 2023. VoMBaT: A Tool for Visualising Evaluation Measure Behaviour in High-Recall Search Tasks. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3591802

## 1 INTRODUCTION

High-recall search tasks or High-Recall Information Retrieval (HRIR) are terms describing the objective of locating all or nearly all relevant documents in a collection. One common approach to carry out high-recall search tasks is Technology-Assisted Review (TAR), which leverages information retrieval (IR) and machine learning (ML) techniques to improve the efficiency, accuracy, and consistency of reviewing large volumes of documents. TAR systems aim to support human reviewers by automating repetitive tasks and highlighting the most relevant documents for review, thus helping organisations save time and resources.

Citation screening for systematic literature reviews is one of the most popular TAR tasks [14, 19]. In this task, researchers screen a large number of publications initially identified through a literature search to determine those relevant to the review. When conducted manually, this process is time-consuming and labour-intensive, and involves making thousands of eligibility decisions. Other examples of high-recall search tasks are legal electronic discovery [24], construction of evaluation collections [16], and responses to the freedom of information laws requests [17].

Workshops such as LegalAIIA [4] and ALTARS [8] have popularised HRIR applications among the research community. The TREC Legal [1, 6, 18, 20], TREC Total Recall [10], and the CLEF eHealth Technology-Assisted Review Tasks [11, 12] have provided researchers with access to datasets and standardised evaluation methods.

One of the most critical aspects of HRIR systems is recall, which measures the proportion of relevant documents that are retrieved by the system. Yet, evaluating TAR systems can be challenging due to the interplay between the precision and recall of the models. One of the key goals of TAR systems is to detect as many relevant documents (True Positives, $TP$s) as possible while excluding as many irrelevant documents (True Negatives, $TN$s) as possible. By reducing the number of TNs, TAR systems can save time and

resources for human reviewers. However, caution must be exercised in implementing TAR systems as poor performance could result in legal sanctions, personal liability, and economic costs, as demonstrated in legal discovery scenarios [9].

Several evaluation measures have been proposed to measure the effectiveness of TAR systems [21]. One popular approach is to evaluate the system's precision at a fixed recall level (Precision at r% recall, *Precision@r%*), representing the percentage of relevant retrieved documents [13, 15]. Another approach is to evaluate the work saved compared to the random ordering of the documents, using Work Saved over Sampling at r% recall evaluation measure (*WSS@r%*) [3]. Evaluating TAR systems at a fixed recall level helps to determine the trade-off between precision and recall. It is particularly useful when we assume a minimum acceptable level of recall for the TAR system. Recall versus effort plots using the *knee method* [5] have been used as a more generalised extension, plotting the scores over the full range of values of recall.

The dynamics between the values of true negatives and recall can be difficult to comprehend when using measures such as *WSS@r%* or *Precision@r%*. These measures do not provide a clear understanding of the number of true negatives found by the model and the time saved as a result. Additionally, it can be challenging to translate a particular score of these measures into real time and money benefits. The complexity of these measures can pose a challenge for practitioners in effectively utilising TAR systems and accurately evaluating their performance.

As the number of potential applications of TAR grows, so will the need for better evaluation techniques. To this end, this paper introduces VoMBaT, a new *Visual Analytics* tool for analysing evaluation measure scores in the high-recall setting. Our visual analytics tool for high-recall search task evaluation addresses the current limitations in understanding the behaviour of evaluation measures, especially at different recall levels, by bridging the gap between technical experts and non-experts in the field of HRIR and TAR. Our tool allows researchers and practitioners to understand the behaviour of evaluation measures better, simulate savings in time and money, and compare manual versus automatic assessments for different datasets. By making this tool open-source and freely available, we hope to improve the effectiveness of HRIR and TAR systems, and facilitate the use of visualisation as a means of communicating complex technical information to a broader audience. The toolbox we developed offers an interface to compare different evaluation measures, providing insights into their impact. The tool does not compare scores from actual runs but instead takes only two dataset parameters: dataset size and a percentage of relevant documents in the dataset, making it domain agnostic and applicable for many TAR applications.

The target users for the tool are researchers, practitioners and other stakeholders involved in high-recall search tasks who want to understand the evaluation measures better. These users may include data scientists, machine learning engineers, legal professionals, and academic researchers. Additionally, the tool can be used to help users in their decision-making process about the quality of TAR models and to evaluate the potential savings in time and resources in a variety of settings.

We pre-implemented 18 evaluation measures based on confusion matrix terms, including general evaluation measures like *Precision*, *Accuracy*, *F−score*, and *TNR*. Additionally, the widespread evaluation measures in TAR systems, Work Saved over Sampling (*WSS*), and Depth for Recall (*DFR*) have also been implemented. Furthermore, we introduce a variation of the True Negative Rate, rectified True Negative Rate (*reTNR@r%*), which penalises models that perform worse than a random ordering of the documents.

Before demonstrating the tool, we briefly describe Technology-Assisted Reviews.

## 2 BACKGROUND

All TAR automation models can be coarsely categorised into either classification or ranking approaches [19]. An effective TAR algorithm aims to maximise the number of relevant documents found and save the reviewers' time by removing irrelevant documents.

When TAR is treated as a classification task, measures based on the confusion matrix and the notion of Precision and Recall are commonly used [19, 21]. Aside from Precision and Recall, measures include variations of the harmonised mean between the two, i.e., $F_\beta$−score, Yield, Burden [23], Utility$_\beta$ [22], sensitivity-maximising thresholds [7], and AUC [2]. Another measure, Work Saved over Sampling (*WSS*), measures the amount of work saved when using machine learning models to screen irrelevant publications [3]. The True Negative Rate (*TNR*) was proposed as an alternative as it addresses some of the limitations of *WSS* regarding averaging scores from multiple datasets [15].

When treating the TAR as a ranking task (e.g., for the sub-task of screening prioritisation or stopping prediction), then rank-based measures and measures at a fixed cut-off are commonly used, e.g., *nDCG@n*, *Precision@n*, *Recall@n*, R-Precision [10], and last relevant found. However, retrospectively evaluating models at different levels of recall might better suit the TAR task, because it takes into account the number of relevant documents found and the trade-off between reviewing more documents and potentially finding more relevant ones, versus stopping the review and potentially missing some relevant documents.

Consider an example scenario in which a review contains a collection of $N = 2,000$ documents. Of these, 200 are relevant to the study and should be included in the final review (we call them *includes*, $\mathcal{I}$), while the remaining $1,800$ are irrelevant and should be excluded (we call them *excludes*, $\mathcal{E}$). In manual screening, annotators must review all $2,000$ documents to identify only the 200 relevant ones. Defining a recall level for the assessment of TAR systems assumes that the number of true positive and false negative documents remains constant. For instance, a recall of 95% would be achieved when the model accurately identifies 190 relevant documents (*TP*) and misclassifies the remaining 10, i.e., these are false negatives (*FN*).

The domain and characteristics of the review influence the choice of recall level. Past studies on the automation of citation screening in medicine typically used 95% recall as the threshold to preserve a satisfactory quality of the systematic literature review in medicine [3]. In other technology-assisted review systems, recall levels might be lower, for instance, in e-discovery, a commonly used recall is 80% [25]. Sometimes the choice of recall is influenced by the time or money limitations of the task—that is why understanding how specific evaluation measures behave depending not only on the dataset but also on the recall level is crucial.

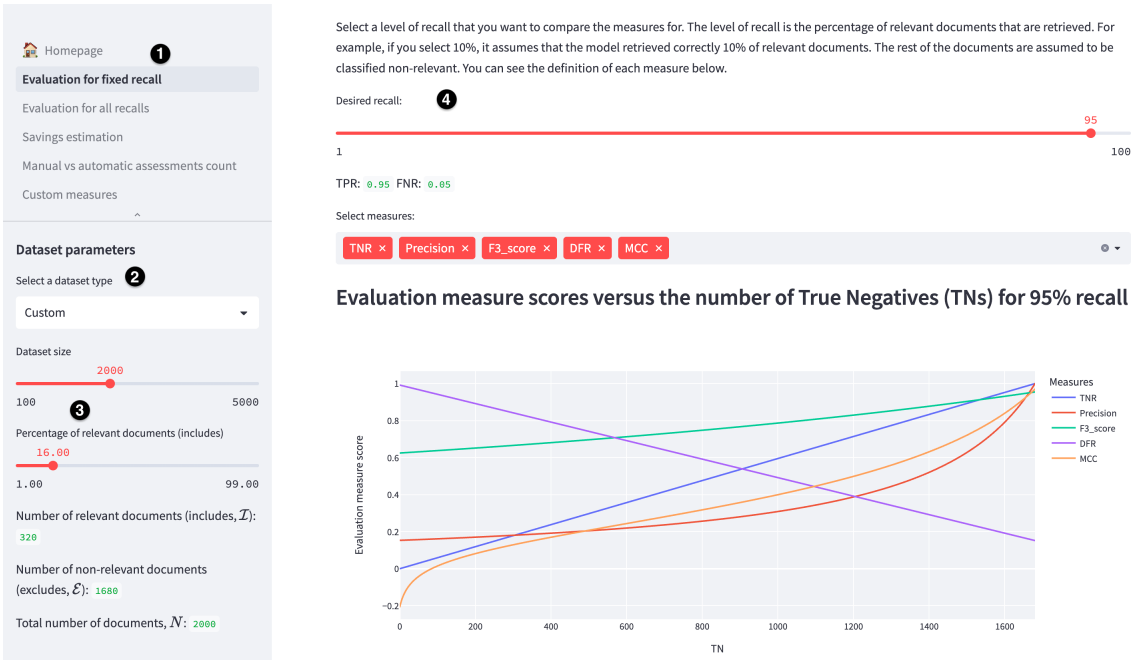


**Figure 1: A screenshot of a web page for comparing evaluation measures at a fixed recall level. Navigation to other web pages and dataset parameters selection are on the left, while recall level and evaluation measure selection are on the top.**

## 3 DEMO

Our toolbox is made using Python 3.10, Plotly and Streamlit and is available as an open-source package[1]. The interface consists of five subpages described in detail in this section. We first define the implemented evaluation measures.

### 3.1 Implemented Evaluation Measures

We implemented twelve evaluation measures based on the confusion matrix terms: *Precion*, *Accuracy*, *Balanced Accuracy*, $F_1 - score$, $F_3 - score$, $F_{0.5} - score$, *TNR*, *MCC* (Matthews correlation coefficient), *FDR* (False discovery rate), *NPV* (Negative predictive value), *FOR* (False omission rate), and *DOR* (Diagnostic odds ratio). Most of these measures have been previously applied to the evaluation of TAR models.

Furthermore, we implemented two measures specific to the evaluation of TAR systems, which can also be calculated for a fixed recall level: Work Saved over Sampling (*WSS*) and Depth for Recall (*DFR*):

$$WSS@r\% = \frac{TN + FN}{N} - (1 - r) \tag{1}$$

$$DFR@r\% = \frac{Prevalence \cdot r}{Precision} \tag{2}$$

The problem with evaluating models at a specific recall value is that most measures will not be bounded between [0, 1], and their minimum and maximum values will depend on the class imbalance. Hence, we also implemented a normalised version of $F_{beta}@r\%$ and $Precision@r\%$:

$$nF_{beta}@r\% = \frac{(r + \beta^2) \cdot T \cdot TN}{F \cdot (r \cdot T + \beta^2 \cdot T + FP)} \tag{3}$$

$$nPrecision@r\% = \frac{TP \cdot TN}{\mathcal{E} \cdot (TP + FP)} \tag{4}$$

Additionally, we introduce a new measure for analysis: rectified True Negative Rate ($reTNR$) and its min-max normalised version ($nreTNR$):

$$reTNR@r\% = \begin{cases} TNR@r\%, & \text{if } \frac{FP@r\%}{\mathcal{E}} < r\% \\ \frac{TN@r\%}{\mathcal{E}}, & \text{otherwise} \end{cases} \tag{5}$$

$$nreTNR@r\% = \frac{reTNR - \min(reTNR)}{\max(reTNR) - \min(reTNR)} \tag{6}$$

$reTNR$ penalise models which perform worse than a random ordering of the documents, i.e., when, for a given $r\%$ of recall, the true negative rate is lower than $(1 - r)$, $reTNR$ score is equal to the $(1 - r)$. This threshold is equal to a simple random sampling, as savings of $(1 - r)\%$ are achieved when on average, $r\%$ of the dataset is randomly sampled. An intuition for this measure is that all models performing worse than random sorting are equally bad, and, especially when averaging scores, they should not influence the actual work saved.

### 3.2 Interface

The user interface consists of five web pages, accessible through a sidebar on the left-hand side of the screen (❶ on Figure 1). A set of predefined datasets' parameters (the total number of documents $N$ and a percentage of relevant documents $\mathcal{I}$) was prepared for each

[1] https://github.com/WojciechKusa/VoMBaT

of these pages ❷. Users can also define custom dataset size and a percentage of relevant documents ❸. There are two types of predefined datasets:

- Three synthetic examples of dataset parameters showing extreme options for the distribution of relevant documents ($\mathcal{I}$) in the dataset: balanced, heavily unbalanced towards positive class (example of a very good search query), and heavily unbalanced towards negative class (very typical in systematic reviews).
- Fifteen datasets which use the $N$ and $\mathcal{I}$ values from systematic reviews in the field of medicine introduced by Cohen et al. [3].

*3.2.1 Evaluation for a fixed recall level.* This web page presents a comparison of evaluation measures for a fixed level of recall (❹ on Figure 1). Users need to select a level of recall to compare the measures first. The level of recall is the percentage of relevant documents that are retrieved. For example, if one selects 10%, it assumes that the model retrieved 10% of relevant documents correctly. The rest of the documents are assumed to be classified as non-relevant.

*3.2.2 Evaluation for all recall levels.* This web page presents 3D plots of possible evaluation measure scores for all recall and TN levels. First, up to four measures from the set of predefined ones can be selected. Each measure is plotted in a separate interactive 3D plot, and the x-axis and y-axis represent the number of TNs and the estimated recall level, respectively. The score of the selected measure is presented on the z-axis.

*3.2.3 Savings estimation.* This web page presents the simulation of time and money savings that can be achieved depending on the value of evaluation measures. Users can use this simulation to determine the minimum threshold for the evaluation measures that can be accepted in order to reduce the manual screening time and the cost of the evaluation. Users can adjust factors such as the average time per document, the number of manual assessments per document, and the cost of annotators. During the manual document review, each document is assessed by annotators, and in the extreme case of systematic reviews, by at least two people. Savings can be achieved when the model's automatic assessments are accurate enough to replace manual checks for certain documents, effectively eliminating True Negatives. The more TNs the model can remove, the greater the potential for cost and time savings.

*3.2.4 Manual vs automatic assessments comparison.* This web page assumes that the number of relevant and irrelevant documents to be reviewed manually or automatically is fixed once the desired recall level is established. The relevant documents included in the automatic assessment are equal to the true positives. In contrast, the remaining relevant documents that need to be reviewed manually are the false negatives. The number of irrelevant documents that will be reviewed automatically or manually depends on the model's TNR score. The higher the TNR score, the more irrelevant documents will be automatically excluded, representing the true negatives. The remaining irrelevant documents will need to be reviewed manually, which are the false positives ($FP$). This page provides a visual representation of the expected number of documents that will be reviewed automatically or manually based on a specified recall level. The values are presented as stacked bar plots for eleven different TNR scores.

*3.2.5 Custom measures.* Finally, we allow users to write and test custom evaluation measures using confusion matrix terms as building blocks. The written equation is converted using reverse polish notation to support basic mathematical operations. The user has the option to select two variables (by default, it is recall and TN, as it is on other pages) which will be plotted for comparison with the evaluation measure. The interface is presented similarly to the one described in Section 3.2.2.

## 3.3 Analysis

Based on the analysis of the plots, it is apparent that there are two main types of evaluation measures relevant to this task. Measures such as $WSS$, $DFR$, and $TNR$ are linearly correlated with the number of $TN$ and $FP$ predicted by the algorithm. On the other hand, $Precision$ (and analogically, $F_\beta$–score) have different, non-linear characteristics. For datasets that consist of a significant number of non-relevant documents, $Precision$ values only start to increase as the number of $TN$ increases (due to the constant value of $TP$ and the $(\mathcal{E} - TN)$ term in the denominator).

This leads us to conclude that $TNR$–style measures would be more directly transferable to cost savings. However, measures focusing on $Precision$ can be more useful when evaluating models for a fully automated final stage of the screening process, for instance during full-text screening in systematic reviews. In this case every document successfully screened by the TAR system is of high importance. This is because, in practice, filtering the last few percent of documents can bring the most significant gains to users, as the remaining, not relevant documents can be easily screened using other techniques. This behaviour can be observed from the analysis of web pages described in Sections 3.2.3 and 3.2.4, where the $TNR$ score grows linearly with decreasing time and cost of conducting the review[2].

# 4 SUMMARY AND FUTURE WORK

This paper introduces an interface to analyse and understand behaviours of evaluation measures used in a high-recall setting. We implemented a dashboard with 18 evaluation measures, focusing on the ones used in technology-assisted review tasks. The interface enables a comparison of how these measures behave depending on specific values of recall and true negatives. Furthermore, for $TNR$, it also provides the estimate of saving when using automatic models and a count of documents that need to be screened automatically versus manually. Our tool helps to increase the understanding of evaluation measures used in high-recall search tasks and especially TAR systems. In the future, we will focus on incorporating active learning measures and comparing scores from actual runs.

# ACKNOWLEDGMENTS

This work was supported by the EU Horizon 2020 ITN/ETN on Domain Specific Systems for Information Extraction and Retrieval – DoSSIER (H2020-EU.1.3.1., ID: 860721).

[2]Under the assumption that every document takes the same amount of time for screening.

# REFERENCES


[1] Jason R Baron, David D Lewis, and Douglas W Oard. 2006. TREC 2006 Legal Track Overview.. In *TREC*. Citeseer.

[2] Aaron M Cohen, Kyle Ambert, and Marian McDonagh. 2010. A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review. In *AMIA annual symposium proceedings*, Vol. 2010. American Medical Informatics Association, 121.

[3] A. M. Cohen, W. R. Hersh, K. Peterson, and Po Yin Yen. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association* 13, 2 (3 2006), 206–219. https://doi.org/10.1197/jamia.M1929

[4] Jack G Conrad and Jeremy Pickens. 2021. Second International Workshop on AI and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2021).. In *ASAIL/LegalAIIA@ ICAIL*. 48.

[5] Gordon V. Cormack and Maura R. Grossman. 2016. Engineering quality and reliability in technology-assisted review. *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (7 2016), 75–84. https://doi.org/10.1145/2911451.2911510

[6] Gordon V Cormack, Maura R Grossman, Bruce Hedin, and Douglas W Oard. 2010. Overview of the TREC 2010 Legal Track.. In *TREC*.

[7] Siddhartha R Dalal, Paul G Shekelle, Susanne Hempel, Sydne J Newberry, Aneesa Motala, and Kanaka D Shetty. 2013. A pilot study using machine learning and domain knowledge to facilitate comparative effectiveness review updating. *Medical Decision Making* 33, 3 (2013), 343–355.

[8] Giorgio Maria Di Nunzio, Evangelos Kanoulas, and Prasenjit Majumder. 2022. Augmented Intelligence in Technology-Assisted Review Systems (ALTARS 2022): Evaluation Metrics and Protocols for EDiscovery and Systematic Review Systems. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II* (Stavanger, Norway). Springer-Verlag, Berlin, Heidelberg, 557–560. https://doi.org/10.1007/978-3-030-99739-7_69

[9] David Dowling. 2020. Tarpits: The Sticky Consequences of Poorly Implementing Technology-Assisted Review. *Berkeley Tech. LJ* 35 (2020), 171.

[10] Maura R Grossman, Gordon V Cormack, and Adam Roegiest. 2016. TREC 2016 Total Recall Track Overview.. In *TREC*.

[11] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2017. CLEF 2017 technologically assisted reviews in empirical medicine overview. *CEUR Workshop Proceedings* 1866 (9 2017), 1–29. https://pureportal.strath.ac.uk/en/publications/clef-2017-technologically-assisted-reviews-in-empirical-medicine-

[12] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2018. CLEF 2018 technologically assisted reviews in empirical medicine overview. *CEUR Workshop Proceedings* 2125 (7 2018). https://pureportal.strath.ac.uk/en/publications/clef-2018-technologically-assisted-reviews-in-empirical-medicine-

[13] Georgios Kontonatsios, Sally Spencer, Peter Matthew, and Ioannis Korkontzelos. 2020. Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews. *Expert Systems with Applications: X* 6 (7 2020), 100030. https://doi.org/10.1016/j.eswax.2020.100030

[14] Wojciech Kusa, Allan Hanbury, and Petr Knoth. 2022. Automation of Citation Screening for Systematic Literature Reviews Using Neural Networks: A Replicability Study. In *Advances in Information Retrieval*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer International Publishing, Cham, 584–598. https://doi.org/10.1007/978-3-030-99736-6_39

[15] Wojciech Kusa, Aldo Lipani, Petr Knoth, and Allan Hanbury. 2023. An Analysis of Work Saved over Sampling in the Evaluation of Automated Citation Screening in Systematic Literature Reviews. *Intelligent Systems with Applications* 18 (2023), 200193. https://doi.org/10.1016/j.iswa.2023.200193

[16] Dawn Lawrie, James Mayfield, Douglas W Oard, and Eugene Yang. 2022. HC4: A new suite of test collections for ad hoc CLIR. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*. Springer, 351–366.

[17] Graham Mcdonald, Craig Macdonald, and Iadh Ounis. 2020. How the accuracy and confidence of sensitivity classification affects digital sensitivity review. *ACM Transactions on Information Systems (TOIS)* 39, 1 (2020), 1–34.

[18] Douglas W Oard, Bruce Hedin, Stephen Tomlinson, and Jason R Baron. 2008. *Overview of the TREC 2008 legal track*. Technical Report. MARYLAND UNIV COLLEGE PARK COLL OF INFORMATION STUDIES.

[19] Alison O'Mara-Eves, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. 2015. Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews* 4, 1 (1 2015), 5. https://doi.org/10.1186/2046-4053-4-5

[20] Stephen Tomlinson, Douglas W Oard, Jason R Baron, and Paul Thompson. 2007. Overview of the TREC 2007 Legal Track.. In *TREC*.

[21] Raymon van Dinter, Bedir Tekinerdogan, and Cagatay Catal. 2021. Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology* 136 (8 2021), 106589. https://doi.org/10.1016/j.infsof.2021.106589

[22] Byron C. Wallace, Kevin Small, Carla E. Brodley, and Thomas A. Trikalinos. 2010. Active learning for biomedical citation screening. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2010), 173–181. https://doi.org/10.1145/1835804.1835829

[23] Byron C Wallace, Thomas A Trikalinos, Joseph Lau, Carla Brodley, and Christopher H Schmid. 2010. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics* 11, 1 (2010), 1–11.

[24] Eugene Yang and David D. Lewis. 2022. TARexp: A Python Framework for Technology-Assisted Review Experiments. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 3256–3261. https://doi.org/10.1145/3477495.3531663

[25] Eugene Yang, Sean MacAvaney, David D. Lewis, and Ophir Frieder. 2022. Goldilocks: Just-Right Tuning of BERT for Technology-Assisted Review. In *Advances in Information Retrieval*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer International Publishing, Cham, 502–517.