Technology Assisted Review Systems: Current and Future Directions

Giorgio Maria Di Nunzio[1]

[1]Department of Information Engineering, University of Padua, Italy, Via Gradenigo 6a, 35131, Padova, Italy

Abstract

Technology-Assisted Review (TAR) systems are becoming indispensable in domains demanding extensive document screening with high precision, notably in eDiscovery and systematic biomedical reviews. Recent advancements in machine learning, particularly the emergence of Large Language Models (LLMs), have expanded the capabilities of TAR systems, enabling them to handle voluminous text data more efficiently and accurately. Despite these strides, significant challenges remain, including the development of effective stopping criteria, availability of high-quality domain-specific datasets, and robust evaluation metrics to ensure reproducibility and defensibility in high-stakes applications. This paper surveys recent trends and emerging methodologies in TAR, with an emphasis on approaches aimed at improving document relevance screening, query generation, and validation protocols across active learning (AL) and reinforcement learning (RL) frameworks. We examine the utilization of LLMs for Boolean query refinement and abstract screening, particularly in enhancing systematic review workflows. Additionally, we discuss the role of specialized datasets and data-driven approaches in addressing the unique requirements of TAR systems in fields like biomedical research and eDiscovery.

Keywords

Technology Assisted Review Systems, eDiscovery, Systematic Reviews

1. Introduction

Technology-Assisted Review (TAR) systems are increasingly essential in domains where extensive document screening and high precision are paramount, such as eDiscovery and systematic review in biomedical research [1, 2, 3, 4]. The advent of advanced machine learning techniques, particularly Large Language Models (LLMs), has expanded the potential of TAR systems, enabling them to handle large volumes of text more efficiently and accurately than before. Despite these advancements, key challenges persist, including the need for reliable stopping rules, high-quality domain-specific datasets, and robust evaluation metrics to ensure reproducibility and defensibility in high-stakes applications.

This paper presents a survey on recent trends and emerging methodologies in TAR, focusing on techniques and frameworks aimed at enhancing document relevance screening, query formation, and validation protocols across various active learning (AL) and reinforcement learning (RL) settings. We review recent innovations in leveraging LLMs for Boolean query training and abstract screening, shedding light on how these models can improve systematic review workflows through more effective query generation and precision in relevance assessment. Additionally, we examine the role of specialized datasets and data-driven approaches, underscoring their importance in addressing TAR challenges specific to biomedical research and eDiscovery.

Beyond the technical advancements in model design and data handling, the survey explores the critical issue of evaluation metrics and reproducibility in TAR systems. By analyzing different metrics, including precision-at-recall and task-aligned model selection, we aim to provide insights into best practices that enhance TAR system reliability. Our discussion extends to established protocols and

3rd Workshop on Augmented Intelligence for Technology-Assisted Review Systems (ALTARS 2024): Evaluation Metrics and Protocols for eDiscovery and Systematic Review Systems. March 28, 2024, Glasgow, UK.

*Corresponding author.

giorgiomaria.dinunzio@unipd.it (G. M. Di Nunzio)

https://www.dei.unipd.it/~dinunzio/ (G. M. Di Nunzio)

0000-0002-0877-7063 (G. M. Di Nunzio)

frameworks that support the implementation of TAR in complex legal environments, where concerns about consistency, privacy, and defensibility are paramount.

Finally, we conduct a comparative analysis of current TAR tools, with attention to feedback mechanisms and dense retrieval models that improve document screening prioritization. By reviewing these tools across varying TAR contexts, we highlight both the strengths and areas for further improvement, contributing to the field's ongoing efforts to refine TAR systems for greater accuracy and applicability.

This survey offers a comprehensive view of the current landscape and future directions for TAR systems, providing a foundation for practitioners and researchers aiming to harness advanced technologies for enhanced document screening efficiency and precision in systematic reviews and eDiscovery.

1.1. Methodology

In this review of the current state of the art of technology assisted review systems, we focus on the last year of research papers presented in international conferences and journals. Given the nature of this work which had the goal to bring the most recent material for a discussion at the third workshop on Augmented Intelligence for Technology Assisted Review Systems (ALTARS 2024) [3], we decided to search on Google Scholar all the paper in 2024 that contained the string "technology assisted review". After filtering work that did not fit our goal and our criteria (international conference and journals), we ended up with 16 research papers.

We would also like to mention a work that is not included in the following sections but that we consider an important element for a broader perspective on the past, present, and future of TAR systems. It is the work presented by [5] offer a foundational survey that synthesizes AI-driven methods for systematic review (SR) automation, covering NLP, ML, and DL applications across SR stages. Their study includes an analysis of 52 works, showing most efforts center on automating screening, with less focus on the search, data extraction, and risk of bias assessment. To enhance this overview, the authors conducted an online survey among SR practitioners and researchers, gathering practical insights on current practices, challenges, and future needs.

The survey results, paired with literature analysis, expose key gaps—such as limited use of large language models (LLMs), lack of multilingual datasets, and insufficient integration of image processing. Saeidmehr et al. propose future directions, including expanding AI to new SR stages, incorporating domain knowledge, developing language-agnostic models, and enhancing data extraction from visual data in scientific publications. Their survey thus serves as a dual resource: equipping researchers with an understanding of AI in SR automation and guiding computer scientists to innovate within this evolving field. This work will serve as our foundation for assessing and advancing the methods and protocols for AI-enhanced SR automation.

2. Current State-of-the-Art

2.1. Large Language Models (LLMs) and Boolean Query Training

In this section, we discuss innovative methods for leveraging LLMs to improve the efficiency of document relevance screening in systematic reviews. Two significant studies contribute to this field, each emphasizing unique aspects of query formation and question-answering frameworks in LLMs.

First, the work by [6] focuses on utilizing LLMs to generate structured Boolean queries that resemble those used by systematic reviewers. Their methodology involves creating a dataset with natural language review descriptions paired with Boolean searches, allowing for the training and evaluation of LLMs like MistralInstruct-7b to map these descriptions effectively. Empirical evaluations and librarian interviews reveal that while the model demonstrates promising results, it is not yet fully ready for integration into review workflows. Medical librarians provided valuable feedback, suggesting further improvements to the dataset, including fields like population and intervention, which could enhance performance. The study emphasizes the importance of continuous engagement with expert librarians, as well as review methodologists and students, to better understand user needs. These interactions

could inform improvements, such as post-processing integration and database compatibility, facilitating simultaneous query refinement and abstract screening.

The other work authored by [7] introduce an LLM-driven question-answering framework specifically designed to automate abstract screening in a zero-shot setting. Their approach represents a pioneering method for transforming selection criteria into binary questions, scoring answers using LLMs, re-ranking based on question and criteria, and ultimately prioritizing studies for review. By applying LLMs to screen abstracts according to predefined inclusion and exclusion criteria, they achieve substantial efficiency improvements in reference prioritization. GPT-3.5, for instance, demonstrates a high capability for managing mixed criteria, accurately processing complex queries to isolate relevant studies. Comparative analyses across different LLMs underscore the potential for model diversity to enhance accuracy, suggesting that each model's strengths could be selectively harnessed to improve screening outcomes.

These studies suggest a promising future for integrating LLMs into systematic review workflows. By generating high-quality queries and facilitating precise abstract screening, these models hold significant potential for improving the speed and accuracy of systematic review processes, though further refinement and stakeholder feedback are necessary to optimize their utility.

2.2. Stopping Rules in Technology-Assisted Review (TAR)

The "Stopping Rules in Technology-Assisted Review (TAR)" section explores innovative approaches for determining when to halt document screening in TAR, addressing both reinforcement learning (RL) and heuristic methods tailored to different active learning settings.

In [8], the authors introduce RLStop, a pioneering TAR stopping rule grounded in reinforcement learning. The RLStop approach is designed to assist reviewers in deciding whether to continue screening documents or halt when a predefined target recall is met. The RL model treats each decision as a sequence, examining ranked documents in batches to assess recall levels collectively. Once a batch's relevance judgments are obtained, the RL agent determines if screening should proceed or stop. Empirical tests demonstrate RLStop's substantial workload reduction in comparison to existing methods, as it effectively identifies stopping points across multiple datasets and recall levels. While this approach shows notable promise, RLStop requires training data from previous relevance screenings and specific recall level configurations, which limits its applicability in settings without such resources. Future research will address these constraints, potentially enhancing RLStop's adaptability across diverse TAR environments.

The paper written by [9] propose SAFE, a conservative stopping heuristic that merges various heuristics to prevent early termination and missed relevant records. SAFE's procedure is composed of four steps: generating a training set, applying active learning, expanding the search with a different model, and quality evaluation. SAFE is designed to balance the risks of continued screening with the probability of missing relevant data, thus supporting reviewers in making data-driven stopping decisions. Notably, SAFE aligns with PRISMA 2020 and Open Science standards to ensure AI transparency and reproducibility. Unlike RL-based approaches, SAFE is suitable for active learning-assisted screening and can adapt to different review types, such as updating systematic reviews. The method is accessible to non-experts, making it a practical tool for researchers across varied fields. However, SAFE may not be universally applicable and is limited to specific datasets and active learning models, calling for further research to generalize its use.

In [10], the authors introduce a stopping strategy called SALτ within the context of Active Learning (AL) workflows in TAR. They tackle the critical "when-to-stop" issue, which balances the need for cost-effective annotation with reaching a desired recall level, by adapting the Saerens–Latinne–Decaestecker (SLD) algorithm to TAR processes. This approach effectively addresses earlier limitations noted in prior works by improving stopping precision, particularly for medium to high recall targets (0.8 and 0.9). However, for higher recall levels, SALτ may halt prematurely, which Molinari et al. address with the refined SALRτ, achieving a more balanced trade-off without significantly escalating annotation costs. Future work on SALτ aims to explore more nuanced methods to prevent recall overestimation and to test its compatibility with additional sampling techniques, potentially enhancing its application across varied TAR and eDiscovery tasks. Integrating SALτ and SALRτ into TAR workflows exemplifies how

strategic stopping decisions can optimize both cost and recall, addressing the critical question of "when to stop" efficiently in large-scale annotation settings.

Finally, [11] examines an alternative target-based method inspired by Cormack and Grossman's "reliability score," which measures the likelihood of reaching sufficient recall within TAR. Their approach involves setting a target recall level and using a reliability score to guide when to stop screening. This algorithm leverages mathematical formulas to gauge performance and allows stopping when all target records are retrieved or resources are exhausted. Their results indicate strong reliability, with both theoretical and practical validations, though future work is needed to refine the algorithm for broader application. This target-based method enhances the precision of stopping rules, providing reviewers with quantifiable stopping criteria based on recall probability, though it is limited by initial assumptions regarding the distribution of relevant records.

These approaches offer varied yet complementary stopping solutions, from RL-based strategies for adaptive decisions to heuristic and reliability-score-based methods that prioritize transparency and performance. In this regard, these papers contribute valuable tools for optimizing document screening in TAR.

2.3. Dataset and Data-Driven Approaches

In this section, we explore specialized datasets for Technology-Assisted Reviews in biomedical research and simulation-based studies focusing on active learning (AL) models for identifying elusive documents in systematic reviews.

In [12], the authors introduce FASS-BSLR, a dataset tailored to support TAR in biomedical systematic literature reviews (SLRs), with a focus on Covid-19. This dataset includes 111 biomedical SLRs, their relevant studies, and Boolean queries derived from PubMed based on titles and keywords. It also incorporates Boolean queries generated by generative language models, including a subset crafted by experts, known as Set-B, to enable efficient searches. The authors also explore the use of ChatGPT for generating Boolean queries (referred to as CGT) and demonstrate its efficacy over traditional and automated search methods. Specifically, CGT significantly outperforms other models on most metrics in FASS-BSLR, except when measured at rank 1000, where manual queries show slight advantages. A distinctive contribution of the paper is its assessment of CGT's performance in scenarios with varying numbers of seed documents; results indicate that even a small number of seed documents substantially enhance CGT's effectiveness. This study offers a comprehensive dataset and an innovative Boolean query generation method, laying a foundation for further TAR research in the biomedical domain.

[13] addresses the challenges of identifying hard-to-find documents within AL frameworks. Their simulation study assesses how different AL models and levels of prior knowledge influence the "Time to Discovery" (TD) for elusive relevant documents. The study highlights that certain models, particularly those combining random forests and support vector machines, perform better at reducing TD for these hard-to-find documents. Conversely, other AL models exhibit variability in finding these challenging papers, with TD fluctuating significantly across simulations, indicating inconsistency in model performance. The findings suggest that documents with higher TD values vary significantly across different AL models, hinting at content-based factors—like title and abstract characteristics—that may affect discoverability. The authors advocate for future research focused on analyzing these content features to understand why certain relevant documents remain challenging to locate, which could inform model improvements and better data preprocessing techniques.

These works provide valuable insights for dataset development and model optimization in TAR, with FASS-BSLR facilitating systematic review searches and AL models potentially addressing challenges in retrieving elusive papers. These approaches underscore the importance of both high-quality, domain-specific datasets and targeted model adjustments to enhance TAR outcomes.

2.4. Evaluation Metrics and Reproducibility

To enhance TAR methodologies, understanding the impact of various evaluation metrics and ensuring reproducibility in research are crucial elements. Different evaluation metrics, as well as validation protocols used in reproducibility studies, directly shape TAR performance outcomes, model reliability, and their defensibility in high-stakes settings like eDiscovery and systematic review screening.

In [14], the authors emphasize the importance of balanced evaluation metrics, particularly introducing normalized Precision at Recall (nP@r%) and a variant, snP@r%. Their work illustrates that while recall is a key measure, nP@r% - the product of precision and true negative rate - provides a more comprehensive metric that aligns with high-recall TAR scenarios. The authors argue that nP@r% and snP@r% are less affected by dataset characteristics, making them more suitable for assessing TAR system performance, particularly where both precision and the exclusion of false positives are critical. This robustness allows for more accurate benchmarking of IR models and supports more efficient decision-making for document screening, especially in cases where incorrect retrieval could have legal or financial implications.

On the reproducibility front, [15] explores the challenges in replicating results across different TAR tasks, finding that while a "Goldilocks epoch" for pretraining BERT-based models often enhances TAR outcomes, the exact optimal epoch can vary significantly depending on the dataset and pre-processing. In the study, the authors underscore the limitations of "domain-mismatch" between general pre-trained models (like BERT) and specific tasks, and highlights BioLinkBERT as a domain-optimized alternative that performs well without further pre-training, suggesting that simpler, task-aligned models can effectively enhance reproducibility and reduce computational demands. This study supports a transparent, domain-sensitive model selection as a more effective approach than exhaustive model tuning, especially for medical and scientific literature reviews where reproducibility is essential for reliable results.

Finally, [16] examines the effectiveness of proprietary TAR tools and their associated validation protocols, particularly focusing on the Elusion test, a common method for estimating recall. Their study, conducted with live litigation data, reveals that certain vendor software can be less effective than open-source algorithms like Continuous Active Learning in real-world applications. Additionally, they critique the Elusion test for its tendency to yield inconsistent recall results, potentially undermining defensibility in legal contexts. This highlights the need for reproducible, transparent validation practices to ensure reliability and prevent overestimation of recall in professional eDiscovery.

The papers presented in this section show the importance of tailored evaluation metrics, reproducible model pipelines, and reliable validation practices for robust TAR system performance. By focusing on metrics like nP@r%, task-aligned model selection, and defensible validation protocols, the TAR field can better support accuracy and reliability in real-world applications, from biomedical research to litigation.

2.5. Protocols and Frameworks for TAR Implementation

In this section, we present studies that cover established protocols in TAR, especially those used in e-discovery, focusing on categorization, validation methods, and challenges in legal contexts.

[17] introduce an error-tolerant TAR protocol that incorporates accountability mechanisms by designing a label-verification process within the Continuous Active Learning (CAL) framework. Their work addresses challenges within non-realizable classification cases, where no perfect classifier exists, by accommodating potential human errors in labeling and detecting strategic mislabeling attempts by parties in litigation. This approach adapts secure multi-party computation concepts, aiming to balance high recall with minimized disclosure of non-responsive documents, which is especially important for preserving privacy and confidentiality. Notably, this label-verification protocol detects errors with high probability and provides a way to counteract attempts at misclassification, enhancing both accuracy and accountability in eDiscovery settings.

In [18], the authors investigate categorization efficiency within large-scale eDiscovery, focusing on the trade-off between using numerous specific Requests for Production (RFPs) versus broader

composite categories (CCs). Conducted within a complex litigation case, their study reveals that while more granular RFPs aim for precision, they can reduce reviewer speed and consistency, creating inefficiencies in document categorization. Reviewers showed variability in assigning documents to categories, suggesting that automated categorization may offer a more reliable alternative. This research underscores the need for balance in categorization strategies, as excessive specificity may hinder consistent review outcomes.

The authors in [19] propose an enhancement in systematic review protocols, advocating for a "spiral approach" that leverages machine learning more effectively in staged review processes. Traditionally, systematic reviews follow the PRISMA two-stage protocol, beginning with title/abstract screening before progressing to full-text retrieval. However, the authors suggest that integrating full-text retrieval earlier enables machine learning to operate on more comprehensive data sooner, enhancing the efficiency and precision of reviews. They also recommend that machine learning classifiers in systematic reviews favor logistic regression, with TF-IDF for feature extraction and probability-based query prioritization. This approach aims to address the limitations of sequential filtering and optimize TAR workflows by incorporating full-text screening earlier, thus maximizing the utility of machine learning without increasing workload. The integration of a protocol that maintains contextualization while improving efficiency aligns with the goal of optimizing TAR performance, particularly when large data collections are involved.

These studies highlight the nuances of implementing TAR protocols in complex legal environments, where balancing recall, privacy, and consistency is critical to maintaining defensibility and efficiency in document review.

2.6. Comparison and Performance Analysis of TAR Tools

In this section, we address the variability in performance across TAR tools, highlighting the role of feedback mechanisms and dense retrieval models in improving screening prioritization for systematic review automation.

[20] underscores the challenge of achieving reliable recall and precision when transitioning from paper-based to electronic document reviews. They propose that TAR methods should strive to meet the hypothetical standard of a "reasonable reviewer," with recall and precision evaluated relative to this benchmark. The authors note significant variance in recall and precision estimates across different review teams, attributing discrepancies to low prevalence rates and the random ordering of documents.

[21] investigates dense retrieval models, specifically BERT-based retrievers, for TAR screening prioritization. Unlike traditional methods, this approach benefits from iterative, explicit relevance feedback from reviewers without requiring frequent re-fine-tuning, thus improving efficiency. Their results indicate that dense retrieval with feedback can match or exceed specialized prioritization methods in effectiveness, offering a more efficient alternative by eliminating the need for constant retraining.

3. Conclusions

In this survey, we have explored recent advancements and critical areas in TAR systems, particularly focusing on methods for optimizing document screening in medical systematic reviews and eDiscovery. Through an analysis of emerging technologies, including Large Language Models (LLMs), stopping rule mechanisms, data-driven approaches, and robust evaluation metrics, we have outlined the multifaceted nature of TAR research and development.

Specifically, we reviewed the role of LLMs in Boolean query training and relevance screening, where recent studies show the promise of these models in generating high-quality queries and accurately filtering abstracts. On the other hand, the research on stopping rules demonstrated the importance of both reinforcement learning (RL) and heuristic methods in determining optimal points for halting document review, which remains a critical efficiency factor in TAR workflows.

We further explored the significance of evaluation metrics and reproducibility in TAR systems, emphasizing metrics and defensible validation protocols that can uphold accuracy and reliability.

Lastly, we examined protocols for implementing TAR in legal contexts and conducted a comparative performance analysis of popular TAR tools, revealing the role of feedback and dense retrieval models in enhancing screening prioritization.

The analysis has highlighted several directions for future research in this field. For example, while LLMs have shown potential in query formation and relevance screening, future studies should explore ways to refine these models with domain-specific knowledge bases. This would further improve the precision and contextual relevance of document screening in specialized areas such as legal and biomedical research. In addition, developing adaptive, context-sensitive stopping criteria will be crucial for maximizing TAR efficiency without compromising accuracy. This might involve hybrid models that combine heuristic transparency with RL adaptability, allowing more flexible application across diverse TAR settings. Of course, the optimization of the models as well as the evaluation of the heuristics need the creation of standardized, high-quality datasets tailored to TAR applications remains a critical need. Openly accessible and annotated datasets would greatly facilitate model benchmarking, training, and reproducibility, while also enabling fair performance comparisons across different TAR approaches. Ensuring that TAR models perform consistently across different domains and use cases requires the development of more granular evaluation metrics and reproducibility protocols. This includes creating metrics that reflect both task-specific requirements and broader model reliability, as well as validation techniques that can standardize the TAR evaluation process.

References

[1] G. M. Di Nunzio, E. Kanoulas, P. Majumder, Augmented Intelligence in Technology-Assisted Review Systems (ALTARS 2022): Evaluation Metrics and Protocols for eDiscovery and Systematic Review Systems, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2022, pp. 557–560. doi:10.1007/978-3-030-99739-7_69.

[2] G. M. Di Nunzio, E. Kanoulas, P. Majumder, 2nd Workshop on Augmented Intelligence in Technology-Assisted Review Systems (ALTARS 2023), in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 384–387. doi:10.1007/978-3-031-28241-6_41.

[3] G. M. Di Nunzio, E. Kanoulas, P. Majumder, Third Workshop on Augmented Intelligence in Technology-Assisted Review Systems (ALTARS 2024), in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 432–436. doi:10.1007/978-3-031-56069-9_59.

[4] G. M. Di Nunzio, E. Kanoulas, Special issue on technology assisted review systems, Intelligent Systems with Applications 20 (2023) 200260. URL: https://www.sciencedirect.com/science/article/pii/S2667305323000856. doi:10.1016/j.iswa.2023.200260.

[5] A. Saeidmehr, P. D. G. Steel, F. F. Samavati, Systematic review using a spiral approach with machine learning, Systematic Reviews 13 (2024) 32. URL: https://doi.org/10.1186/s13643-023-02421-z. doi:10.1186/s13643-023-02421-z.

[6] G. P. Adam, J. DeYoung, A. Paul, I. J. Saldanha, E. M. Balk, T. A. Trikalinos, B. C. Wallace, Literature search sandbox: a large language model that generates search queries for systematic reviews, JAMIA Open 7 (2024) ooae098. URL: https://doi.org/10.1093/jamiaopen/ooae098. doi:10.1093/jamiaopen/ooae098.

[7] O. Akinseloyin, X. Jiang, V. Palade, A question-answering framework for automated abstract screening using large language models, Journal of the American Medical Informatics Association 31 (2024) 1939–1952. URL: https://doi.org/10.1093/jamia/ocae166. doi:10.1093/jamia/ocae166.

[8] R. Bin-Hezam, M. Stevenson, RLStop: A Reinforcement Learning Stopping Method for TAR, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY,

USA, 2024, pp. 2604–2608. URL: https://dl.acm.org/doi/10.1145/3626772.3657911. doi:10.1145/3626772.3657911.

[9] J. Boetje, R. van de Schoot, The SAFE procedure: a practical stopping heuristic for active learning-based screening in systematic reviews and meta-analyses, Systematic Reviews 13 (2024) 81. URL: https://doi.org/10.1186/s13643-024-02502-7. doi:10.1186/s13643-024-02502-7.

[10] A. Molinari, A. Esuli, SALτ: efficiently stopping TAR by improving priors estimates, Data Mining and Knowledge Discovery 38 (2024) 535–568. URL: https://doi.org/10.1007/s10618-023-00961-5. doi:10.1007/s10618-023-00961-5.

[11] Z. Hou, E. Tipton, Enhancing recall in automated record screening: A resampling algorithm, Research Synthesis Methods 15 (2024) 372–383. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1690. doi:10.1002/jrsm.1690, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jrsm.1690.

[12] L. Budau, F. Ensan, Fully Automated Scholarly Search for Biomedical Systematic Literature Reviews, IEEE Access 12 (2024) 83764–83773. URL: https://ieeexplore.ieee.org/abstract/document/10539102. doi:10.1109/ACCESS.2024.3405529, conference Name: IEEE Access.

[13] F. Byrne, L. Hofstee, J. Teijema, J. De Bruin, R. van de Schoot, Impact of Active learning model and prior knowledge on discovery time of elusive relevant papers: a simulation study, Systematic Reviews 13 (2024) 175. URL: https://doi.org/10.1186/s13643-024-02587-0. doi:10.1186/s13643-024-02587-0.

[14] W. Kusa, G. Peikos, M. Staudinger, A. Lipani, A. Hanbury, Normalised Precision at Fixed Recall for Evaluating TAR, in: Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 43–49. URL: https://dl.acm.org/doi/10.1145/3664190.3672532. doi:10.1145/3664190.3672532.

[15] X. Mao, B. Koopman, G. Zuccon, A Reproducibility Study of Goldilocks: Just-Right Tuning of BERT for TAR, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 132–146. doi:10.1007/978-3-031-56066-8_13.

[16] T. O'Halloran, B. McManus, A. Harbison, M. R. Grossman, G. V. Cormack, Comparison of Tools and Methods for Technology-Assisted Review, in: S. Li (Ed.), Information Management, Springer Nature Switzerland, Cham, 2024, pp. 106–126. doi:10.1007/978-3-031-64359-0_9.

[17] J. Dong, J. D. Hartline, L. Shan, A. Vijayaraghavan, Error-Tolerant E-Discovery Protocols, in: Proceedings of the Symposium on Computer Science and Law, CSLAW '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 24–35. URL: https://dl.acm.org/doi/10.1145/3614407.3643703. doi:10.1145/3614407.3643703.

[18] B. McManus, T. O'Halloran, A. Harbison, M. R. Grossman, G. V. Cormack, Limitations of the Utility of Categorization in eDiscovery Review Efforts, in: S. Li (Ed.), Information Management, Springer Nature Switzerland, Cham, 2024, pp. 301–311. doi:10.1007/978-3-031-64359-0_24.

[19] E. Yang, Contextualization with SPLADE for High Recall Retrieval, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 2337–2341. URL: https://dl.acm.org/doi/10.1145/3626772.3657919. doi:10.1145/3626772.3657919.

[20] G. V. Cormack, M. R. Grossman, A. Harbison, T. O'Halloran, B. McManus, Unbiased Validation of Technology-Assisted Review for eDiscovery, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 2677–2681. URL: https://dl.acm.org/doi/10.1145/3626772.3657903. doi:10.1145/3626772.3657903.

[21] X. Mao, S. Zhuang, B. Koopman, G. Zuccon, Dense Retrieval with Continuous Explicit Feedback for Systematic Review Screening Prioritisation, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 2357–2362. URL: https://dl.acm.org/doi/10.1145/3626772.3657921. doi:10.1145/3626772.3657921.