DEMONSTRATION

RelEx: An XAI-Enhanced Relevance Feedback Model for User-Adaptive Explanations

SAYANTAN POLLEY, Otto von Guericke University Magdeburg, Magdeburg, Sachsen-Anhalt, Germany

GOVIND SHUKLA, Otto von Guericke University Magdeburg, Magdeburg, Sachsen-Anhalt, Germany

PRITHA GHOSAL, Otto von Guericke University Magdeburg, Magdeburg, Sachsen-Anhalt, Germany

A. NÜRNBERGER, Otto von Guericke University Magdeburg, Magdeburg, Sachsen-Anhalt, Germany

Published: 13 July 2025

SIGIR '25: The 48th International ACM SIGIR Conference on Research and Development in Information Retrieval
July 13 - 18, 2025
Padua, Italy

Conference Sponsors:
SIGIR

https://doi.org/10.1145/3726302.3730132
ISBN: 9798400715921

RelEx: An XAI-Enhanced Relevance Feedback Model for User-Adaptive Explanations

Sayantan Polley
Otto-von-Guericke University
Magdeburg, Germany
sayantan.polley@ovgu.de

Govind Shukla
Otto-von-Guericke University
Magdeburg, Germany
govind.shukla@ovgu.de

Pritha Ghosal
Otto-von-Guericke University
Magdeburg, Germany
pritha.ghosal@ovgu.de

Andreas Nürnberger
Otto-von-Guericke University
Magdeburg, Germany
andreas.nuernberger@ovgu.de

Abstract

The rise of Gen-AI and LLMs often makes it difficult for users to trust retrieved results. The risk of IR systems using LLMs and being susceptible to misinformation can be tackled under the lens of explainability in AI. The topic of explainability in AI, machine learning (XAI) and information retrieval (IR) has been explored through various methods, yet few incorporate user feedback to adapt explanations. In this work, we present an XAI-driven extension to the classic relevance feedback model in IR, incorporating user feedback in the process of explaining the model behavior to the user. Our proposed model, RelEx, introduces XAI-specific elements, including key phrase vectors, text summaries, and contextual phrases combined with a neural ranker. RelEx interactively gathers user feedback, adapting search results based on the modified query and contextual vectors. We further introduce a novel additive similarity scheme that combines document similarity with key-phrase overlap. Retrieval performance is empirically evaluated on multiple benchmark datasets. In the absence of ground truth explanations, we assess explainability and assessability via user studies, where RelEx exhibit promising results.

CCS Concepts

• **Information systems → Users and interactive retrieval**.

Keywords

Explainable Information Retrieval, ExIR, XAI, Relevance Feedback

ACM Reference Format:
Sayantan Polley, Govind Shukla, Pritha Ghosal, and Andreas Nürnberger. 2025. RelEx: An XAI-Enhanced Relevance Feedback Model for User-Adaptive Explanations. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3726302.3730132

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730132

1 Introduction

One of the major goals of explainable AI (XAI) [11] is to help non-AI experts understand how AI systems make decisions so that users can trust the decisions of "black box" AI models such as retrieval or search systems. In the context of explainable information retrieval [2, 24, 31], XAI typically focuses on explaining how the retrieved results are similar to the user's query. The fundamental research question is - can we explain to a non-AI expert user, how documents retrieved from the database are relevant or similar to the user query? There have been multiple works in the explainable information retrieval (XIR) community [21, 31, 32] that attempted to explain the notion of similarity of text, and thereby explain relevance rankings to users. However, few systems[9] could effectively incorporate user feedback and explain how the feedback was used in search.

In this work, we incorporate user feedback as part of explanation generation and propose, RelEx or Relevance Explainer (Fig.1) - a novel XAI-specific extension of the classic relevance feedback model (the Rocchio algorithm) [12, 13]. We argue that key phrases [30] and summary embeddings [8], contextual words - when adapted to generate explanations can support the user better. To facilitate replication and reuse of our work, our code is available [23] with a demo-video [22]. Our work makes the following contributions.

- Incorporate user feedback and adapt search explanations via re-ranking and XAI elements such as key-phrase vectors, contextual words and summaries, using open-source tools/LLMs.
- We investigate how XAI elements such as key phrases in comparison to the summary of a document help users select relevant documents as part of Relevance Feedback.
- We qualitatively evaluate specific XAI aspects, such as assessability and explainability via user studies, by adapting the existing XAI community definitions. We quantitatively evaluate retrieval performance on multiple public datasets.

2 Background and Related work

Relevance feedback is a technique designed to improve search result relevance by incorporating user preferences. By adapting to user needs, it enhances precision and relevance in information retrieval. The process begins with an initial set of results from a user's query, after which the user marks documents as relevant or non-relevant.

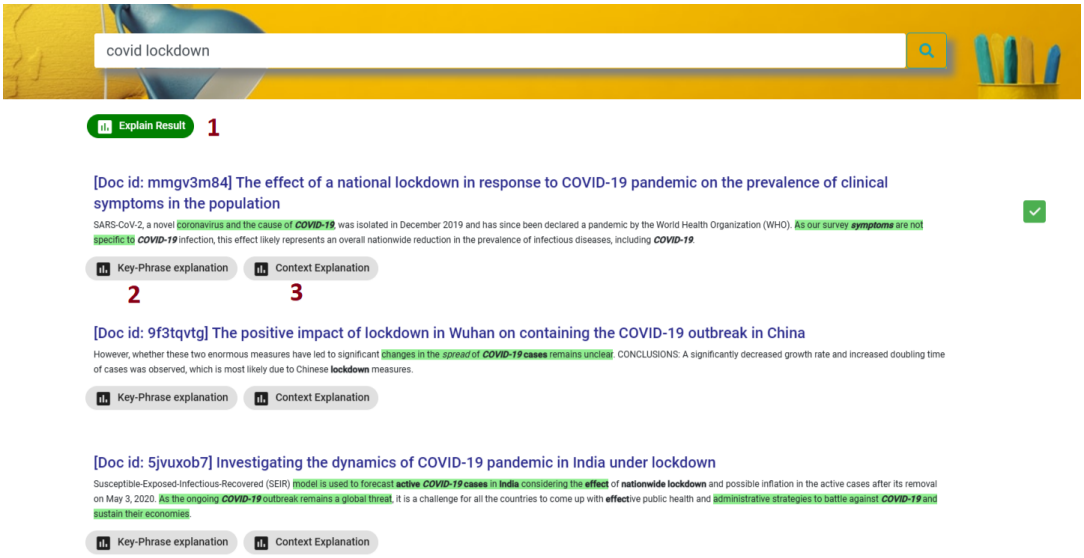

Figure 1: This is the RelEx SUI after user feedback, showing results of extended Relevance Feedback for the query 'covid lockdown'. The component marked by '1' explains the Rocchio classification visually with original and modified query. Component '2' referred to as 'Key-phrase explanation' (see Fig: 3) shows the matching key-phrase with 'chosen relevant document'. Component '3' highlights the contextual words as contextual explanation.

The system then refines the query representation based on this feedback, displaying an updated result set. The Rocchio algorithm is a widely used implementation of relevance feedback, enhancing recall and precision, with several extensions [1, 13] to the original algorithm [29]. The Rocchio algorithm modifies the query vector by adding a weighted combination of the vectors of the relevant and irrelevant documents to the original query vector. The weights used for the relevant and irrelevant documents are determined by three factors: alpha, beta, and gamma. The equation of modified query vector $\vec{Q}_m$ is given by

$$\vec{Q}_m = \alpha\vec{Q}_0 + \frac{\beta}{|D_r|}\sum_{\vec{D}_j \in D_r}\vec{D}_j - \frac{\gamma}{|D_{nr}|}\sum_{\vec{D}_k \in D_{nr}}\vec{D}_k \qquad (1)$$

where $\vec{Q}_0$ is the original query vector D_r and D_{nr} are sets of vectors containing coordinates of relevant and non-relevant documents. Generally, the hyper-parameters α is set to 1, β to 0.8 and γ to 0.3.

A categorization of explainability methods in IR is here [7], when dealing with fairness in ranking systems. The categorization is based on data pre-processing [25], ranker in-processing [6], and result list post-processing [24, 26, 32] methods. Works such as Explainable Search (EXS) [32] extends LIME [28] in the context of text-search, to display words that explain search as a posthoc analysis. Recent systematic surveys of explainable information retrieval can be found here [2, 5]. Mi and Jiang et al. [18] highlighted the importance of explainable approaches, noting that low transparency prevents users from understanding retrieved results, and low assessability limits their ability to judge a result's usefulness, reducing click rates. They used key-phrase matching and document summaries to measure attractiveness, concluding that a search result summary is crucial for explaining result relevance and influencing click decisions. Building on this, Chios and Verberne et al. [8] addressed explainability at both the query and document levels, using Deep Relevance Matching Model (DRMM) to create an interface where query term relevance is visualized.

3 Method and Materials

This section presents the two benchmark datasets and approach of our work (see Fig. 2 for the entire pipeline). Firstly, we used the TREC-COVID-19 dataset [33], a collection of scholarly articles and research documents with relevance judgments for specific queries. From 192,509 records, we removed 400 duplicates and 54,621 records with null values, retaining document-number, title, and abstract for analysis. Besides TREC-COVID, we used about 200,000 documents from Robust04 dataset. We use an XAI search baseline for Explainable Search system (EXS) [32] that uses the popular LIME [28] method to compare explanations with RelEx.

Summarizing, Key-phrasification and Context-window We summarized the abstract text into three sentences using various summarizers such as LSASummarizer [10] from the Sumy Python library (Latent Semantic Analysis - LSA) along with popular abstractive LLM summaries from huggingface (AutoModelForSeq2SeqLM). We then extracted key phrases [30], ranging from one to four words, using the TopicRank algorithm on the 'abstract' field. This method identifies noun-adjective sequences, clusters them into topics, and ranks them via a random walk to produce the final set of key phrases. For each key phrase, we extracted context [3] by capturing a window (see colored or shaded tokens in Fig. 1) of five words before and after the phrase, stored in a separate field. These contexts are later used in the re-ranking with relevance feedback phase.

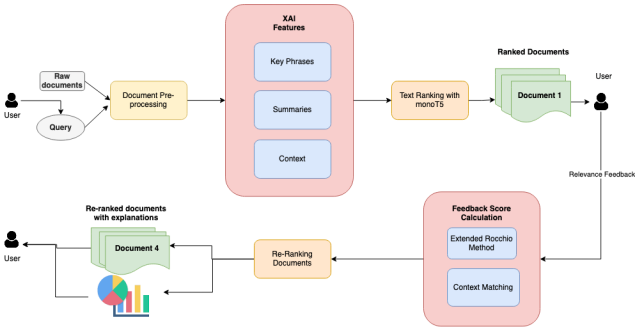

Figure 2: The architecture used for our experiments.

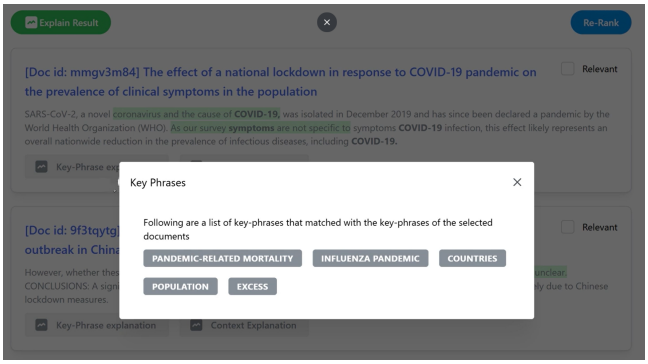

Figure 3: Key-phrase Explanations of a sample text

Text Ranking with User-Feedback: For selecting a neural ranker with SOTA performance, we evaluated two popular models, text-to-text transformers, MonoT5 [20] and DuoT5 [27], against a baseline model BM25. For our use case, we used PyTerrier Python transformer plugins [16] with our preprocessed TREC-COVID-19 dataset with relevance judgments for predefined queries. We performed experiments by building the PyTerrier pipeline [17] and running the PyTerrier experiments API [17]. We used metrics such as MAP, Reciprocal Rank, NDCG at 10, to evaluate the models, and selected the best-performing model. MonoT5 outperformed BM25 and DouT5 in all metrics across both TREC-COVID and Robust04. Table 1 displays the results for TREC-COVID-19 which was used in the user studies for evaluation of the explanations.

Table 1: Experiments with different IR models

Model	MAP	Recip. Rank	P10	NDCG at 10
BM25	0.076117	0.807094	0.678	0.601847
MonoT5	0.081632	0.869000	0.740	0.676241
DuoT5	0.015044	0.813000	0.720	0.646023

A novel XAI extension of the Rocchio Algorithm: We propose a novel XAI-based extension to the Rocchio algorithm for processing user feedback on 50 records, classified as relevant (top 10 marked by the user), irrelevant (top 10 marked irrelevant) and neutral (the remaining 40). The extended Rocchio algorithm (Eqn. 2), re-ranks using the extended query vector over selected relevant and irrelevant documents. We converted the query, abstracts, and key phrases into vectors using Doc2Vec [14] for the representation of abstract text. We modified the Rocchio equation with XAI components, a weighted sum of key phrase vectors from relevant and irrelevant documents to refine the query. Hence, the popular Rocchio as per 1 is modified with XAI flavor as -

$$\vec{Q}_m = \alpha\vec{Q}_0 + \frac{\beta}{|D_r|}\sum_{\vec{D}_j \in D_r}\vec{D}_j - \frac{\gamma}{|D_{nr}|}\sum_{\vec{D}_k \in D_{nr}}\vec{D}_k + \frac{\delta}{|P_r|}\sum_{\vec{P}_j \in K_r}\vec{P}_j - \frac{\eta}{|P_{nr}|}\sum_{\vec{P}_k \in P_{nr}}\vec{P}_k \quad (2)$$

Here, P_r and P_{nr} represent vectors of relevant and non-relevant key phrases, respectively, with hyper-parameters δ and η analogous to β and γ. The enhanced query vector is used to compute cosine similarity between unmarked and user-selected relevant documents. We aggregate key-phrase occurrences, storing them as a key-phrase context match count. Finally, each document's feedback score is calculated as a linear combination of normalized cosine similarity and normalized key-phrase context match count as below.

$$feedbackScore = a \cdot \frac{CoSim}{\| CoSim \|} + b \cdot \frac{Key - PhrCnt}{\| Key - PhrCnt \|}$$

The parameters ranging from 0 to 1, control the weighting of cosine, and key-phrase similarity in re-ranking, which were determined empirically (a = 0.3 and b= 0.7).

4 Results and Discussion

The evaluation was focused on two objectives: first, quantitatively assessing retrieval performance using standard metrics like Mean Average Precision (MAP) and NDCG to select the ranker; Secondly, as ground truth explanations are unavailable [19], we aligned our user study on qualitative XAI evaluation focused on user-mental models [4] to evaluate the explanations.

User Study Goal and Design: The study involved 18 participants (aged 22–32), primarily bachelor's and master's students, conducting COVID-19-related searches and rating outcomes on a 7-point Likert scale. Initially, 30 participants were recruited without any payment, but 50% with IR backgrounds were removed to minimize bias. A within-participant design with Latin block ordering was used, alternating two search UI interfaces to reduce ordering bias [15]. Users were asked to assess **explainability** ("The explanations help me understand result ranking based on my feedback" [18]) and **assessability** ("I can identify relevant results without opening documents based on the feedback page" [8]). We developed two comparable interfaces: one with explanation features and one without (See Fig.4). We investigate various research questions (RQs) in this context as follows.

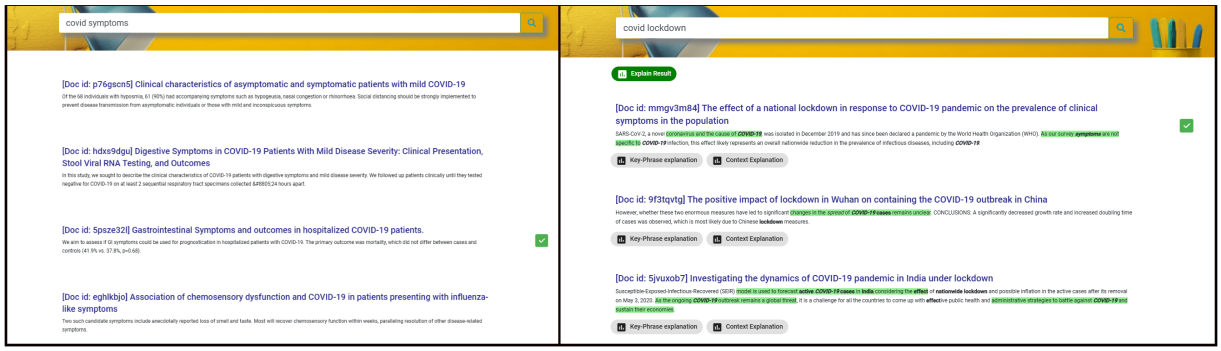

Figure 4: Left SUI with no explainability and Right SUI with explainability.

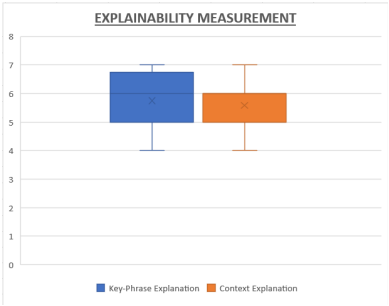

Figure 5: Explainability of Key-phases and context

RQ1. How well do the key phrases in comparison to the summary of a document help the user select relevant documents during Relevance Feedback?
To address this question, users rated whether the highlighted key phrases in each document helped them assess their relevance to their query in the feedback section. The highlighted key phrases received a mean rating of 6.25 (SD = 0.97), and summaries were rated with a mean of 6.08 (SD = 0.79). The WSR (Wilcoxon Signed-Rank) show that there is statistical evidence (p-value = 0.03611, and 95% confidence), indicating the helpfulness of key-phrases over summaries.

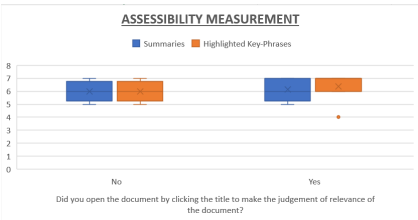

Figure 6: Evaluating Assessability

RQ2. How do users rate the explanation of feedback in the explainable search interface, after Relevance Feedback? We wanted to investigate if users could comprehend how their feedback (Rocchio algorithm) was being incorporated to refine the results. To address this, we included explanation elements on the final results page and asked participants to rate their helpfulness in clarifying document rankings on a Likert scale (0–7). Key-phrase explanations received a mean rating of 5.75 (SD = 0.97), and context explanations averaged 5.58 (SD = 0.79). For System 1 (Regular Interface), 25% of the participants reported understanding the ranking principle, while 75% did not comprehend why certain documents ranked higher. Our evaluation suggests that the graphical representation of the Rocchio algorithm did not meet user expectations, although key phrases and context helped (see Fig. 5 and 6) to a degree.

RQ3. Does the extended Rocchio Algorithm in RelEx provide the user with more relevant results than the original Rocchio Algorithm? In the user study, we experimented to assess whether the extended RelEx outperformed the traditional Rocchio method in relevance and precision. Users' queries and selected document IDs were tested in two systems (See Fig.4): one with the standard Rocchio algorithm (SUI-one) and the other with the XAI-extended version (SUI-two). SUI-two consistently retrieved more relevant documents than SUI-one; We observed that it is statistically significant in WSR (p-value = 0.03715, and 95% confidence), indicating that the extended Rocchio algorithm is more effective to retrive relevant results. We investigated some results manually, for instance, in response to multiple queries (e.g. "Long COVID symptoms", "pandemic death") - SUI-two retrieved 2 of 3 selected documents, while SUI-one retrieved none. For comparison of RelEx with another explainable search baseline, we manually inspected a couple of queries in the proposed RelEx and EXS baseline [32].

Limitations: Although, there was no clear winner between LIME [28] based EXS explanations versus RelEx, synthetically generated LIME features can sometimes be difficult to be consumed by a non-AI expert user, compared to key-words in context. However, we are aware of the limitations - results of the above XAI evaluation are anecdotal evidence for a small set of users.

5 Conclusion & Future Work

We propose RelEx, an explainable search engine combining user feedback with an extended Rocchio model and XAI elements such as key-phrases, summaries, and context terms. Empirical results show improved retrieval performance and user-understandability, though future work is needed to enhance stability, scalability, and completeness, using fine-tuned LLMs and larger-scale user studies with eye-tracking.

References

[1] Mohannad ALMasri, Catherine Berrut, and Jean-Pierre Chevallet. 2016. A comparison of deep learning based query expansion with pseudo-relevance feedback and mutual information. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*. Springer, 709–715.

[2] Avishek Anand, Procheta Sen, Sourav Saha, Manisha Verma, and Mandar Mitra. 2023. Explainable Information Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3448–3451.

[3] Peter G Anick and Shivakumar Vaithyanathan. 1997. Exploiting clustering and phrases for context-based information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 314–323.

[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[5] Nolwenn Bernard and Krisztian Balog. 2025. A Systematic Review of Fairness, Accountability, Transparency, and Ethics in Information Retrieval. *Comput. Surveys* 57, 6 (2025), 1–29.

[6] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 405–414.

[7] Carlos Castillo. 2019. Fairness and transparency in ranking. In *ACM SIGIR Forum*, Vol. 52. ACM New York, NY, USA, 64–71.

[8] Ioannis Chios and Suzan Verberne. 2021. Helping results assessment by adding explainable elements to the deep relevance matching model. *arXiv preprint arXiv:2106.05147* (2021).

[9] Azin Ghazimatin, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. ELIXIR: Learning from user feedback on explanations to improve recommender models. In *Proceedings of the Web Conference 2021*. 3850–3860.

[10] Nikolaos Giarelis, Charalampos Mastrokostas, and Nikos Karacapilidis. 2023. Abstractive vs. extractive summarization: An experimental review. *Applied Sciences* 13, 13 (2023), 7620.

[11] David Gunning and David Aha. 2019. DARPA's explainable artificial intelligence (XAI) program. *AI Magazine* 40, 2 (2019), 44–58.

[12] Donna Harman. 1992. Relevance feedback revisited. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1–10.

[13] Thorsten Joachims et al. 1997. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *ICML*, Vol. 97. Citeseer, 143–151.

[14] Petros Karvelis, Dimitris Gavrilis, George Georgoulas, and Chrysostomos Stylios. 2018. Topic recommendation using Doc2Vec. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–6.

[15] Diane Kelly et al. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval* 3, 1–2 (2009), 1–224.

[16] Craig Macdonald and Nicola Tonellotto. 2020. Declarative experimentation in information retrieval using PyTerrier. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. 161–168.

[17] Rui Meng, Debanjan Mahata, and Florian Boudin. 2022. From fundamentals to recent advances: A tutorial on keyphrasification. In *European Conference on Information Retrieval*. Springer, 582–588.

[18] Siyu Mi and Jiepu Jiang. 2019. Understanding the interpretability of search result summaries. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 989–992.

[19] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11, 3-4 (2021), 1–45.

[20] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713* (2020).

[21] Sayantan Polley. 2022. Towards Explainable Search in Legal Text. In *European Conference on Information Retrieval*. Springer, 528–536.

[22] Sayantan Polley et al. 2025. Demo Video of Relevance-Explainer - RelEx. https://vimeo.com/1026112074/. [Online; Last accessed 30-Apr-2025].

[23] Sayantan Polley et al. 2025. Github code for Relevance-Explainer - RelEx. https://github.com/sayantanpolley/Information-Retrieval-Project. [Online; Last accessed 30-Apr-2025].

[24] Sayantan Polley, Atin Janki, Marcus Thiel, Juliane Hoebel-Mueller, and Andreas Nuernberger. 2021. ExDocS: Evidence based Explainable Document Search. In *Workshop on Causality in Search and Recommendations co-located with 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[25] Sayantan Polley, Rashmi Raju Koparde, Akshaya Bindu Gowri, Maneendra Perera, and Andreas Nuernberger. 2021. Towards trustworthiness in the context of explainable search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2580–2584.

[26] Sayantan Polley, Subhajit Mondal, Venkata Srinath Mannam, Kushagra Kumar, Subhankar Patra, and Andreas Nürnberger. 2022. X-vision: Explainable image retrieval by re-ranking in semantic space. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4955–4959.

[27] Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2021. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *arXiv preprint arXiv:2101.05667* (2021).

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386* (2016).

[29] Joseph John Rocchio. 1971. The smart retrieval system: Experiments in automatic document processing. *Relevance feedback in Information Retrieval* (1971), 313–323.

[30] Agata Savary and Yue Zhang. 2020. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020): Tutorial Abstracts. (2020).

[31] Procheta Sen, Debasis Ganguly, Manisha Verma, and Gareth JF Jones. 2020. The curious case of IR explainability: Explaining document scores within and across ranking models. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2069–2072.

[32] Jaspreet Singh and Avishek Anand. 2019. Exs: Explainable search using local model agnostic interpretability. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 770–773.

[33] Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. TREC-COVID: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, Vol. 54. ACM New York, NY, USA, 1–12.