Explainable Information Retrieval

Anand, Avishek; Sen, Procheta; Saha, Sourav; Verma, Manisha; Mitra, Mandar

DOI
10.1145/3539618.3594249

Publication date
2023

Document Version
Final published version

Published in
SIGIR 2023 - Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval

Citation (APA)
Anand, A., Sen, P., Saha, S., Verma, M., & Mitra, M. (2023). Explainable Information Retrieval. In *SIGIR 2023 - Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 3448-3451). (SIGIR 2023 - Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval). Association for Computing Machinery (ACM). https://doi.org/10.1145/3539618.3594249

Explainable Information Retrieval

Avishek Anand
Delft Institute of Technology
Delft, The Netherlands
Avishek.Anand@tudelft.nl

Procheta Sen
University of Liverpool, UK
procheta.sen@liverpool.ac.uk

Sourav Saha
Indian Statistical Institute, India
sourav.saha_r@isical.ac.in

Manisha Verma
Amazon, New York
mvr@amazon.com

Mandar Mitra
Indian Statistical Institute, India
mandar@isical.ac.in

ABSTRACT

This tutorial presents explainable information retrieval (ExIR), an emerging area focused on fostering responsible and trustworthy deployment of machine learning systems in the context of information retrieval. As the field has rapidly evolved in the past 4-5 years, numerous approaches have been proposed that focus on different access modes, stakeholders, and model development stages. This tutorial aims to introduce IR-centric notions, classification, and evaluation styles in ExIR, while focusing on IR-specific tasks such as ranking, text classification, and learning-to-rank systems. We will delve into method families and their adaptations to IR, extensively covering post-hoc methods, axiomatic and probing approaches, and recent advances in interpretability-by-design approaches. We will also discuss ExIR applications for different stakeholders, such as researchers, practitioners, and end-users, in contexts like web search, patent and legal search, and high-stakes decision-making tasks. To facilitate practical understanding, we will provide a hands-on session on applying ExIR methods, reducing the entry barrier for students, researchers, and practitioners alike.

CCS CONCEPTS

• **Information systems** → **Explainable AI**.

KEYWORDS

explainable information retrieval, posthoc interpretability, interpretable by design, axiomatic ranking, probing

ACM Reference Format:
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 (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3539618.3594249

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

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3594249

1 MOTIVATION

Explainability and transparency are paramount in fostering the responsible and trustworthy deployment of machine learning systems, as they facilitate interpretability, accountability, and trust [30]. Particularly in the context of information retrieval (IR), there has been a surge of interest in explainability over the past 4-5 years. By ensuring that the inner workings of these models are accessible and understandable, researchers and practitioners can detect and rectify biases, that may unintentionally skew the information landscape and limit diverse perspectives [44, 45]. Consequently, the research community can effectively scrutinize IR models leading to more accurate, efficient, and fair information retrieval systems. There has been a large amount of growing yet unorganized work that covers many tasks and aspects of explainable information retrieval. Inspired by [18, 27], early works in ExIR focussed on posthoc approaches to understand already learned models [12, 22, 31, 35]. Later on, there has been a large variety of approaches that have focussed on different explanation styles [4, 23], different modalities [33] and investigations and grounding to IR-specific abilities [29, 36]. In IR, explainability was initially used to explain the retrieval output to a non-IR expert. ExIR can be broadly categorized into three areas. They are *post-hoc explanations*, *grounding to IR properties*, or methods that are *interpretable by design* [2]. There are numerous applications to IR depending the target stakeholder of the system. Researchers, model developers, and practitioners are interesting discovering new science or debugging the validity and correctness of their models in various ranking and QA tasks [28]. For end-users, its particularly useful in high-stakes decision making retrieval tasks like patent and legal search [22].

1.1 Objectives

The first aim of this tutorial is to introduce the IR-centric notions, classification, and evaluation styles in ExIR in the last few years. Different from existing general-purpose tutorials, our tutorial focuses mainly on IR-specific tasks – ranking, text classification, and learning-to-rank systems – and IR-specific explanations. Secondly, we will delve into the method families and their adaptations to IR – specifically ranking tasks. We will extensively cover posthoc methods, axiomatic and probing approaches, and most recent advances in interpretability-by-design approaches. Finally, for Web search that has a large user base, ExiR primitives and tools can be used to identify and control for user undesirable biases that might exist. To this extent, one of our aims is to reduce the entry barrier for students, researchers, and practitioners by providing a hands-on session on application of ExIR methods.

1.2 Scope of the Tutorial

Many of the methods in ExIR have methodological overlap with those invented in ML, natural language processing (NLP), and recommender systems (RS) communities. We only focus on core-IR issues in this tutorial and, wherever possible, clearly spell out the distinctions from similar approaches in NLP, RS and ML in general. Explainable methods are also related to methods addressing adversarial attacks in retrieval models to some extent. However, in our tutorial we will focus on only explainability approaches.

2 RELEVANCE TO THE IR COMMUNITY

We first detail similar tutorials in the IR and other related communities, and argue why and how our tutorial is both timely and fills the gap of a missing tutorial on *Explainable IR*.

2.1 Related tutorials

From the past few there has been a significant number of tutorials on the topic of explainability in many disciplines in computer science. In KDD 2021 [8] there was a tutorial on explainability for NLP. In AAAI 2021 and 2022 there were tutorials on explainable AI. There was also a tutorial on the evaluation of explanations in NAACL 2022. Specific to the IR community, the last tutorial that related to explainability was held four years prior on the topic of "explainable recommendation and search" in WWW 2019 [42], SIGIR 2019 [41] and ICTIR 2019 [40] by the same set of authors. However, this tutorial's focus was not on search tasks and models. This was perhaps because of the limited amount of literature in ExIR. Very recently, in ECIR 2023, there will be a tutorial on Neuro-symbolic approaches for IR that focus partially on interpretability [9].

2.2 A need for a new one

In spite of the above tutorials and attempts, there has not been a single tutorial solely on the topic of ExIR. Many IR-specific tasks need special adaptations and are not covered in earlier tutorials. We believe that this is a gap that our tutorial on ExIR can plug. For example, explaining rankings entails interpreting pairwise [32] or listwise decisions [38] and not solely pointwise decisions. In this tutorial, we will focus on IR-specific notions, and approaches in explainable AI from an IR point of view. Additionally, in the last few years, there have been many papers and approaches that address topics of explainability in information retrieval. Specifically, there in the last five years there has been lot of progress in rationale-based models [15, 16, 43] that claim to be interpretable by design. In our tutorial we also present both classical IR approaches – like axiomatic IR [1] – and modern probing methods [3] to understand ranking models. Both these approach families are fairly recent and have not been covered in earlier tutorials. We aim to help IR researchers and practitioners (through a hands-on session) to explore this domain and it will also provide a platform to discuss future research directions in this domain.

3 NOTIONS AND CLASSIFICATION

Explainability has different notions based on the output produced by an explainer module, the explanation provided by the explainer module, or the methodology used to explain a model. We note that, *interpretability* and explainability are two subtly related concepts in the literature. While interpretability refers to the ability of a machine learning model to be understood and analyzed by a human, *explainability* refers to the ability of a model to provide an explanation of its decision-making process. In this paper we use both these concepts interchangeably.

Explainability can be local or global depending on the explainability the explainer module provides. Local interpretability aims to explain a retrieval model's output in a specific query's locality. On the other hand, global interpretability does not differentiate between queries in terms of model parameters, input spaces, etc. It aims to provide a global perspective to the model. Similarly, interpretability in IR can be pointwise, listwise, or pairwise depending on the nature of the output provided by the explainer module. Ranking models output a ranked candidate list for a given query. The explanation of pointwise methods can only explain the models' decision of a single element in the list, while pairwise methods intend to explain the model's preference of a candidate pair. The explanation of listwise methods, however, aims to cover all individual decisions in the entire ranking list. Interpretability approaches can be also categorized into black box and white box approaches depending on whether we have access to the retrieval model or not. Black box approaches are mainly posthoc explanations where the explainer module only takes the query and the ranked list of documents as input. White box approaches are the ones where the explainer model has access to the model parameters.

In this tutorial we categorized the explainability approaches into posthoc explanations, axiomatic strategies, probing strategies, and interpretable by design approaches.

4 POSTHOC INTERPRETABILITY

Post-hoc explanations broadly use feature attribution approach or generative approach to explain a retrieval model.

4.1 Feature Attribution based Approaches

For feature attribution-based approaches, the importance scores corresponding to a feature are commonly visualized using a heatmap or a bar chart, informing the user about which features the model's prediction is most sensitive to. The work in [24] computed the importance of tokens to interpret a BERT based ranking model. The study in [31] proposed a LIME (i.e. a state-of-the-art local explanation generation framework) based explanation approach named EXS to address questions like 1) Why is a document relevant to the query, 2) Why is a document ranked higher than another document, and 3) What the intent of the query is according to the ranker? The work in [21] compared EXS with their evidence-based explainable document search system, ExDocS, which performs reranking using interpretable features. Similarly, the study in [35] adapted LIME to create locally interpretable ranking model explanations (LIRME). In contrast to EXS, LIRME trains the local surrogate model directly on the query-document scores and does not transform them into class probabilities.

Apart from explaining a retrieval model, there exist approaches to explain learning to rank (LTR) models. The study in [33] distills an already trained black-box LTR model into an interpretable global surrogate model that is used to generate explanations. The work in [33] proposed a simple, yet effective greedy search-based

approach to find a subset of explanatory features that maximizes two measures, validity and completeness.

There has been a few attempts to explain a retrieval model with gradient based explanations. The study in [23] used simple gradient-based feature attribution to find the most important features used by LTR models using saliency maps. The work in [39] used integrated gradients [34] to obtain feature attributions for a BERT-based ranking model. All the approaches discussed above attempted to explain a retrieval model from the perspective of a common user who does not have IR expertise. In contrast, the work in [30] proposed a regression framework to explain a retrieval model from an IR practitioner's perspective.

Feature attribution-based approaches are in general evaluated by model fidelity score or by removing the top k feature and observing the performance of a model compared to a scenario where no features were removed.

4.2 Generating Free Text Explanations

In free text explanation-based approaches, rather than selecting top k features or attributes, a set of words or phrases are generated to explain a retrieval model [19, 32]. The study in [25] proposed a transformer-based explanation model GenEx which learns to generate a text sequence that explains the relevance of a document corresponding to a query. Similarly, the study in [38] proposed a listwise explanation generator (LiEGe) that for a given query jointly explains all the documents contained in a ranked result list. [19, 32] propose model-agnostic approaches to interpret a query intent as understood by a black-box ranker. The goal is to identify a set of query expansion terms such that most of the pairwise preferences in the output ranking are preserved. The study in [39] introduced a Query-to-Intent-Description task for query understanding. Given a query and a set of both relevant and irrelevant documents, the goal is to generate a natural language intent description.

Free-text explanations are generally evaluated using human annotated ground-truth data. Most IR datasets do not include explanations. As a result of this, proxy explanations can be created from query descriptions, query aspect annotations, topic annotations, or click logs [25]. With recent advancements in generative models, one of the important research questions is to investigate the explanation units for generative retrieval models.

5 AXIOMATIC AND PROBING STRATEGIES

5.1 Axiomatic Analysis

The formal study of explainability in IR dates back to the seminal work by [11]. They used heuristics in terms of axioms to explain several statistical retrieval models (*term-weighting*). In general, this framework provides axioms to explain why a particular document D_i is ranked above D_j for a given query Q.

Recently, Hagen et al. [14] relaxed several popular axioms and in a subsequent work by [36] these were used to explain neural ranking models. In specific, they study to what extent neural models obey the axiomatic constraints. Rennings et al. [26] created a diagnostic dataset for IR. The objective is to evaluate the axiomatic efficacy of several IR models. This was extended further by [5] for analyzing DistillBERT and the observation was that the existing semantic axioms are not sufficient enough to analyze the performance of BERT.

5.2 Probing

Probing classifiers are widely adapted to analyze the content of latent embeddings – information encoded in the model's parameters and its representations. Typically, probing involves training a lightweight separate classifier to directly predict some specific property (e.g., part-of-speech tags, relevance, matching, etc.) from the learned representation. Early work on probing based techniques for IR/QA [7]trained a multilayered LSTM on passage retrieval tasks. The objective is to analyze what sort of NLP features, i.e., POS, NER, is encoded at the intermediate neural representations. In a similar thread, recent work by MacAvaney et al. [20] proposed three probing strategies to analyze neural IR models. They observed how sensitive the ranking models are with reference to various textual properties, e.g., fluency, succinctness, typos, paraphrases, etc. Specific to text ranking, probing tasks ask different questions – what is the world knowledge contained is fine-tuned rankers ? [6, 13] What are the IR abilities of rankers ? [10, 37]

6 INTERPRETABLE BY DESIGN APPROACHES

The general architecture of these models involves intermediate feature extraction, and a task-specific decision structure. not all components are fully interpretable to ensure competitive task performance. Therefore, most of the interpretable by design approaches resort to making only specific components interpretable or transparent.

There has been a very few attempts to explain a retrieval model or its components using interpretable by design approach. The study in [46] employs an isolated black-box (e.g., neural networks) model to generate a score indicating the contribution (or importance) of the feature in a LTR model. Similarly, the work in [17] shares a similar structure as [46], while using LambdaMART as the sub-model. [17] starts from learning a set of trees, with each dealing with one single distinct feature only. This step enables to identify a small yet crucial set of features and exclude the rest.

In general, interpretable by design approach category methods evaluate the goodness of explanations using anecdotal examples. Additionally, [46] compares the features to a referenced tree-model, and justifies the faithfulness of explanations by a similar trend.

7 SUPPORT FOR ATTENDEES AND SCHEDULE

We will provide the attendees with a link to the tutorial slides and preparatory reading material. Upon acceptance, we will prepare a webpage with all updated information and the necessary reading material, and python notebooks well in time before the conference. A detailed schedule for our proposed *half-day tutorial* (three hours plus breaks), which is aimed to meet a high-quality presentation within the chosen time period, is as follows:

- **Part I.** *Motivation, Scope and Notions of IR Explainability* (**35 minutes**)
  - Motivation (10 mins)
  - Scope (i.e. breadth/depth) of the tutorial. (10 mins)

  - Notions of IR Explainability (15 mins)
- **Part II.** *Posthoc explanations and evaluation* (**45 minutes**)
  - Feature-attributions (15 minutes)
  - Free-text explanations(15 minutes)
  - Evaluation of explanations (15 minutes)
- QA Session (**10 minutes**)

––Coffee Break––

- **Part III.** *Axiomatic and probing strategies* (**25 minutes**)
  - Interpretability using Axiomatic IR (15 minutes)
  - Probing for IR abilities (10 minutes)
- **Part IV.** *Interpretable-by-design approaches*(**20 minutes**)
  - Rationale-based models (10 mins)
  - Select-then-Rank models (5 mins)
  - Learning-to-rank (5 mins)
- **Part V.** *Hands-on session and New horizons* (**30 minutes**)
  - Demonstration & Hands-on session (15 mins)
  - New horizons (15 minutes)
- QA Session (**10 minutes**)

ACKNOWLEDGEMENT

This work is partially supported by BMBF grant no. 13N16052 (Qubra).

REFERENCES

[1] Enrique Amigó, Hui Fang, Stefano Mizzaro, and ChengXiang Zhai. 2017. Axiomatic Thinking for Information Retrieval: And Related Tasks. In *Proc. of SIGIR 2017*. 1419–1420.
[2] Avishek Anand, Lijun Lyu, Maximilian Idahl, Yumeng Wang, Jonas Wallat, and Zijian Zhang. 2022. Explainable Information Retrieval: A Survey.
[3] Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Comput. Linguistics* 1 (2022), 207–219.
[4] Alexander Bondarenko, Maik Fröbe, Jan Heinrich Reimer, Benno Stein, Michael Völske, and Matthias Hagen. 2022. Axiomatic Retrieval Experimentation with ir_axioms. In *Proc. of SIGIR 2022*. 3131–3140.
[5] Arthur Câmara and Claudia Hauff. 2020. Diagnosing BERT with Retrieval Heuristics. In *Proceedings of ECIR 2020*. 605–618.
[6] Jaekeol Choi, Euna Jung, Sungjun Lim, and Wonjong Rhee. 2022. Finding Inverse Document Frequency Information in BERT. *ArXiv preprint* (2022).
[7] Daniel Cohen, Brendan O'Connor, and W. Bruce Croft. 2018. Understanding the Representational Power of Neural Retrieval Models Using NLP Tasks. In *Proc. 2018 ACM ICTIR*. 67–74.
[8] Marina Danilevsky, Shipi Dhanorkar, Yunyao Li, Lucian Popa, Kun Qian, and Anbang Xu. 2021. Explainability for Natural Language Processing. In *Proc. of SIGKDD 2021*. 4033–4034.
[9] Laura Dietz, Hannah Bast, Shubham Chatterjee, Jeff Dalton, Edgar Meij, and Arjen de Vries. 2023. Neuro-Symbolic Approaches for Information Retrieval. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland*. 324–330.
[10] Yixing Fan, Jiafeng Guo, Xinyu Ma, Ruqing Zhang, Yanyan Lan, and Xueqi Cheng. 2021. A Linguistic Study on Relevance Modeling in Information Retrieval. 1053–1064.
[11] Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A Formal Study of Information Retrieval Heuristics. In *Proceedings of SIGIR 2004*. 49–56.
[12] Zeon Trevor Fernando, Jaspreet Singh, and Avishek Anand. 2019. A study on the Interpretability of Neural Retrieval Models using DeepSHAP. In *Proc. of SIGIR 2019*. 1005–1008.
[13] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. A White Box Analysis of ColBERT. In *Proc. of ECIR 2021*. 257–263.
[14] Matthias Hagen, Michael Völske, Steve Göring, and Benno Stein. 2016. Axiomatic Result Re-Ranking. In *Proc. of CIKM 2016*. 721–730.
[15] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing Neural Predictions. In *Proc of EMNLP 2016*. Austin, Texas, 107–117.
[16] Jurek Leonhardt, Koustav Rudra, and Avishek Anand. 2021. Extractive Explanations for Interpretable Text Ranking. *ACM Transactions on Information Systems* (2021).
[17] Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, and Alberto Veneri. 2022. ILMART: Interpretable Ranking with Constrained LambdaMART. In *Proc. of SIGIR 2022*. 2255–2259.
[18] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proc. of NIPS 2017*. 4765–4774.
[19] Lijun Lyu and Avishek Anand. 2023. Listwise Explanations for Ranking Models Using Multiple Explainers. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland*. Springer, 653–668.
[20] Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey, and Arman Cohan. 2020. ABNIRML: Analyzing the Behavior of Neural IR Models. *ArXiv preprint* (2020).
[21] Sayantan Polley. 2022. Towards Explainable Search in Legal Text. In *European Conference on Information Retrieval*. Springer, 528–536.
[22] Sayantan Polley, Atin Janki, Juliane Thiel, Marcusand Hoebel-Mueller, and Andreas Nuernberger. 2021. ExDocS: Evidence based Explainable Document Search. In *Proc. of SIGIR Workshop on Causality in Search and Recommendation 2021*.
[23] Alberto Purpura, Karolina Buchner, Gianmaria Silvello, and Gian Antonio Susto. 2021. Neural feature selection for learning to rank. In *Proc. of ECIR 2021*. 342–349.
[24] Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. Understanding the Behaviors of BERT in Ranking. *ArXiv preprint* (2019).
[25] Razieh Rahimi, Youngwoo Kim, Hamed Zamani, and James Allan. 2021. Explaining Documents' Relevance to Search Queries. *ArXiv preprint* (2021).
[26] Daniël Rennings, Felipe Moraes, and Claudia Hauff. 2019. An Axiomatic Approach to Diagnosing Neural IR Models. In *Proceedings of ECIR 2019*. 489–503.
[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proc.of SIGKDD 2016*. 1135–1144.
[28] Rishiraj Saha Roy and Avishek Anand. 2021. Question Answering for the Curated Web: Tasks and Methods in QA over Knowledge Bases and Text Collections. *Synthesis Lectures onSynthesis Lectures on Information Concepts, Retrieval, and Services* 13, 4 (2021), 1–194.
[29] Procheta Sen, Debasis Ganguly, Manisha Verma, and Gareth J. F. Jones. 2020. The Curious Case of IR Explainability: Explaining Document Scores within and across Ranking Models. In *Proc. of SIGIR 2020*. 2069–2072.
[30] Procheta Sen, Sourav Saha, Debasis Ganguly, Manisha Verma, and Dwaipayan Roy. 2022. Measuring and Comparing the Consistency of IR Models for Query Pairs with Similar and Different Information Needs. In *Proc of CIKM 2022*. 4449–4453.
[31] Jaspreet Singh and Avishek Anand. 2019. EXS: Explainable Search Using Local Model Agnostic Interpretability. In *Proc. of WSDM 2019*. 770–773.
[32] Jaspreet Singh and Avishek Anand. 2020. Model agnostic interpretability of rankers via intent modelling. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 618–628.
[33] Jaspreet Singh, Megha Khosla, Wang Zhenye, and Avishek Anand. 2021. Extracting per Query Valid Explanations for Blackbox Learning-to-Rank Models. In *Proc. of ICTIR 2021*. 203–210.
[34] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proc of ICML 2017 (Proceedings of Machine Learning Research)*. 3319–3328.
[35] Manisha Verma and Debasis Ganguly. 2019. LIRME: Locally Interpretable Ranking Model Explanation. In *Proc. of SIGIR 2019*. 1281–1284.
[36] Michael Völske, Alexander Bondarenko, Maik Fröbe, Benno Stein, Jaspreet Singh, Matthias Hagen, and Avishek Anand. 2021. Towards Axiomatic Explanations for Neural Ranking Models. In *Proc. of ICTIR 2021*. 13–22.
[37] Jonas Wallat, Fabian Beringer, Abhijit Anand, and Avishek Anand. 2023. Probing BERT for ranking abilities. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland*. Springer, 255–273.
[38] Puxuan Yu, Razieh Rahimi, and James Allan. 2022. Towards Explainable Search Results: A Listwise Explanation Generator. In *Proc. of SIGIR 2022*. 669–680.
[39] Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2020. Query Understanding via Intent Description Generation. In *Proc. of CIKM 2020*. 1823–1832.
[40] Yongfeng Zhang. 2019. Tutorial on Explainable Recommendation and Search. In *Proc. of ICTIR 2019*. 255–256.
[41] Yongfeng Zhang, Jiaxin Mao, and Qingyao Ai. 2019. SIGIR 2019 Tutorial on Explainable Recommendation and Search. In *Proc. of SIGIR 2019*. 1417–1418.
[42] Yongfeng Zhang, Jiaxin Mao, and Qingyao Ai. 2019. WWW'19 Tutorial on Explainable Recommendation and Search. In *Proc. of WWW 2019*. 1330–1331.
[43] Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021. Explain and Predict, and then Predict Again. In *WSDM '21, Israel, March 8-12, 2021*. ACM, 418–426.
[44] Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021. FaxPlainAC: A Fact-Checking Tool Based on EXPLAINable Models with HumAn Correction in the Loop. In *Proceedings of the 30th ACM CIKM*. 4823–4827.
[45] Zijian Zhang, Vinay Setty, and Avishek Anand. 2022. SparCAssist: A Model Risk Assessment Assistant Based on Sparse Generated Counterfactuals. In *Proc. of SIGIR*. 3219–3223.
[46] Honglei Zhuang, Xuanhui Wang, Michael Bendersky, Alexander Grushetsky, Yonghui Wu, Petr Mitrichev, Ethan Sterling, Nathan Bell, Walker Ravina, and Hai Qian. 2021. Interpretable Ranking with Generalized Additive Models. In *Proc. of WSDM 2021*. 499–507.