On the Application of a Common Theoretical Explainability Framework in Information Retrieval

Andrea Albarelli[1], Claudio Lucchese[1], Matteo Rizzo[1] and Alberto Veneri[1,2,*]

[1]*Ca' Foscari University of Venice*
[2]*ISTI-CNR*

Abstract

Most of the current state-of-the-art models used to solve the search and ranking tasks in Information Retrieval (IR) are considered "black boxes" due to the enormous number of parameters employed, which makes it difficult for humans to understand the relation between input and output. Thus, in the current literature, several approaches are proposed to explain their outputs, trying to make the models more explainable while maintaining the high level of effectiveness achieved. Even though many methods have been developed, there is still a lack of a common way of describing and evaluating the models and methods of the Explainabile IR (ExIR) field. This work shows how a common theoretical framework for explainability (previously presented in the biomedical field) can be applied to IR. We first describe the general framework and then focus on specific explanation techniques in the IR field, focusing on core IR tasks: search and ranking. We show how well-known methods in ExIR fit into the framework and how specific IR explainability evaluation metrics can be described using this new setting.

Keywords

Explainable Information Retrieval, Theoretical Explainability Framework, Search and Ranking

1. Introduction

The emergence of Deep Learning (DL), and in particular the application of Pretrained Language Models (PLMs), have drastically changed the Information Retrieval (IR) landscape. Previously skeptically considered by part of the community [1], the advent of the first publicly available PLM (BERT [2]) has completely changed the adoption of DL models in the IR field, especially in the core IR tasks, i.e. *search* and *ranking*, given their high effectiveness. Even though PLM-based approaches are highly effective, they are also opaque and way more challenging to analyze, debug, and understand than the traditional IR methods, such as the well-known BM25 [3]. Therefore, in pursuit of ensuring more reliable and trustworthy IR systems, recent years have witnessed a growing interest in the field of Explainable Information Retrieval (ExIR) [4]. The motivation to go beyond the opacity of the current state-of-the-art method is not purely technical (e.g., to create more robust and simple to debug IR systems) or ethical (e.g., to easily investigate possible unfair behavior in the model), but it also has a compliance nature, given the current or

IIR2024: The 14th Italian Information Retrieval Workshop, 5th - 6th September 2024, Udine, Italy

[*] Corresponding author.

albarelli@unive.it (A. Albarelli); claudio.lucchese@unive.it (C. Lucchese); matteo.rizzo@unive.it (M. Rizzo); alberto.veneri@unive.it (A. Veneri)

0000-0002-3659-5099 (A. Albarelli); 0000-0002-2545-0425 (C. Lucchese); 0000-0002-9723-211X (M. Rizzo); 0000-0003-2094-3375 (A. Veneri)

upcoming international regulation for Artificial Intelligence (AI) systems, such as the AI Act, in which the concept of explainability is a crucial one [5], even though most of the state-of-the-art eXplainable Artificial Intelligence (XAI) methods can provide only a limited answer to the compliance requirements [6].

As commonly happens in a fast-growing field, one of the problems faced by ExIR is the difficulty of relating different approaches since different terminology is used and different evaluation metrics have been proposed. For example, some methods are called "interpretable" while others "explainable" and some works present evaluation metrics related to the same concept. However, it is unclear how they should be compared or which characteristic of the model they try to underline. Some recent surveys [4, 7] tried to organize the relevant literature; however, they did not create a common framework useful to compare and evaluate the explainability of different IR models. This is problematic, especially if we want to evaluate various methods that belong to different categories, which usually have different evaluation metrics.

With this work, we aim to show how one of our recent contributions presented in the biomedical domain, a theoretical framework for explainability [8], can be suitably employed also in the IR domain, and thus describe all the methods present in the ExIR literature through the same lens. By applying the framework to IR, we show how the current most popular explanation methods in IR fit the framework. Even though we highlight that the framework is not a novel contribution per se, we claim that starting to apply it to the new explanation techniques and model presented in IR is an important step towards building more rigorous explanation methods that take into account all the aspects related to this interdisciplinary and complex subfield. We present how three so-called post-hoc explanation methods fit the framework: LIRME[9], MULTIPLEX [10], and the explainability techniques based on IR axioms, while also analyzing two so-called "intrinsically interpretable" (or "intrinsically explainable") models such as ColBERT[11] (not explicitly interpretable but considered interpretable) and Interpretable LambdaMART (ILMART)[12].

The paper is organized as follows: in Sec. 2, we present the most related works; in Sec. 3, we briefly summarize the framework; in Sec. 4 we highlight the peculiarities of applying the framework in the IR presenting 5 case studies, and, finally, in Sec. 5 we present our conclusions.

2. Related Works

Even though the focus on the explainability aspects in IR is relatively new, numerous works have been published, but no general framework has been proposed. To the author's knowledge, this work presents the application of a general framework to explainability techniques in IR for the first time in the literature. Nonetheless, other works have been presented in the form of surveys, categorizing ExIR works and trying to create a common taxonomy. However, they do not directly try to create a common framework to describe and compare ExIR methods but mainly focus on categorizing them.

The more relevant survey is the one presented by Anand et al. [4], in which the authors have nicely categorized 32 ExIR approaches into three general categories: *i)* post hoc, *2)* grounding to IR properties, and *3)* interpretable by design. The survey presents various explanation aspects,

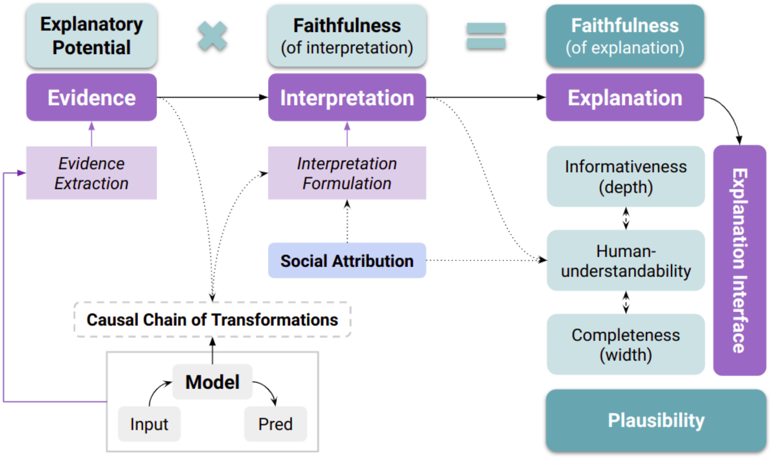

Figure 1: Overview of the theoretical framework of explainability. Schema presented in [8].

including the difference between local and global explanations and the difference between pointwise, pairwise, and listwise explanations. One of the main conceptual differences between their survey and our work is that they differentiate between “interpretability" (interpretability by design) and “explainability" (mainly post hoc explainability) while we approach the problem differently; we disagree with this categorization, and we claim that every model has some explainability degree, and differentiating between interpretable by design and post hoc explainable by design models is not well defined and sometimes misleading.

Saha et al. [7], in another survey, provide an overview of ExIR methods similar to [4] but also add some methods from the related Natural Language Processing (NLP) field. Similarly to Anand et al., they differentiate between interpretable and explainable models and further categorize the approaches into categories, including embeddings, sequence models, attention, transformers, and BERT. Similarly to [7], their work is designed to solve the categorization problem of the various explanation techniques, and they do not provide a common framework to describe and compare the ExIR methods available.

More broadly, some attempts have been made to outline the differences in the terminology used in XAI related field, such as in [13] and in [14]. Still, they lack a definition of an explanation’s inner structure and meaning.

In brief, as we have common frameworks to describe the learning phase of various Machine Learning (ML) tasks, including supervised learning, unsupervised learning, and reinforcement learning, among others [15], we aim to present a common framework to create an explanation for a ML model decision. This work shows that the framework is also suitable for all the existing ExIR techniques.

3. Theoretical Framework for Explainability

In this section, we briefly recap each framework component, accompanying the descriptions using, as an example, the well-known explanation technique based on analyzing the attention weights used in a transformer model, e.g., [16]. Since the transformer-based models are really popular and the worthiness of the explanations based on the attention weights is debated [17, 18], we aim to show how to analyze explanations techniques for which the consensus on their usefulness is not shared across the community using a common framework. A schema of the whole framework and the relation between the components taken from the original paper is presented in Fig. 1.

3.1. Explanation Framework components

In the following paragraphs, we present the fundamental components of the framework: *evidence*, *interpretation*, *explanation*, and *explanation interface*.

Evidence The *evidence*, used to create explanations for an AI system, is any information that we can retrieve from a model and that can give some understanding of its inner workings (*e.g.*, model parameters, gradients, input/output values, etc.). Two related concepts of *evidence* are the *evidence extractor* and the *explanatory potential*. The former is the process of retrieving the evidence from the model and/or its input/output data, while the latter is described as "how much of a model the selected type of evidence can explain". In particular, the concept of *explanatory potential* is fundamental in the framework since it helps the developer of the explanation technique to understand the maximum expected faithfulness level of the explanation. The measure of the explanatory potential can change case by case, just as we use different metrics for evaluating different tasks. However, if we retrieve the evidence directly from the model, a good metric for the explanatory potential can be the ratio between the number of parameters analyzed and the total number of parameters of the model. We identify the evidence with the symbol e, and with $pot(\cdot)$ the function computing the explanatory potential, and thus with $pot(e)$, we identify the explanatory potential of the evidence.

Example. In the case of attention-based explanation, the evidences are the attention weights, which can be extracted just by retrieving the weights of the attention modules in the model. The explanatory potential of this explanation can be measured as the ratio between the number of parameters in the attention modules and the total number of parameters of the net. In the case of BERT with only 12 transformer blocks[1], we have that the parameters used in the attention modules related to the "weight" and "bias" for queries and keys are 14,174,208[2] while the total number of parameters of the model is 109,482,240, and thus, we can say that the explanatory potential is approximately 13%. In this way, the explanation technique's explanatory potential is between 0% and 100%, and an explanatory potential of 13% seems insufficient to provide a faithful explanation.

Interpretation An *interpretation* is a function applied to some evidence and mapping its instances into explanations. Interpretation can be any function applied to the evidence. Still, it is usually based on some social attribution, which makes it easier for humans to understand the content and lowers the cognitive load. Interpretation can also be as simple as the identity

[1] https://huggingface.co/google-bert/bert-base-uncased

[2] We only consider the weights directly involved in computing the attention weights.

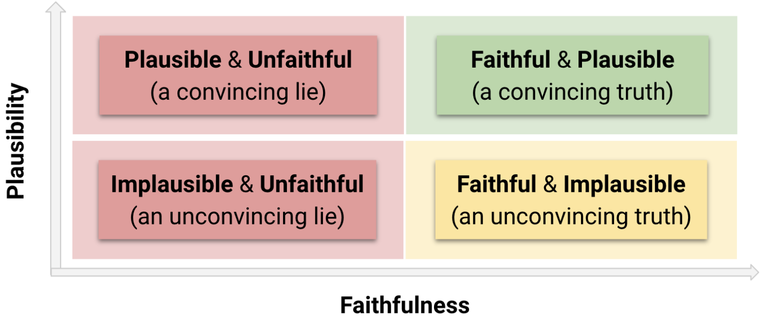

Figure 2: Overview of the outcome on the user of the interaction between faithfulness and plausibility. Schema presented in [8].

function, as in the so-called "white box" models, where no actual interpretation is needed; only the evidence is sufficient. We identify the interpretation function with the symbol $i(\cdot)$.

Example. In the case of attention-based explanation, the common interpretation is that the weights of the model's attention module can summarize the importance of the input token to the final prediction. There are two main caveats to this interpretation. First, each attention head at each transformer block gives a different weight, and it is tricky to combine this information. Second, each transformer block's input and output tokens do not necessarily relate to the same token. It is common to assume that the token in the j-th position always refers to the same concept, but this is not necessarily true; furthermore, assuming that each token is contextualized after each block of the transformer, we could have that they replace the concept of the token itself.

Explanation The concept of “explanation” is defined as “the output of an *interpretation function* applied to some *evidence*, providing the answer to a “why question” posed by the user”. In mathematical terms, the explanation ϵ results from applying the interpretation function to the retrieved evidence, $\epsilon = i(e)$.

Example. In our attention-based example, the explanation is the triples formed by the input token, the output token, and the associated attention weight for each head for each transformer block.

Explanation interface Last but not least, the explanation has to be presented to the end user with an adequate user interface characterized by three main properties: (i) human understandability, (ii) informativeness, and (iii) completeness. The *human-understandability* measures how easily the user can understand the explanation provided. The *informativeness* (i.e., depth) measures how much information is given to the user to understand the behavior of the AI system for a specific user need. Finally, the *completeness* (i.e., width) describes how well the explanation pictures the entire inner workings of the model.

Example. In the case of attention-based explanation, every attention weight is presented by transformer block and attention head interactively; see [16] for an example.

3.2. Explanation Framework Evaluation

The framework presents two main aspects to take into account during the evaluation of an AI system, namely the *plausibility* and the *faithfulness* of the explanation.

Faithfulness We define the property of *faithfulness of an interpretation* as the degree to which an interpretation accurately reflects the behavior of the transformation function applied by a ML model. Various measures of faithfulness can be associated with different types of explanations, analogous to the metrics used to evaluate an ML model's performance on a given task.

When designing a faithful explanation method, we can opt for two approaches. Faithfulness can be achieved by design incorporating this property into pre-selected interpretations during the model design process (white box models), or alternatively, we can ignore the explanation during the design and propose an explanation after the creation of the model (post hoc explanation). Although formal proofs are currently lacking in the literature, several tests for faithfulness have been recently proposed [19, 17, 18, 20].

Plausibility Plausibility is the degree to which an explanation aligns with the user's understanding of the model's partial or overall inner workings. It is worth highlighting that plausibility is mainly a property influenced by users. Unlike faithfulness, the plausibility of explanations can mainly be assessed via user studies. We highlight that plausibility and faithfulness can be competing properties of an explanation. Interestingly, an unfaithful but plausible explanation may deceive a user into believing that a model behaves according to a rationale when this is not the case. Given the attention weights' low explanatory potential and tricky interpretation, we claim that they are unfaithful but plausible. Fig. 2 provides a simplified problem overview.

Summing up, in this section, we presented the main components and evaluation criteria of the framework proposed in [8]. The components are general enough to include any explanation technique, and it is worth highlighting that the framework does not distinguish between post hoc explanation techniques and explainable by design models. All the models have to be interpreted given some evidence. The models usually called "intrinsically interpretable" are simply models in which the interpretation of the evidence is trivial.

4. Applying the Framework to Search and Ranking Applications

The proposed framework has been presented in the realm of bioinformatics, focusing on the most common explanation techniques and explainable models in the literature. This section shows how the framework can be successfully adapted to the particular case of ExIR. We describe five explanation techniques and associated evaluation metrics from five different works in the literature. In particular, we chose three so-called post hoc explanation methods: LIRME [9], MULTIPLEX [10], and the explanation technique based on the adherence to IR axioms presented in [21]. In addition, we analyze two considered white-box models: ColBERT [11] and ILMART [12]. For each explanation method proposed, we identify *i)* the evidence used, *2)* the interpretation, *3)* how the evaluation strategies relate to the concept of faithfulness and plausibility. In this case, we leave the analysis of the explanation interface for future work.

4.1. LIRME

In [9], the authors present a method to locally approximate the function of a complex text ranking model with a simple local surrogate function, similar to LIME [22] but considering sampling strategies more suitable for the ranking task. The local surrogate model is created as a scoring function $S(D, Q) = \sum_{t \in D \cap Q} w(t, D)$ in which D represents a document, Q the query, t a term in the document and query and $w(t, D)$ the term weight, learned with an optimization function to approximate the real and complex function $f(D, Q)$ of the text ranker.

Evidence After having fixed the query Q, the optimization function (Eq. 1 in [9]) to define $S(\cdot, \cdot)$ uses as evidence e the predictions made on a sample of the documents, thus $e = \{(D_i, s_i)\}_1^N$ where N is the total number of sampled documents, D_i is a sampled document and $s_i = f(D, Q)$ is the score of the complex model. The explanatory potential of this evidence $pot(e)$ can be estimated by the ratio of the documents sampled (N) over the total number of possible sampled documents in the (possibly infinite) neighborhood of the document. In [9], it is assumed that the sampled documents are only those obtained by removing terms from the document, thus limiting the neighborhood.

Interpretation The interpretation function $i(\cdot)$, in this case, takes in input the whole evidence $i\left(\{(D_i, s_i)\}_1^N\right)$ a return the model $S(\cdot, \cdot)$. In other words, the assumption is that we can explain the model's behavior just by looking at its output and using a simple model that mimics the behavior of $f(\cdot, \cdot)$. $S(\cdot, \cdot)$, as defined in [9] gives an estimate of the term importance for the model decision for each term, so if $w(t, D)$ is relatively high, it is important for the scoring function, and the opposite otherwise.

Evaluation The authors assessed the quality of their explanation with two metrics, the *explanation consistency* and the *explanation correctness*. On the one hand, evaluating the *explanation consistency* measures how a "particular choice of samples around the pivot document, D, should not result in considerable differences in the predicted explanation vector.". This metric measures faithfulness since it measures how much the sampling can change the explanation provided. Thus, this is a proxy for how difficult the function is in that particular subspace; the higher the consistency, the higher the faithfulness of the explanation. On the other hand, the *explanation correctness* measures how many terms of high contributions in the surrogate models correspond to terms occurring in documents that are judged relevant by assessors. Since its formulation does not consider the original model and is explicitly linked with the relevances given by the assessors, it can be considered a sort of user study and, thus, a measure of plausibility.

4.2. MULTIPLEX

In [10], the authors present an explanation method designed to find the subset of terms most impacting in the prediction of a text-ranker. The problem statement defined the term subset to be identified as "small," and that can explain most of the preference pairs $\{D_i \succ D_j\}$ from the original ranking π produced by a complex model, where D_i is the i-th document in the ranking π. To find the subset of terms, multiple simple ranking models are used to rank the most important features.

Evidence Similarly to LIRME, the evidence e can be defined as $e = \{(D_i, D_j, p_i)\}_1^N$ where N is the number of preference pairs are sampled from π, D_i and D_j are document in position i and j and p_i represents the preference of the complex model for the pair (either positive or negative). The explanatory potential of this evidence $pot(e)$, as for LIRME, can be estimated by the ratio of the sampled preference N with respect to the total number of possible preference pairs. We have full explanatory potential for this particular task if all the pairs are sampled.

Interpretation The interpretation function $i(\cdot)$ for MULTIPLEX takes as input the whole evidence and returns a subset of terms that explain most of the preference pairs. During the interpretation, various assumptions and heuristics were taken into account to find the subset of terms, including using only a limited subset of the terms used by the documents, using three simple ranking models (term matching, position-aware, and semantic similarity) to identify the utility of each term, and using the approximation introduced by the optimization algorithm to combine the found utility of each term.

Evaluation In [10], the evaluation is mainly measured by the "fidelity" of the explanation. The fidelity is computed with "the fraction of the maintained preference pairs by the explainers given the explanation terms." Naturally, in our framework, this is a measure of faithfulness in the explanation. Besides an anecdotal example, no evaluation has been performed using information from the end-user, and thus, no plausibility evaluation has been performed.

4.3. Explanation by IR axioms

Since other works, as [4], consider the explanation through IR axioms a completely different category, as a last example of so-called "post-hoc" explanation, we analyze the work by Câmara and Hauff [21]. In the aforementioned work, the authors used the concept of diagnostic datasets to analyze if BERT fulfills the retrieval axioms proposed by [23]. In particular, they created one diagnostic dataset, one for each axiom, and checked if the rank produced by the model was aligned with the one artificially created using the heuristic. The explanation aimed to explain the model predictions using one or more heuristics.

Evidence As in the case of LIRME and MULTIPLEX, the evidence is only based on the model score attributed to a document-query pair, i.e., $e = \{(D_i, s_i)\}_1^N$ for a fixed query, where D_i is the i-th document of a diagnosing dataset, and s_i is the associated score. The explanatory potential $pot(e)$ can be, therefore, estimated with the ratio between the number of documents taken into account N and the number of documents in the countable (in general case, possibly infinite) document space that can be created for the particular diagnosing dataset.

Interpretation The interpretation of the evidence $i(e = \{(D_i, s_i)\}_1^N)$ says that if the document order produced by BERT is aligned with the order of the diagnosing dataset, BERT follows the particular axiom with which the dataset was created.

Evaluation The evaluation is based only on the agreement between BERT ranking and the order of the documents in the diagnosing dataset. Therefore, there is no measure of the explanation's faithfulness but only a quantitative measurement of its plausibility since measuring the agreement with a diagnosing dataset is a proxy for measuring the agreement with a fictitious user.

4.4. ColBERT

Even though not explicitly presented as an explainable model, ColBERT is usually considered an "intrinsically interpretable" model [4] in which the weights on the so-called late interaction between tokens are considered term importance [24]. The late interaction is implemented using the MaxSim operator, in which for each query token representation after the last transformer block q_i, with $1 < i < M$, where M is the number of query tokens, the maximum similarity to all the other document tokens is computed and then summed up. The MaxSim between a query Q and a document D is therefore defined as: $\Phi(Q, D) = \sum_{i=1}^{M} w_i$, with $w_i = \max_{j \in \{1,\ldots,N\}} \phi(q_i, d_j)$, and where N is the number of document tokens, d_j is the j-th token of D, and $\phi(\cdot, \cdot)$ is a similarity function.

Evidence The evidence used during the explanation is the subset of similarity values resulting from the *MaxSim* operator. Thus, the evidence is the set of similarity values $e = \{w_i\}_1^M$. Since we can consider the set of similarities as a part of the model weights, the explanatory potential is equal to $pot(e) = N/W$, where W is the total number of model weights. Since N in the small BERT version is 512 and $W > 100,000,000$, we have $pot(e) < 5.12 \cdot 10^{-6}$, where $pot(e)$ is theoretically bounded between 0 and 1.

Interpretation In the common interpretation, the similarity of the query-document token pairs contributing to the summation $\{\max_{j \in \{1,\ldots,M\}} \phi(q_i, d_j)\}$, represents the importance of the term association between q_i and d_j. We can, therefore, rank the most important query and document token for each query.

Evaluation The original paper does not evaluate the explainability of those scores. However, other papers have explored their properties, e.g. [24]. In this case, we highlight that the explanatory potential of those weights is minimal and that the "contextualization" brought after each transformer block might result in a token at the end of the transformer in which there is more "context" than the token itself, as mentioned in the previous section.

4.5. ILMART

In [12], an explainable by-design model for ranking with hand-crafted features has been presented. The authors presented a simple additive model by constraining the well-known LambdaMART algorithm [25] using only one or two features per tree, creating a scoring function that is a sum of univariate or bivariate functions. Formally, the score of a query-document pair (Q, D) is defined as $S(Q, D) = \sum_{i \in \mathcal{M}} \tau_i(D) + \sum_{\{j,k\} \in \mathcal{I}} \tau_{jk}(D)$, where $\mathcal{M}$ is the set of feature, $\mathcal{I}$ is the set of all the possible pair of features, and $\tau_i(D)$ and $\tau_{jk}(D)$ are respectively the univariate and bivariate functions. In addition, to limit the complexity of the function, the author presents a greedy way to limit the number of univariate and bivariate functions.

Evidence The evidence in this type of model considered explainable by design (as others in the literature as NeuralRankGAM [26] or BM25) is the entire model itself, and thus this type of model has full explanatory potential.

Interpretation Given the explainable design, interpreting the evidence is trivial and can be formalized as the identity function since the output of $i(e)$ is the model itself. We highlight that even though the interpretation is considered trivial, the final explanation provided might not be plausible for the user.

Evaluation The authors claim that the model does not need explanation and thus does not provide any evaluation that can be mapped in our framework, but just a measure of the model's effectiveness (similarly to [26]). The plausibility aspect is only mentioned with an anecdotal example of model visualization. In this case, both faithfulness and plausibility evaluation are missing.

5. Conclusion

In this paper, we presented how a theoretical explainability framework presented to be applied in the biomedical domain can also be suitable for the IR field, with a particular focus on the core IR tasks, i.e., search and ranking. We first summarize the framework and then show how a selection of explanation techniques presented in the IR literature can easily fit in the framework. We selected three so-called "post hoc" techniques (including one in the category of explanation based on IR axioms) and two "white box" models and highlighted that the explanation procedure is the same for all the methods considered. All explanations start from evidence and are provided to the user through an interpretation; the main difference is that the interpretation can be convoluted or trivial. We also showed that the evaluation performed in the papers analyzed can always be mapped to one of the two evaluation categories we identified, namely, faithfulness or plausibility. We claim that this unified view can be a starting point to create a common vocabulary for the ExIR field and to allow a better comparison between explanation techniques previously thought to be diametrically opposed, helping to pave the path to a more structured and robust field development.

References

[1] J. Lin, The Neural Hype and Comparisons Against Weak Baselines, ACM SIGIR Forum 52 (2019) 40–51. URL: https://dl.acm.org/doi/10.1145/3308774.3308781. doi:10.1145/3308774.3308781.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[3] S. Robertson, H. Zaragoza, The Probabilistic Relevance Framework: BM25 and Beyond, Foundations and Trends® in Information Retrieval 3 (2009) 333–389. URL: https://www.nowpublishers.com/article/Details/INR-019. doi:10.1561/1500000019, publisher: Now Publishers, Inc.

[4] A. Anand, L. Lyu, M. Idahl, Y. Wang, J. Wallat, Z. Zhang, Explainable Information Retrieval: A Survey, 2022. URL: http://arxiv.org/abs/2211.02405. doi:10.48550/arXiv.2211.02405, arXiv:2211.02405 [cs].

[5] L. Nannini, A. Balayn, A. L. Smith, Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1198–1212. URL: https://doi.org/10.1145/3593013.3594074. doi:10.1145/3593013.3594074.

[6] C. Panigutti, R. Hamon, I. Hupont, D. Fernandez Llorca, D. Fano Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. Soler Garrido, E. Gomez, The role of explainable AI in the context of the AI Act, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1139–1150. URL: https://dl.acm.org/doi/10.1145/3593013.3594069. doi:10.1145/3593013.3594069.

[7] S. Saha, D. Majumdar, M. Mitra, Explainability of Text Processing and Retrieval Methods: A Critical Survey, 2022. URL: http://arxiv.org/abs/2212.07126. doi:10.48550/arXiv.2212.07126, arXiv:2212.07126 [cs].

[8] M. Rizzo, A. Veneri, A. Albarelli, C. Lucchese, M. Nobile, C. Conati, A Theoretical Framework for AI Models Explainability with Application in Biomedicine, in: 2023 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE, Eindhoven, Netherlands, 2023, pp. 1–9. URL: https://ieeexplore.ieee.org/document/10264877. doi:10.1109/CIBCB56990.2023.10264877.

[9] M. Verma, D. Ganguly, LIRME: Locally Interpretable Ranking Model Explanation, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1281–1284. URL: https://doi.org/10.1145/3331184.3331377. doi:10.1145/3331184.3331377.

[10] L. Lyu, A. Anand, Listwise Explanations for Ranking Models Using Multiple Explainers, 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. 653–668. doi:10.1007/978-3-031-28244-7_41.

[11] O. Khattab, M. Zaharia, ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 39–48. URL: https://doi.org/10.1145/3397271.3401075. doi:10.1145/3397271.3401075.

[12] C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, A. Veneri, ILMART: Interpretable

Ranking with Constrained LambdaMART, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 2255–2259. URL: https://dl.acm.org/doi/10.1145/3477495.3531840. doi:10.1145/3477495.3531840.

[13] M. Graziani, L. Dutkiewicz, D. Calvaresi, J. P. Amorim, K. Yordanova, M. Vered, R. Nair, P. H. Abreu, T. Blanke, V. Pulignano, J. O. Prior, L. Lauwaert, W. Reijers, A. Depeursinge, V. Andrearczyk, H. Müller, A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences, Artificial Intelligence Review 56 (2023) 3473–3504. URL: https://doi.org/10.1007/s10462-022-10256-8. doi:10.1007/s10462-022-10256-8.

[14] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu, Definitions, methods, and applications in interpretable machine learning, Proceedings of the National Academy of Sciences 116 (2019) 22071–22080. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1900654116. doi:10.1073/pnas.1900654116, company: National Academy of Sciences Distributor: National Academy of Sciences Institution: National Academy of Sciences Label: National Academy of Sciences Publisher: Proceedings of the National Academy of Sciences.

[15] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016.

[16] J. Vig, A Multiscale Visualization of Attention in the Transformer Model, in: M. R. Costa-jussà, E. Alfonseca (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Florence, Italy, 2019, pp. 37–42. URL: https://aclanthology.org/P19-3007. doi:10.18653/v1/P19-3007.

[17] S. Jain, B. C. Wallace, Attention is not Explanation, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3543–3556. URL: https://aclanthology.org/N19-1357. doi:10.18653/v1/N19-1357.

[18] S. Wiegreffe, Y. Pinter, Attention is not not Explanation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 11–20. URL: https://aclanthology.org/D19-1002. doi:10.18653/v1/D19-1002.

[19] S. Serrano, N. A. Smith, Is Attention Interpretable?, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2931–2951. URL: https://aclanthology.org/P19-1282. doi:10.18653/v1/P19-1282.

[20] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, Curran Associates Inc., Red Hook, NY, USA, 2018, pp. 9525–9536.

[21] A. Câmara, C. Hauff, Diagnosing BERT with Retrieval Heuristics, in: J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, F. Martins (Eds.), Advances in Information Retrieval, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2020, pp. 605–618. doi:10.1007/978-3-030-45439-5_40.

[22] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, 2016, pp. 1135–1144. URL: https://dl.acm.org/doi/10.1145/2939672.2939778. doi:10.1145/2939672.2939778.

[23] H. Fang, C. Zhai, An exploration of axiomatic approaches to information retrieval, in: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, Salvador Brazil, 2005, pp. 480–487. URL: https://dl.acm.org/doi/10.1145/1076034.1076116. doi:10.1145/1076034.1076116.

[24] T. Formal, B. Piwowarski, S. Clinchant, A White Box Analysis of ColBERT, in: D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2021, pp. 257–263. doi:10.1007/978-3-030-72240-1_23.

[25] C. J. Burges, From ranknet to lambdarank to lambdamart: An overview, Learning 11 (2010) 81.

[26] H. Zhuang, X. Wang, M. Bendersky, A. Grushetsky, Y. Wu, P. Mitrichev, E. Sterling, N. Bell, W. Ravina, H. Qian, Interpretable Learning-to-Rank with Generalized Additive Models, arXiv:2005.02553 [cs, stat] (2020). URL: http://arxiv.org/abs/2005.02553, arXiv: 2005.02553.