Discovery: an Explainable AI Powered Academic Search Engine for Knowledge Workers

Robin Weitemeyer[1], Jun Ma[2], Yu Cao[2], Sinan Sen[2], Jens Beyer[3], and Lena Kölmel[4]

[1] Institut für Lernen und Innovation in Netzwerken, Karlsruhe University of Applied Sciences
robin.weitemeyer@h-ka.de
[2] Datalyxt GmbH
{jun.ma, yu.cao, sinan.sen}@datalyxt.com
[3] LAVRIO.solutions GmbH
jens@lavrio.solutions
[4] Institut für Arbeitswissenschaft und Betriebsorganisation, Karlsruher Institute for Technology
lena.koelmel@kit.edu

Abstract. In recent years, transformer models were able to achieve astonishing results in various natural language processing (NLP) tasks. Especially with the rise of ChatGPT and the countless alternatives following its triumphal march, using large language models (LLM) for information retrieval has established itself in both the private, work and research context. For scientific search however, semantic analysis with LLMs is an underutilized tool for enhancing the work process of literature research. We therefore propose an academic search engine called Discovery, which uses BERT to semantically analyze arbitrary text queries in order to recommend fitting search results for scientific publications. Through explainable artificial intelligence (XAI), additional information about the AI output is provided to the user with the goal to decrease the time needed for evaluating the suitability of a recommended paper.

Keywords: Large Language Model, Transformer, Knowledge Work, Academic Search Engine, Semantic Search, Explainable Artificial Intelligence

1 Introduction

Knowledge workers became valuable assets for today's organizations as forecast by Drucker [1]. By applying their expert knowledge to tasks such as general problem solving and developing of new products, they accelerate progress in both science and technology [2]. It is thereby in the interest of every company to maximize the productivity of its knowledge workers.

A knowledge task where we see major potential for efficiency improvement is the acquirement of new knowledge to keep up with new technical and scientific advances. The International Data Corporation (IDC) estimated in 2001 that a knowledge worker spends "roughly 30% of the workday searching for information" [3]. In 2012, Borlund et al. stated that "prior knowledge or no knowledge of the topic in question do also influence searching time" [4]. In addition, Toms et al. found out in 2013 that two-third of the time needed for finding relevant information "was mainly spent reviewing documents that had already been retrieved" [5]. We investigated those insights further and developed a tool, which aims to significantly decrease the time spend on searching for information by providing meaningful recommendations for relevant literature. Due to their state-of-the-art performance in natural language processing (NLP) tasks [6], we believed that a

search engine utilizing transformer-based AI, i.e. large language models (LLM), would be suitable for accomplishing this. Therefore, we introduce our novel explainable AI powered academic search engine called Discovery. However, our goal was not solely to map a search query to helpful recommendations, but also to provide knowledge workers with additional support for finding information on a topic of which the user has little to no prior knowledge and for quickly evaluating the usefulness of the recommended literature. Both of this present additional difficulties during the search for information as stated by Borlund et al. [4] and Toms et al. [5] and would be crucial for increasing the efficiency of the task of literature research. Especially regarding recommendation evaluation, we argue that explaining the decisions of an AI application with methods of explainable artificial intelligence (XAI) is a critical addition.

In the following we give an overview on related work in Section 2, looking at technologies we deployed and introducing existing LLMs-based search engines. In Section 3, we describe the methodology with which we performed the requirements analysis and implemented our search engine approach. Concluding, we discuss our achievements and give a look into future works.

2 Related Work

Transformer-based AI models have been successfully deployed on a wide range of NLP tasks like machine translation, sentiment analysis, question answering, language modelling and named entity recognition [7]. This is also a result of efforts on pre-training language models on large text datasets. Two of the most popular transformer model families are Bidirectional Encoder Representations from Transformers (BERT) from Google [8] and OpenAI's Generative Pre-trained Transformers (GPT) [9]. Although both achieve outstanding results on NLP benchmarks, there are significant differences between those two model types. BERT as well as GPT are based on the original transformer architecture [10] which follows an encoder-decoder structure. However, BERT is an bidirectional encoder that processes the input in both directions and therefore considers the left and right context of each word. In contrast, GPT is an autoregressive decoder, which predicts the next word only depending on its left context. It is thus arguable that BERT tends to be better in understanding the overall context of a text and therefore learn meaningful representations. GPT on the other hand, excels at generating text to a given input [11]. Thou there have been introduced several different variants for BERT and GPT with their own advances [12–14], traditionally speaking BERT performs well for natural language understanding (NLU) and GPT for natural language generation (NLG) [15].

Since 2001, a significant rise can be observed in the number of releases of new LLM models [15] indicating an increased interest and usage of such AI models. Thus, it is unsurprising that there exist several commercial solutions for literature research utilizing LLMs. The academic search engines Consensus[5], Elicit[6], Scispace[7] and Zeta Alpha[8] use transformers for question answering and summarization of abstracts or insights of several scientific publications. In addition, they also list fitting literature for a search query. However, not all of those search engines leverage the full potential of LLMs by not at all or not entirely replacing a keyword-based search with a vector search, i.e. a semantic search. For example, Consensus performs their search in multiple steps. First, they run

[5] Consensus, https://consensus.app/, October 2023
[6] Elicit, https://elicit.com/, October 2023
[7] Scispace, https://typeset.io/, October 2023
[8] Zeta Alpha, https://www.zeta-alpha.com/, October 2023

a simple keyword-based search on the paper database. Only afterwards, a vector search is run on the reduced dataset. This artificially narrows down the extent of the semantic search.

One feature that all of these search engines lack is the explanation of their AI recommendations. By summarizing the content of relevant papers, they potentially reduce the amount of information that has to be review by the user. However, it provides no insight into why the AI model estimates the recommended papers to be relevant for a search query. XAI can be utilized to generate such insight into the AI models behaviour and logic [16]. Danilevsky et al. [17] give an overview of XAI methods and visualization techniques for NLP. However, in the case of transformer models, generating good explanations still poses a challenging task due to their increasing size and complex architecture [18].

3 Methodology

This section will give an overview of the methods used for requirements analysis, computing the literature recommendations and explaining the AI output for better understanding why a certain scientific publication is recommended.

3.1 Requirements Analysis

For requirements analysis, we conducted interviews with research associates of different experience and disciplines to gain more insight in their search strategies as well as most common pain points while conduction literature research. We interviewed six knowledge workers with an average age of 27 years and of whom three were female (standard deviation = 4.6). By selecting the interview partners, we aimed for a diverse distribution in terms of research type (basic research (n = 3) or applied research (n = 3)), scientific working methods (empirical (n = 4) or non-empirical (n = 2)), professional orientation (human (n = 1), technical (n = 2) or organizational (n = 3)) and current career stage (graduate (n = 1), PhD student (n = 4) or post-doctoral (n = 1)). Participants had only superficial knowledge about AI-related topics and did not work in an AI-related field. On average, the interviews took 71 minutes and were conducted as semi-structured interviews divided into four segments. First, we asked about the procedure of a prototypical literature research including details about the used literature databases, the formation of search words and the documentation of results. In the next segment, we asked about other search strategies aside from the prototypical literature research. We wanted to know which kind of different research scenarios exist and what changes regarding the prototypical procedure those entail. Next, we asked about pain points while conducting literature research and their causes. Lastly, we asked the participants to describe the perfect literature research support system, without any barriers due to technical feasibility. We asked about useful features and what kind of pain points could be solved by implementing those.

To analyse the data we transcribed the interviews from speech to text via f4x. Due to the diverse answers between the different interview partners, we decided against a standardized qualitative analysis and instead isolated the most relevant text passages, summarized them into key aspects and finally grouped them into semantically similar clusters. This procedure resulted in a suitable database from which we derived empirically substantiated system requirements. Participants reported that they usually follow an unstructured procedure while conducting a literature research. They adapt their strategies

along the way, depending on the success rate of e.g. the prior search words. They could differentiate between different literature research scenarios (e.g. planning of an experiment, preparation of a manuscript, identifying interesting research gaps, informing oneself about the status quo regarding a topic, learning about new methods or theories, identifying hot topics or looking for literature on behalf of third parties). None of our interview partners identified conscious adaptions to the prototypical procedure depending on the specific research scenario. Most participants reported that during the search process, they mostly only screen title and abstract, full text reading is a subsequent process. Further, the analysis of the interviews led to the following key insights:

- Searching for literature on an interdisciplinary topic or topics in general about which the interviewee has little prior knowledge is stated as one of the biggest pain point.
- Additional information about the search results is requested in order to speed up the process of the immediate evaluation whether or not a recommended paper should be considered for a more thorough inspection.

We then derived three key requirements from the interview results for the listed insights:

1. The search engine should be able to perform a semantic search on continuous text written in natural language so that the user has the possibility to only describe what to search for instead of knowing the exact terminology.
2. The search engine should be able to recommend terms which are related to the search query but possibly unknown to the user since they may be primarily used in different disciplines.
3. The search engine should provide additional information to the user about why the deployed AI considers the recommendations fitting to the search query.

Based on these requirements, the first prototype of our novel AI powered academic search engine was developed. The use of LLMs to compute recommendations addresses the first two requirements while the application of XAI aims to generate the desired additional information of the third.

3.2 Semantic Sentence Similarity

To meet the key requirement 1 we decided to utilize LLMs for replacing a keyword-based search with a semantic search [19]. For this, a transformer model should be deployed to learn meaningful sentence embeddings for a semantic search on research literature by computing the semantic similarity between those vector representations. Recommending publications based on this semantic sentence similarity approach allows for a more intuitive use of our search engine by knowledge workers of different experience. It is no longer required to carefully construct an effective search string with precise combinations of correct terms and their synonyms in order to receive helpful recommendations. Instead, our AI powered academic search engine Discovery can process arbitrary text written in natural language and automatically consider related terms. Thus, making it possible to provide a rather descriptive search query. In Discovery, the literature recommendations are determined by calculating the cosine similarity between the sentence embeddings of the search text and the titles and abstracts of the papers.

Due to BERT's suitability for NLU tasks, we implemented Discovery using the SentenceTransformer[9] Python framework which originates from the Sentence-BERT (sBERT)

[9] SentenceTransformer, https://www.sbert.net/, October 2023

network [20]. sBERT builds on the base BERT model and fine-tunes it in a Siamese network architecture. This extension allows for better performances when combined with similarity measures such as the cosine similarity. For our approach, DistilBERT [13] was chosen as the pre-trained BERT model, which was then fine-tuned in a Siamese network on the online contrastive loss with the cosine distance as the similarity measure. Our own database for the search engine and model training consists of over 350.000 publications from IEEE, Springer and Scopus containing publicly available metadata, e.g. title, abstract and keywords. To start with, we confined the database on the topic of AI to make the required data size more manageable. Since there was no ground truth for the similarity between the data points and no reference search queries, we used the self-supervised training method contrastive learning [21]. The training task was to learn sentence embeddings such that the title and abstract of a paper have a high similarity. For positive samples, we kept the original title and abstract pair. For negative samples, we shuffled the titles and abstracts to receive not only incorrect title-abstract pairs but also random title-title and abstract-abstract combinations. The resulting positive and negative samples had a ratio of 1:3. Although, in the face of computation constraints, we were forced to limit our training data to a subset of 150,000 data points. As expected of the contrastive loss, the positive samples were assigned a label of 1, which reduces the distance between the two text inputs, and 0 for the negative samples, increasing the distance. The network used for fine-tuning is shown in Figure 1. For dimensionality reduction, we introduced two fully connected layers with dimensions of 256 and 128 along with Tanh activation functions. This reduces computational complexity while still providing a high enough dimensional vector for our model to effectively capture semantic nuances in a more compact space. We fine-tuned our entire model over five training epochs utilizing the AdamW optimizer with a learning rate of 2e-05 and weight decay of 0.01 to ensure model stability and convergence. The training took a total of 7 hours and 31 minutes on Google Colab using the T4 GPU, highlighting the efficiency and effectiveness of our training pipeline. This efficient use of computational resources underscores the practicality and scalability of our approach.

In order to fulfil the key requirement 2, we introduced an additional functionality to Discovery: term recommendation. The recommended terms are selected from a dataset, which we create by aggregating the keywords over all papers in our database. Since related terms occur in semantically similar texts, the same model can be used which we fine-tuned for the task of paper recommendation. The difference is that instead of searching over the titles and abstracts, we compute the sentence embeddings of the keywords with our BERT-based model and compare them to the input string. The user may submit a search request for either a single term or a descriptive text. Recommending related terms can aid users with limited or no prior knowledge in their search while also introducing them to the appropriate terminology, thereby creating new knowledge.

3.3 Explainable Artificial Intelligence

Key requirement 3 has the intent to decrease the time a knowledge worker needs to evaluate the literature recommendations and to select the set of suitable papers for further consideration. We argue that providing additional explanations about why the AI model recommended the exact papers can improve the user's decision making [22]. However, comprehending the intrinsic logic of LLMs presents multiple challenges. First and foremost is their black box characteristic [23] due to their complexity. With hundreds of millions to billions of parameters, understanding the intricate relationships and dependencies within the model becomes an impossible task without additional explanations.

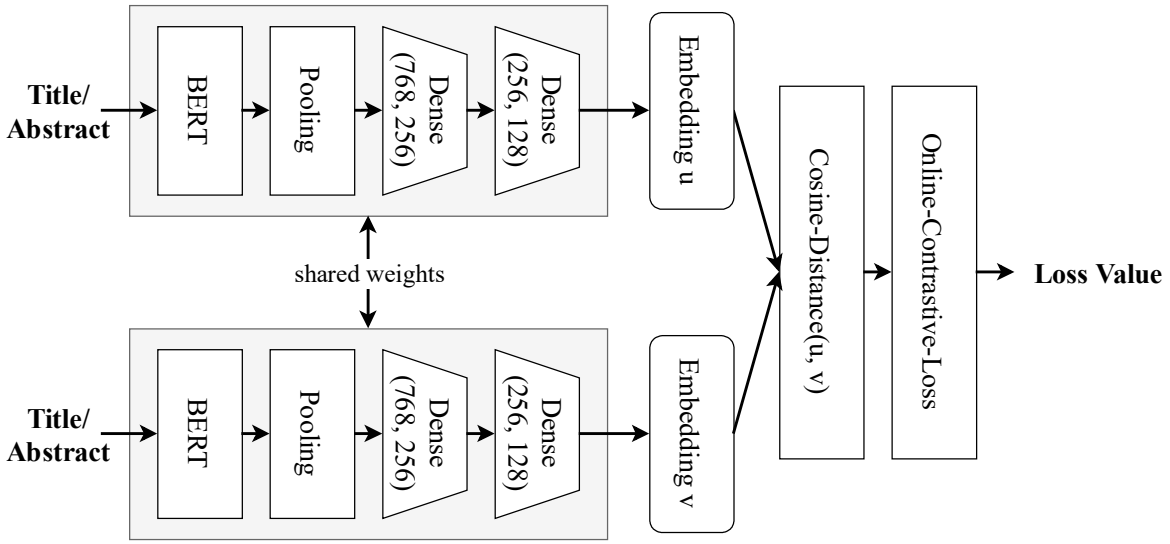

Fig. 1. Visualization of the model architecture used for computing the embeddings between which the consine similarity is calcutated. The model was trained in a Siamese network architecture using constrastive learning. The original title and abstract pairs were used as positive samples, while titles and abstracts were shuffled to synthesise negative samples.

The non-linearity introduced by deep architectures with non-linear activation functions further obscures straightforward interpretations, making it difficult to discern any direct input-output relationships. Additionally, these models are inherently data-driven, meaning their behaviour is shaped by vast and varied training data. This data-driven nature can embed subtle nuances, biases, and patterns from the training data into the model, many of which may not be immediately apparent or interpretable.

In contrast to other use cases of XAI, we do not perform a classification task on the scientific papers but instead measure the semantic similarity between two text fragments. Python packages implementing the XAI methods like LIME[10] [24] or SHAP[11] [25] are highly focused on these more usual use cases of XAI. Thus, we implemented our own XAI algorithm, which was inspired by the visualization of LIME explanations on text data.

For the purpose of our academic search engine, we want to explain to the user how the words in the search string influenced the similarity measurement between the user input and the recommended papers. Hence, we do not want to globally explain the LLM, but instead generate local explanations based on word importance. The main idea of our algorithm follows a similar concept as the one proposed by Landthaler et al. [26] for semantic text matching. The XAI algorithm leaves one word out of the input and then recalculates the similarity score for the top N search results. For each paper, the magnitude of the change in similarity correlates to the importance attributed to the left out word. This step is repeated for each word in the user input. We visualize the word importance by colouring the words of the search query according to their impact on the similarity score. An example for the visualization of our XAI method is shown in Figure 2.

To begin with, we wanted to keep the explanation as simple as possible to not overwhelm the user. Hence, the word importance is only calculated for the user input. How-

[10] LIME Python Package, https://github.com/marcotcr/lime, October 2023
[11] SHAP Python Package, https://github.com/shap/shap, October 2023

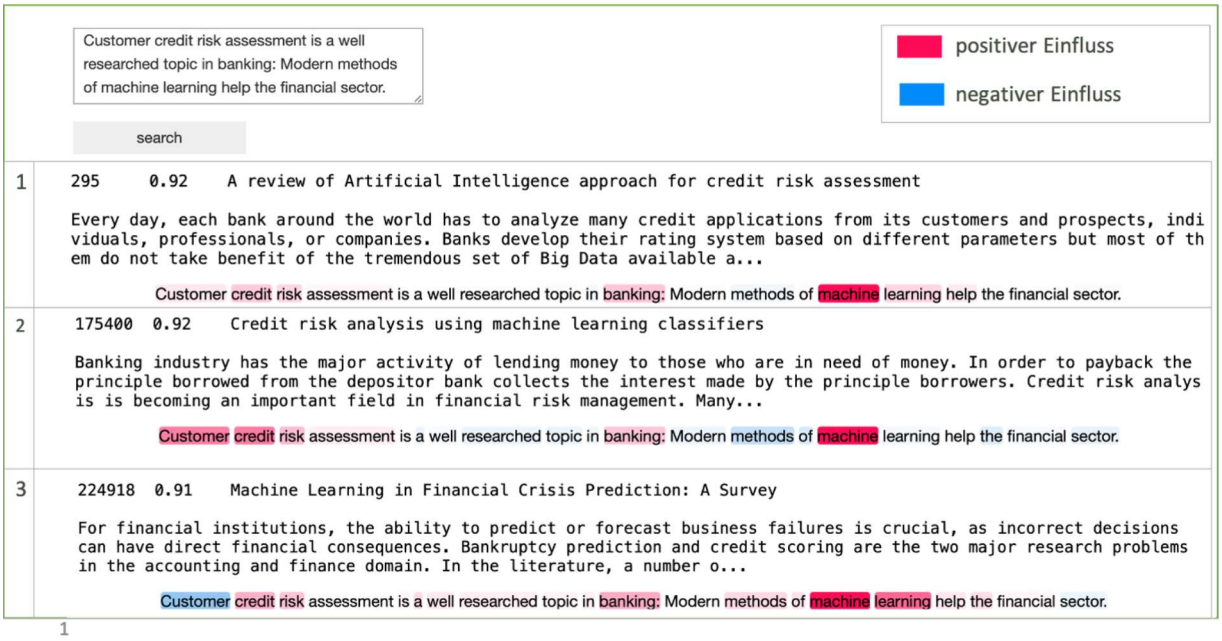

Fig. 2. Visualization of the custom XAI method for a search query example. Red depicts positive and blue negative word importance.

ever, our algorithm can easily be modified to also consider the title and abstract of the recommended papers. Our intention is to gradually increase the complexity of the explanations guided by user feedback.

4 Conclusions

We presented the working prototype of our novel explainable AI powered academic search engine Discovery for recommending literature and related terms to knowledge workers based on semantic sentence similarity. It was implemented using state-of-the-art LLM and XAI techniques by successfully deploying a pre-trained BERT-based model, which was fine-tuned on our own database with contrastive learning in a Siamese network architecture, for the task of measuring sentence similarity. Through the integration of our own XAI method, Discovery is further capable of providing an explanation for every recommended paper to the users based on the word importance of their search queries. Our search engine was developed for three key requirements, which were derived from the results of our requirements analysis consisting of a literature research and interviews with probands of our target group. During so, we could identify significant potential for improving the efficiency of the task of acquiring new knowledge through literature research.

Our approach fully replaces the traditional keyword-based search with a vector search thereby setting Discovery apart from established academic search engines like Google Scholar or Semantic Scholar. Further, even in comparison to other transformer-based search engines, as those discussed in Section 2, we introduce additional features by recommending related terms and explaining the model output through XAI. Other search engines utilize LLMs to summarize abstracts or extract insights from scientific papers. We see this generative approach critical in the context of literature research. For example, ChatGPT has been proven to lack accuracy providing scientific facts [27]. In addition,

it is stated in the FAQ of Elicit that the users should assume that only 90% of the information they see is accurate. Further, summarizing and extracting the insights could potentially mislead the knowledge worker into relying only on the generated information, rather than considering the recommended papers in more detail. Our XAI approach is designed to accelerate the immediate assessment of the recommendations, but not to make the more thorough review of the papers obsolete.

The eligibility of the recommendations provided by Discovery has been initially tested and shows promising results, however, more extensive evaluation of the system is required through further human feedback and comparison with suitable benchmarks [28]. Furthermore, first feedback indicates that the word importance is not satisfying the need for explanation and that a topic-based explanation is assumed to be more effective. Thus, additional XAI approaches have to be implemented and evaluated on the target group.

Acknowledgements

This research and development project is/was funded by the German Federal Ministry of Education and Research (BMBF) within the "The Future of Value Creation – Research on Production, Services and Work" program (funding number: 02L19C250) and managed by the Project Management Agency Karlsruhe (PTKA). The author is responsible for the content of this publication.

References

1. Drucker, P.F.: Landmarks of Tomorrow: A Report on the New "Post-Modern" World. Routledge (1996)
2. Powell, Walter W.; Snellman, K.: The knowledge economy. Annu. Rev. Sociol. (Annual Review of Sociology) (2004) 199–220
3. Feldman, S., Sherman, C.: The high cost of not finding information: An idc white paper. KMWorld Magazin (2001) 1–10
4. Borlund, P., Deier, S., Bystroem, K.: What does time spent on searching indicate? In: Proceedings of the 4th information interaction in context symposium. (2012) 184–193
5. Toms, E.G., Villa, R., McCay-Peet, L.: How is a search system used in work task completion? Journal of Information Science **39**(1) (2013) 15–25
6. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (2020) 28–45
7. Lin, T., Wang, Y., Liu, X., Qiu, X.: A survey of transformers. AI Open **3** (2022) 111–132
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT. Volume 1. (2019) 4171–4186
9. Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018)
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). (2017)
11. Bosley, M., Jacobs-Harukawa, M., Licht, H., Hoyle, A.: Bosley, mitchell, et al. "do we still need bert in the age of gpt? comparing the benefits of domain-adaptation and in-context-learning approaches to using llms for political science research (2023)

12. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019)
13. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (2020)
14. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., eds.: Advances in Neural Information Processing Systems. Volume 33., Curran Associates, Inc. (2020) 1877–1901
15. Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., Mian, A.: A comprehensive overview of large language models (2023)
16. Gunning, D., Aha, D.: Darpa's explainable artificial intelligence (xai) program. AIMag (AI Magazine) **40**(2) (2019) 44–58
17. Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., Sen, P.: A survey of the state of explainable ai for natural language processing. In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. (2020) 447–459
18. Deb, M., Deiseroth, B., Weinbach, S., Schramowski, P., Kersting, K.: Atman: Understanding transformer predictions through memory efficient attention manipulation (2023)
19. Jimmy Lin, Rodrigo Nogueira, A.Y.: Pretrained Transformers for Text Ranking: BERT and Beyond. Synthesis Lectures on Human Language Technologies. Springer Cham (2022)
20. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processingand the 9th International Joint Conference on Natural Language Processing. (2019) 3982–3992
21. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Volume 2. (2006) 1735–1742
22. Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., Ranjan, R.: Explainable ai (xai): Core ideas, techniques, and solutions. ACM Computing Surveys **55**(9) (2023) 1–33
23. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access **6** (2018) 52138–52160
24. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you? In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. (2016) 1135–1144
25. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., eds.: Advances in Neural Information Processing Systems. Volume 30., Curran Associates, Inc. (2017)
26. Landthaler, J., Glaser, I., Matthes, F.: Towards explainable semantic text matching. Proceedings of Jurix 2018: International Conference on Legal Knowledge and Information Systems (2018)
27. Borji, A.: A categorical archive of chatgpt failures (2023)
28. Kalyan, K.S., Rajasekharan, A., Sangeetha, S.: Ammus : A survey of transformer-based pretrained models in natural language processing. (2021)