Zero-Shot Medical Information Retrieval via Knowledge Graph Embedding

Yuqi Wang[1,3], Zeqiang Wang[1], Wei Wang[1(✉)] , Qi Chen[1], Kaizhu Huang[2] ,
Anh Nguyen[3] , and Suparna De[4]

[1] Xi'an Jiaotong-Liverpool University, Suzhou, China
{yuqi.wang17,zeqiang.wang22}@student.xjtlu.edu.cn,
{wei.wang03,qi.chen02}@xjtlu.edu.cn
[2] Duke Kunshan University, Kunshan, China
kaizhu.huang@dukekunshan.edu.cn
[3] University of Liverpool, Liverpool, UK
anh.nguyen@liverpool.ac.uk
[4] University of Surrey, Surrey, UK
s.de@surrey.ac.uk

Abstract. In the era of the Internet of Things (IoT), the retrieval of relevant medical information has become essential for efficient clinical decision-making. This paper introduces MedFusionRank, a novel approach to zero-shot medical information retrieval (MIR) that combines the strengths of pre-trained language models and statistical methods while addressing their limitations. The proposed approach leverages a pre-trained BERT-style model to extract compact yet informative keywords. These keywords are then enriched with domain knowledge by linking them to conceptual entities within a medical knowledge graph. Experimental evaluations on medical datasets demonstrate MedFusionRank's superior performance over existing methods, with promising results with a variety of evaluation metrics. MedFusionRank demonstrates efficacy in retrieving relevant information, even from short or single-term queries.

Keywords: medical information retrieval · Internet of Things · natural language processing · clinical decision-making · medical knowledge graph

1 Introduction

The widespread adoption of the Internet of Things (IoT) has enabled the collection of large amounts of medical text data. By using IoT to identify patients, transfer information to central databases, and search for relevant medical texts such as electronic health records (EHRs) and disease-related papers, we can improve the efficiency of treatment procedures and therapeutic outcomes [8,19]. For instance, the MIMIC-III [12] and MIMIC-IV [11] critical care medical databases use IoT systems to collect structured clinical data and texts. These

medical texts have become the foundation for medical natural language processing, serving as corpora for pre-training large language models and embeddings [1,16,32]. Additionally, the use of IoT in healthcare has the potential to revolutionise patient care by providing real-time monitoring and personalised treatment plans based on individual patient data. This can lead to improved patient outcomes and reduced healthcare costs [7].

A key challenge in healthcare is enabling real-time, personalised clinical decision-making beyond traditional tasks like diagnostic classification and outcome prediction. Effective clinical decision support fundamentally relies on the ability to retrieve relevant information from massive amounts of unstructured EHR data. While earlier work in medical information retrieval relied on statistical methods like BM25 [23] with Term Frequency-Inverse Document Frequency (TF-IDF) features, these techniques struggled with the complexity and sparsity of medical text. Medical notes exhibit pervasive synonym phenomena, with different terms like "*hypertension*" and "*high blood pressure*" denoting identical concepts. Abbreviations and shorthand introductions are also ubiquitous, posing difficulties for simple lexical matching.

Recently, pre-trained large language models (LLMs) like BERT [6], Alpaca [27], and Llama [29] have shown promise by learning generalisable representations of medical language. However, their computational overhead makes deployment directly onto resource-constrained IoT devices impractical. Training with massive LLMs requires substantial data, computing power, and memory exceeding the available on-device. Therefore, an open challenge is adapting the strengths of LLMs for medical search on embedded IoT systems. More efficient methods are needed to extract knowledge from LLMs and make it accessible for medical information retrieval on hardware-friendly architectures.

To address the aforementioned challenges, we propose a novel zero-shot information retrieval approach that integrates the strengths of statistical methods and pre-trained LLMs while mitigating their limitations. Our key insight is to leverage a pre-trained BERT-style model to extract compact yet informative keywords. These keywords are then enriched with domain knowledge by linking them to conceptual entities within a medical knowledge graph. Our method has demonstrated promising results on two benchmark datasets, outperforming a range of existing Information Retrieval models across various evaluation metrics.

2   Related Work

Medical information retrieval (MIR) aims to retrieve relevant medical data from sources such as EHR. However, it faces distinct challenges that extend beyond conventional information retrieval (IR) - complex medical terminology, heterogeneous data, privacy constraints, and difficulties in system evaluation. While leveraging core IR techniques, MIR has specific requirements arising from the medical domain. In this section, we provide an overview of key IR methods that facilitate effective MIR.

2.1   Statistical Information Retrieval

Statistical information retrieval (Statistical IR) is a foundational approach that leverages probabilistic and statistical models to quantify the relevance of documents to user queries. This allows ranking search results by estimated relevance based on mathematical models. Popular statistical IR techniques, including vector space model [3], probabilistic retrieval model [25], and Okapi BM25 [23] rely heavily on weighted keyword matching between query and document terms. They estimate relevance using statistical signals like TF-IDF, and length normalisation. While very effective for many search tasks, these lexical similarity models have limitations. Specifically, they cannot account for semantic matching, failing to recognise synonyms and antonyms.

2.2   Neural Information Retrieval

Neural information retrieval (Neural IR) is a modern paradigm that leverages neural networks and deep learning techniques to overcome the limitations of statistical IR models. Neural IR models can be classified into two main types: first-stage retrieval methods and re-ranking methods.

First-Stage Methods. First-stage methods aim to directly retrieve relevant documents from a large collection using neural networks. These methods can be further categorised into sparse retrieval methods and dense retrieval methods. Sparse retrieval methods use sparse word representations, such as bag-of-words or TF-IDF, as inputs to neural networks and learn to rank documents based on their similarity to queries [5,15]. Dense retrieval methods, on the other hand, use dense vector representations, such as word embeddings or contextual embeddings, as inputs to neural networks and learn to map queries and documents into a common semantic space where their relevance can be measured by distance metrics [10,14,24].

Re-ranking Methods. Re-ranking methods use neural networks to refine the initial ranking results produced by a base retriever, such as BM25 or a sparse/dense retriever. These methods can be categorised into two main approaches: 1)Re-ranking with sentence embeddings: These methods treat each document independently as an instance and learn to score its relevance to the query [22]. They derive vector representations for the query and each document in a separate manner, compare their embeddings and assign relevance scores. 2) Re-ranking using a cross-encoder: These methods consider each query-document pair as an instance and learn to compare their relative relevance [31]. The cross-encoder jointly models the query and document to capture semantic matching.

3   Methodology

We show the overall architecture of our proposed method in Fig. 1. Specifically, it first extracts keywords from medical documents to capture semantic context. Then, medical embeddings for each keyword are constructed based on the
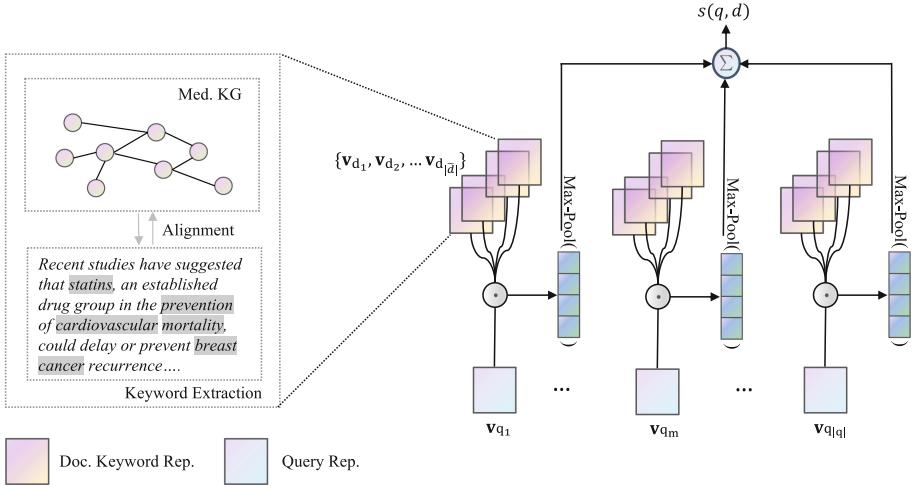
Fig. 1. The overall architecture of our proposed method.

domain-specific knowledge graph. The query and document keywords are compared in the medical embedding space and their similarity scores are aggregated to identify relevant information across query terms for retrieval.

3.1   Document Keyword Extraction

Given the inherent complexity of documents within the medical domain, often encompassing multiple aspects, the necessity of pre-processing before conducting IR becomes evident. One such approach involves the extraction of keywords that aptly describe and summarise the content. By utilising a contextualised attention-based pre-trained language model, the contextual information can be effectively harnessed to discern the document's relatively significant sections. Therefore, we utilise the RoBERTa [18] model for the initial encoding of the corpus documents. RoBERTa is a state-of-the-art language model that has demonstrated exceptional performance in various natural language processing tasks. Specifically, when dealing with a document d comprised of k words, denoted as $d = \{d_1, ...d_k\}$, we leverage the RoBERTa encoding function, $f(\cdot; \theta)$, to transform all the words into a coherent and meaningful semantic space, i.e.

$$\{\mathbf{h}_{<s>}, \mathbf{h}_{d_1}, ...\mathbf{h}_{d_k}, \mathbf{h}_{</s>}\} = f(\{<s>, d_1, ...d_k, </s>\}; \theta) \qquad (1)$$

where $\mathbf{h}_{d_i}$ is the representation of the i-th word in RoBERTa embedding space. $<s>$ and $</s>$ are two special tokens indicating the start and the end positions in the document, respectively. This process enables us to capture the intricate contextual relationships and nuances present within the document.

The comprehensive essence of the document is commonly encapsulated within the hidden state of the special token $<s>$; in order to estimate the significance of individual words within the document, we compute the cosine similarity

between the representation of the special token $<s>$ and the representation of each word. We take the top K ranking words based on their similarity scores, and extract those as the key keywords for the document d. This process is articulated as follows:

$$\tilde{d} = \underset{d_i \in d}{\text{top}} K \left[\text{Sim} \left(\mathbf{h}_{d_i}, \mathbf{h}_{<s>} \right) \right] \tag{2}$$

where $\tilde{d}$ is the keyword set for document d, $\text{Sim}(\cdot)$ is the cosine similarity function. Based on our observation, the top 20 keywords can effectively capture the core semantic content of a document. Hence, we set the number of extracted keywords (K) to 20.

3.2   Medical Embedding Construction

In our work, the challenge posed by zero-shot IR is significant, primarily due to the absence of any prior exposure of the model to the medical domain. In this case, a crucial approach involves enhancing each keyword in the keyword set $\tilde{d}$ with relevant background information. This enrichment encompasses additional context, definitions, and pertinent details sourced from the medical field. In this endeavour, the Medical Subject Headings (MeSH) [17] knowledge graph emerges as an exceptional resource. MeSH is a meticulously structured and high-quality knowledge graph that encompasses a vast spectrum of medical concepts along with their relationships. For instance, the relation "*treatment*" connects the two concepts "*cancer*" and "*chemotherapy*". This indicates that chemotherapy is a type of treatment commonly used for cancer patients.

   To harness the knowledge from MeSH, a method called Node2Vec [9] can be used to generate medical embeddings. The main idea is to treat this graph as a network, where nodes are concepts and edges represent relationships between concepts [32]. This method utilises random walks and learns latent representations of nodes that maximise the probability of the sampled walks. The objective function J for constructing the medical embeddings can be written as follows:

$$J = \max \left[\frac{1}{T} \sum_{i=1}^{T} \sum_{v_j \in \mathcal{C}(v_i)} \log p \left(v_j \mid v_i \right) \right] \tag{3}$$

where T is the number of the MeSH concepts and $\mathcal{C}(v_i)$ is a set containing surrounding words of v_i based on random walks in the knowledge graph. For this study, alignment between the keyword set $\tilde{d}$, the query q, and concepts in the MeSH knowledge graph were performed by matching keywords with concept names. This simple lexical approach to entity linking was chosen for its simplicity. However, it has known limitations, such as ambiguity and lack of semantic matching. Future work should explore more sophisticated techniques to deal with the issue.

3.3 Retrieval with Medical Knowledge

By acquiring all the medical embeddings for document keywords from a corpus in the MeSH knowledge graph embedding space through an offline process, we can retrieve relevant information for each word from a given human-generated query in an efficient manner. In particular, each query term can focus on each word in the document to identify the most relevant information in the document that can be retrieved by that specific query word. We aggregate all the relevance scores for each query term during the retrieval process, i.e.

$$s(q,d) = \sum_{i=1}^{|q|} \max_{j=1}^{|\tilde{d}|} \left[\mathbf{v}_{q_i} \odot \mathbf{v}_{d_j} \right] \tag{4}$$

where $|q|$ and $|\tilde{d}|$ are the number of words in the query and document keyword set, respectively. $\odot$ is the dot product operation symbol. $\mathbf{v}_{q_i}$ and $\mathbf{v}_{d_j}$ are corresponding medical embeddings for the i-th word in the query and j-th word in the document keyword set.

One clear limitation of Retrieval with Medical Knowledge is the equal weighting given to documents whose keyword sets contain query terms, regardless of term frequency. Despite the inclusion of background knowledge corresponding to each word in the document's keywords, factors such as term frequency should also be considered. BM25 [23] is a commonly used unsupervised ranking function, incorporating lexical aspects and statistical information to improve scoring. Leveraging medical embeddings enables the retrieval of candidate-relevant documents while applying BM25, which can further refine the ranking of those initial results by incorporating term frequency statistics. Therefore, we propose fusing the scores yielded by both approaches to improve overall performance, i.e.

$$\hat{s}(q,d) = \begin{cases} s(q,d) + s'(q,d) & \exists s'(q,d) \\ s(q,d) & \nexists s'(q,d) \end{cases} \tag{5}$$

where $s'(q,d)$ represents the BM25 score assigned to a given query q and document d. $\hat{s}(q,d)$ is the final score after the fusion.

4 Results and Evaluation

We evaluated the performance of our proposed models on two medical datasets: NFCorpus [2] and SCIFACT [30]. Both focus on retrieving medical abstracts relevant to search queries. The abstracts are written in technical medical terminology, mostly from PubMed. For each dataset, a range of metrics, including Mean Reciprocal Rank (MRR), Precision, normalised Discounted Cumulative Gain (nDCG), Precision (P) and Recall (R), was employed for a thorough evaluation. Our model was compared against several first-stage retrievers and BM25-based re-rankers to assess its effectiveness.

4.1   Baseline Models

First-Stage Retrievers

– **BioLinkBERT** [13] and **S-BERT** [22]: These are two BERT-based models that generate sentence embeddings using siamese networks. While S-BERT was pre-trained on a general domain question-answering dataset to create universal semantic embeddings, BioLinkBERT utilises contrastive learning on medical texts from PubMed to produce embeddings specialised for the medical domain.
– **DocT5Query** [20]: It leverages a pre-trained T5 [21] model to generate synthetic queries based on the document for text enrichment before indexing.
– **DeepCT** [4]: It employs the BERT model to estimate the weight of each word in the context of the document. These BERT-derived weights are then used to modify the term frequencies of the words.
– **BM25** [23]: It is a traditional unsupervised ranking function. The basic idea is that a more relevant document will contain more of the query terms, and multiple occurrences of a term can indicate higher relevance.

BM25-Based Re-Rankers

– **S-BERT** [22]: We used the same S-BERT model as described previously to re-rank the top 100 candidate documents retrieved in the first-stage for each query.
– **Cross Encoder** [31]: It passes both the query and document sentence simultaneously to a Transformer network, producing an output value between 0 and 1, which indicates the relevance of the sentence pair. In reference to a study by Thakur *et al.*[28], it is highlighted that MiniLM demonstrates the best performance. Therefore, we evaluate the performance when using MiniLM as the Cross Encoder for re-ranking.

4.2   Main Results

The main retrieval results are illustrated in Table 1. It demonstrates that BM25 is an effective baseline for zero-shot IR compared with bi-encoders such as S-BERT and BioLinkBERT. BM25 ranking alone achieves reasonable performance, which can be further improved by re-ranking using a cross-encoder model. This two-stage ranking pipeline achieves the best MRR results on the NFCorpus dataset. However, re-ranking based on BM25 has limitations stemming from BM25's dependence on exact term matching, which can cause relevant documents to be excluded from consideration during later re-ranking stages.

A noteworthy scenario emerged where the precision of MedRetriever at the top 1000 exhibited favourable results among all the baseline retrievers. In contrast, the nDCG at the top 10 demonstrated comparatively suboptimal performance. This disparity between precision and nDCG metrics suggests that although the MedRetriever is capable of retrieving a fair proportion of relevant

Table 1. Performances of first-stage retrievers, BM25-based re-rankers and our proposed models. †The results were cited from [28]. *MedRetriever refers to our proposed method as a standalone approach, distinct from its fusion with BM25.

Method	NFCorpus				SCIFACT			
	MRR	P@10	nDCG@10	R@1k	MRR	P@10	nDCG@10	R@1k
First-stage Retrievers								
BioLinkBERT	0.329	0.132	0.173	0.532	0.519	0.076	0.550	0.979
S-BERT	0.501	0.218	0.300	0.574	0.570	0.082	0.596	0.959
DocT5Query†	–	–	0.328	–	–	–	0.675	–
DeepCT†	–	–	0.283	–	–	–	0.630	–
BM25	0.537	0.233	0.325	0.372	0.635	0.088	0.665	0.980
BM25-based Re-rankers								
Cross Encoder	**0.591**	0.244	0.350	0.250	0.662	0.091	0.688	0.908
S-BERT	0.430	0.170	0.232	0.229	0.539	0.081	0.568	0.864
Our Proposed Models								
MedRetriever *	0.499	0.222	0.298	**0.644**	0.540	0.083	0.581	**0.990**
MedFusionRank	0.552	**0.262**	**0.357**	**0.644**	**0.673**	**0.094**	**0.705**	**0.990**

documents overall, it struggles to rank the most relevant documents at the very top of the list. When we combine scores from two methods, MedRetriever and BM25, the results consistently outperformed nearly all of the baseline methods across all evaluation metrics.

4.3   Out-of-Vocabulary Strategy

Table 2. Performances of using different out-of-vocabulary strategies for MedFusionRank

Method	NFCorpus				SCIFACT			
	MRR	P@10	nDCG@10	R@1k	MRR	P@10	nDCG@10	R@1k
Prefix Approx	0.552	0.262	0.357	**0.644**	0.673	0.094	0.705	0.990
CharLSTM	**0.553**	**0.263**	**0.358**	0.643	**0.684**	0.094	**0.713**	0.990

To handle out-of-vocabulary (OOV) words, this work incorporates two strategies: Prefix Approximation and a Character-level Long Short-Term Memory network (CharLSTM). Prefix Approximation, originally proposed in [26], identifies the longest common prefix between an OOV word and in-vocabulary words, then averages all embeddings sharing that prefix to represent the OOV term. On the other hand, the CharLSTM learns sequential character-level features of

in-vocabulary words to construct a non-linear mapping from character sequences to medical embeddings. As depicted in Table 2, the CharLSTM achieves better overall performance compared to Prefix Approximation. This indicates that modelling the sequential patterns and characters of medical terminology plays a more vital role in estimating representations for OOV words in this domain.

4.4 Case Study

Table 3. Keywords in the retrieved document based on a single term as query

Query	Keywords in retrieved document				
zoloft	*depression*	*depressive*	*antidepressants*	*exercise*	*sertraline*
	aerobic	*therapy*	*anxiety*	*treatment*	*medication*
	therapeutic	*50*	*disorders*	*mental*	*older*
	effects	*67*	*rating*	*mdd*	*diagnostic*
myelopathy	*spinal*	*sclerotic*	*paraplegia*	*cobalamin*	*spine*
	vegetarian	*vegan*	*subacute*	*cervical*	*vitamin*
	degeneration	*hypertonia*	*diagnosed*	*reflexia*	*impairment*
	paresthesias	*rehabilitative*	*hypotrophy*	*neurogenic*	*diet*

To further evaluate the performance of our proposed model, we conducted a case study using short, single-term queries common in human searches. Statistical matching models like BM25 often struggle with these sparse queries, as the single terms may not exist in the corpus. As shown in Table 3, the sample query terms "*zoloft*" and "*myelopathy*" did not appear in any documents. However, our proposed model successfully retrieved relevant documents with medical concepts from the knowledge graph, ranking pertinent documents in the top 10 results for both queries.

In the first example, "*zoloft*" is an antidepressant medication. Therefore, "*depression*", "*depressive*", and "*anxiety*" are closely connected to "*zoloft*" since the medication aims to alleviate the symptoms associated with these conditions. In another example, "*myelopathy*" is a spinal cord pathology that can result from vitamin deficiency, spinal degeneration, or cord compression. The keywords "*spinal*", "*spine*", "*vitamin*" and "*degeneration*" from the retrieved document could be relevant to the query.

This case study highlights the potential of our proposed model to improve the search relevancy of short user queries. Our model effectively utilised associated medical concepts to match user information needs.

5 Conclusion and Future Work

In this paper, we have presented MedFusionRank, a novel zero-shot MIR approach that integrates the strengths of statistical methods and pre-trained lan-

guage models. Our key insight is to leverage a pre-trained BERT-style model to extract compact yet informative keywords. These keywords are then enriched with domain knowledge by linking them to conceptual entities within a medical knowledge graph.

Our experiments on two benchmark medical datasets demonstrate that Med-FusionRank achieves promising results, outperforming a range of existing models across various evaluation metrics. The case study also reveals MedFusionRank's ability to retrieve relevant documents even for short or single-term queries.

There are several exciting directions for future work. First, we plan to expand the coverage of our medical knowledge graph using more comprehensive knowledge resources. Second, we intend to explore more sophisticated entity-linking techniques beyond simple lexical matching. Third, to enable deployment on resource-constrained IoT devices, we will construct a vector database of the encoded document embeddings and load it directly onto the target hardware. This will circumvent the need for inference-time encoding and drastically reduce retrieval latency and memory overhead. Finally, we aim to implement an end-to-end prototype for real-time clinical decision support on medical IoT devices.

References

1. Alsentzer, E., et al.: Publicly available clinical Bert embeddings. arXiv preprint arXiv:1904.03323 (2019)
2. Boteva, V., Gholipour, D., Sokolov, A., Riezler, S.: A full-text learning to rank dataset for medical information retrieval. In: Ferro, N., et al. (eds.) ECIR 2016. LNCS, vol. 9626, pp. 716–722. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30671-1_58
3. Christopher, D., Raghavan, P., Schütze, H., et al.: Scoring term weighting and the vector space model. Introduction Inf. Retrieval **100**, 2–4 (2008)
4. Dai, Z., Callan, J.: Context-aware term weighting for first stage passage retrieval. In: Association for Computing Machinery, SIGIR 2020, pp. 1533–1536. New York, NY, USA (2020). https://doi.org/10.1145/3397271.3401204
5. Dai, Z., Xiong, C., Callan, J., Liu, Z.: Convolutional neural networks for soft-matching n-grams in ad-hoc search. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 126–134 (2018)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. 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), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423

7. Dimitrov, D.V.: Medical internet of things and big data in healthcare. Healthc. Inf. Res. **22**(3), 156–163 (2016)
8. Elhoseny, M., Ramírez-González, G., Abu-Elnasr, O.M., Shawkat, S.A., Arunkumar, N., Farouk, A.: Secure medical data transmission model for IoT-based healthcare systems. IEEE Access **6**, 20596–20608 (2018)
9. Grover, A., Leskovec, J.: node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855–864 (2016)
10. Huang, P.S., He, X., Gao, J., Deng, L., Acero, A., Heck, L.: Learning deep structured semantic models for web search using clickthrough data. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 2333–2338 (2013)
11. Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L.A., Mark, R.: Mimic-iv. PhysioNet (2020). https://physionet.org/content/mimiciv/1.0/. Accessed 23 Aug 2021
12. Johnson, A.E., et al.: Mimic-iii, a freely accessible critical care database. Sci. Data **3**(1), 1–9 (2016)
13. raj Kanakarajan, K., Kundumani, B., Abraham, A., Sankarasubbu, M.: BioSimCSE: biomedical sentence embeddings using contrastive learning. In: Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI), pp. 81–86 (2022)
14. Karpukhin, V., et al.: Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906 (2020)
15. Kim, S.W., Gil, J.M.: Research paper classification systems based on TF-IDF and LDA schemes. HCIS **9**, 1–21 (2019)
16. Li, Y., Wehbe, R.M., Ahmad, F.S., Wang, H., Luo, Y.: A comparative study of pretrained language models for long clinical text. J. Am. Med. Inform. Assoc. **30**(2), 340–347 (2023)
17. Lipscomb, C.E.: Medical subject headings (mesh). Bull. Med. Libr. Assoc. **88**(3), 265 (2000)
18. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
19. Lu, Z.X., et al.: Application of AI and IoT in clinical medicine: summary and challenges. Curr. Med. Sci. **41**, 1134–1150 (2021)
20. Nogueira, R., Yang, W., Lin, J., Cho, K.: Document expansion by query prediction. arXiv preprint arXiv:1904.08375 (2019)
21. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1), 5485–5551 (2020)
22. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using siamese BERT-networks. 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), pp. 3982–3992 (2019)
23. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M., et al.: Okapi at TREC-3. Nist Special Publication Sp **109**, 109 (1995)
24. Shen, Y., He, X., Gao, J., Deng, L., Mesnil, G.: Learning semantic representations using convolutional neural networks for web search. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 373–374 (2014)
25. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. J. Documentation **28**(1), 11–21 (1972)

26. Speer, R., Chin, J., Havasi, C.: ConceptNet 5.5: an open multilingual graph of general knowledge. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)
27. Taori, R., et al.: Alpaca: a strong, replicable instruction-following model. Stanford Center Res. Found. Models **3**(6), 7 (2023). https://crfm.stanford.edu/2023/03/13/alpaca.html
28. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: BEIR: a heterogeneous benchmark for zero-shot evaluation of information retrieval models. In: Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021). https://openreview.net/forum?id=wCu6T5xFjeJ
29. Touvron, H., et al.: Llama: open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
30. Wadden, D., et al.: Fact or fiction: verifying scientific claims. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7534–7550 (2020)
31. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: MiniLM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. Adv. Neural. Inf. Process. Syst. **33**, 5776–5788 (2020)
32. Zhang, Y., Chen, Q., Yang, Z., Lin, H., Lu, Z.: BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci. Data **6**(1), 52 (2019)