From Claim to Evidence: Verifying Chinese Health Claims with Medical Literature

Chaoyuan Zuo[1(✉)][0000−0001−5361−0193], Yishuang Liu[1][0009−0007−3768−2219],
Chenlu Wang[2][0009−0007−6870−5401], and Ritwik Banerjee[2][0000−0003−0336−0258]

[1] School of Journalism and Communication
Nankai University, China
{zuocy,yishuangliu}@nankai.edu.cn
[2] Department of Computer Science
Stony Brook University, USA
{chenlwang,rbanerjee}@cs.stonybrook.edu

Abstract. Ensuring the accuracy of health claims in media is vital for public well-being, and evidence-based claim verification is critical in achieving this goal. However, identifying relevant biomedical literature as evidence for health claims is particularly challenging, especially within cross-genre and cross-lingual contexts. We propose an *ad hoc* information retrieval (IR) task to identify support for Chinese health claims obtained from Chinese news sources. We demonstrate the feasibility of such a task by presenting experiments on a novel dataset of pairs of Chinese health claims and English biomedical literature. We describe a two-step methodology comprising (i) a selection of the most relevant candidates from 764K research papers, and (ii) a final re-ranking of this selection. Our comprehensive experimental research demonstrates that incorporating domain-specific information significantly enhances retrieval accuracy and claim verification efficacy. This strategy is a major step toward improving the credibility of public health information dissemination and reducing the prevalence of falsehoods in health journalism.

Keywords: Chinese health claim · Information retrieval · Dataset

1 Introduction

In the digital era, online health news articles serve as a vital resource for the general public to obtain information on health-related issues [11]. These news websites also function as key platforms for conveying scientific biomedical research to a broader audience [20,28]. The accuracy of these articles is critically important, as errors in the news-gathering process can lead to the spread of false information and misleading health advice [40]. Such misinformation may stem from various sources, including the misinterpretation of scientific findings due to a lack of specialized expertise, the oversimplification of complex medical concepts, exaggerated claims, or a focus on sensationalism to attract a larger audience at the expense of factual accuracy [2,4,14,23,29]. Therefore, it is crucial

Table 1: **Sample health claims (news headlines) from our data.** Health news often cites biomedical literature to lend credibility to the shared information: (1) a health claim not containing a link to biomedical literature; (2) a news headline without any specific check-worthy claim; (3, 4) a health claim worth checking vis-à-vis the linked biomedical literature.

	Health claim (news headline)	**Corresponding biomedical literature**
(1)	哮喘药物可唤醒“丢失”记忆 *Translation: Asthma drug can revive ‘lost’ memories* [news.sciencenet.cn/htmlnews/2023/1/492551.shtm]	*– no research paper cited –*
(2)	远程医疗的转折点 *Translation: A turning point for telemedicine* [www.thepaper.cn/newsDetail_forward_19291736]	**Headline:** Variation in Quality of Urgent Health Care Provided During Commercial Virtual Visits **DOI:** *10.1001/jamainternmed.2015.8248*
(3)	噩梦可能是帕金森病的早期预警 *Translation: Nightmares could be an early warning of Parkinson’s disease* [paper.sciencenet.cn/HTMLpaper/2022/6/202261013585490073449.shtm]	**Headline:** Distressing dreams and risk of Parkinson’s disease: A population-based cohort study **DOI:***10.1016/j.eclinm.2022.101474*
(4)	吃得越咸，死的越早 *Translation: The saltier you eat, the sooner you die* [www.cn-healthcare.com/articlewm/20230214/content-1510348.html]	**Headline:** Adding salt to foods and hazard of premature mortality **DOI:** *10.1093/eurheartj/ehac208*

to verify health claims in news articles using robust verification methods, such as cross-referencing original research studies, consulting experts, and employing automated fact-checking technologies.

Over the years, the verification of health news has evolved significantly, with recent advances introducing methods that leverage external evidence, such as Wikipedia and repositories of verified news [13,34]. An evidence-based approach is crucial for verifying health claims, as it involves cross-referencing these claims with original scientific research to ensure their accuracy [27,37].

However, identifying relevant evidence presents significant challenges. While many news articles support their claims by citing original research through embedded hyperlinks or explicit citations (see Table 1), such sources are not always provided. For instance, in Table 1-(1), a health claim is presented without any cited support from biomedical research. As these articles are disseminated across various platforms, tracing the original sources becomes increasingly difficult.

When scientific research transitions from academic journals to news articles intended for the general public, the information is conveyed in a markedly different language. While news articles aim to be more accessible and engaging for a broad audience, academic papers are often dense, technical, and complex. This linguistic and stylistic disparity complicates the task of aligning claims made in news articles with corresponding evidence in scientific literature. Additionally, cross-lingual challenges further exacerbate this process, as research studies and health news may be published in various languages, necessitating translation and interpretation to verify claims accurately.

We address these challenges by introducing a novel bilingual dataset specifically designed for the information retrieval (IR) task of verifying Chinese health claims. Our dataset comprises 8,647 health claims from Chinese news sources, each associated with biomedical literature documents sourced from international medical journals. The verification task is divided into two phases: (i) retrieving a candidate list of 100 abstracts from a large corpus of 763,956 research abstracts, and (ii) re-ranking these candidates using a transformer-based cross-encoder.

2 Dataset

To verify Chinese health claims, we developed a specialized IR dataset. It was constructed in two stages: first, we created a gold-standard dataset, which was then expanded into a silver-standard corpus.

2.1 Gold-Standard Dataset

We collected 13,748 news articles from the health sections of several Chinese general news platforms, as well as multiple medical news websites. We discarded headlines that were unduly brief or too lengthy. Next, we formatted the headlines using a set of rules, eliminating unnecessary metadata like the author names or the publication source. Furthermore, as a manual filtration step, we removed headlines such as Table 1-(2), which did not contain any information worth verifying. Following Zuo et al. [41], we regarded the remaining headlines as health claims to be verified, and checked each document for hyperlinks to domains listed by Wikipedia or Alexa as biomedical journals. We utilized the PubMed[3] API to retrieve specific information about the relevant publications. Due to the challenges of obtaining full-text articles, we extracted only the titles and abstracts. From this process, we have curated a novel dataset[4], structured in tuples of the form (c, e), where: c signifies the health claim derived from a news article, e represents the relevant publications. Our *gold-standard* dataset features 1,439 unique news headlines and 1,400 entries from the biomedical research literature,

[3] https://pubmed.ncbi.nlm.nih.gov/

[4] https://github.com/chzuo/nlpcc2024_chinese_claim_ir

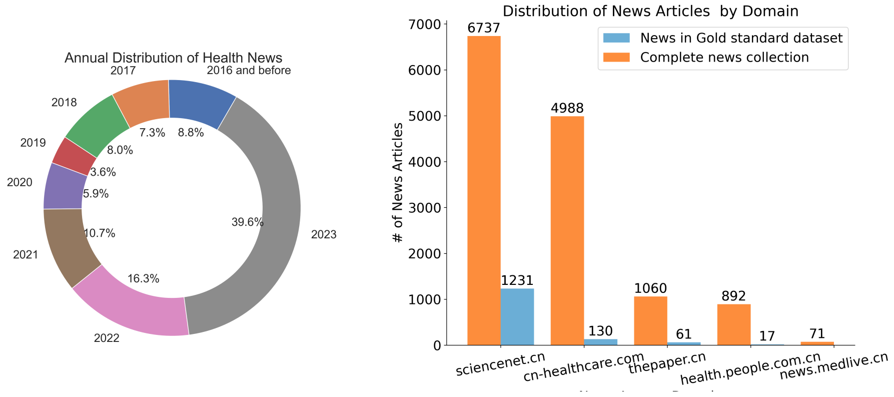

Fig. 1: **Dataset description:** The annual distribution of health news (left), showing the proportion of articles published from 2017 to 2023, with earlier articles combined into a single category; and the gold standard dataset's size against the complete corpus (right), shown across the top 5 news domains.

as some research publications are cited by multiple news articles. Fig. 1 first presents the temporal distribution of health news articles, and then compares this gold-standard corpus against the silver-standard expansion (which we describe next).

2.2 Silver-Standard Dataset

Two main issues with our dataset are (a) unequal distribution of news sources and (b) only one research citation per news article, even though several studies may be relevant. We developed a three-pronged strategy to tackle these concerns, resulting in the creation of our larger *silver-standard* dataset.

Similar news clustering. In our initial collection of 13,748 news articles, we noticed that some claims closely aligned with those in our *gold-standard* dataset but were not paired with biomedical literature. This may be due to the absence of such hyperlinks or the failure of our automatic extraction methods, possibly caused by the websites' anti-scraping mechanisms. This observation aligns with insights about news content redundancy [15], which indicates that multiple news articles often report on the same event. Therefore, we clustered similar news articles using the following steps:

1. After identifying the publication dates of the original articles, we conducted a search for additional health-related news stories published within a ten-day window surrounding these dates.
2. We then compared the headlines of these candidate articles with the original claims to evaluate their semantic similarity using Sentence-BERT [24].

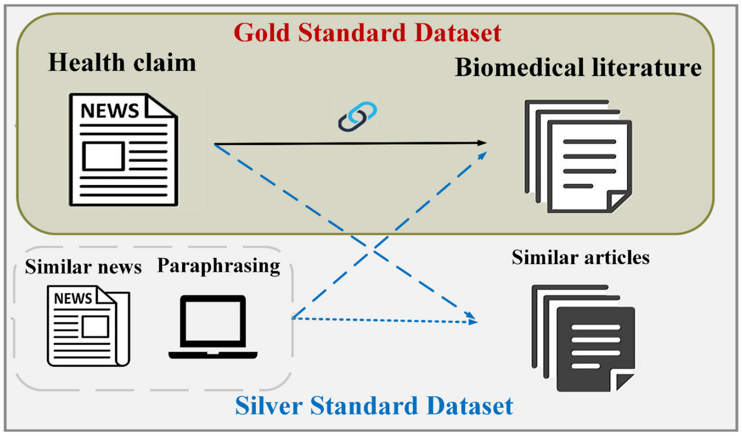

Fig. 2: Dataset construction schema. Direct hyperlinks from news articles to cited medical research literature are integral to the *gold-standard* corpus, which is expanded to the *silver-standard* collection based on inferred associations between claims and scholarly peer-reviewed work.

3. Subsequently, we utilized medical named entity recognition (NER) methods to extract medically relevant entities from these claims. By measuring the degree of overlap of these entities between the original and candidate claims, we assessed the relevance and accuracy of the matches.

Claim augmentation via paraphrasing. To improve the lexical and syntactic diversity of our dataset, we employed advanced paraphrasing techniques. After evaluating a range of tools [1,7], we selected the ChatGPT Paraphraser [36] model for its superior performance to maintain semantic integrity while providing high-quality paraphrasing. This tool allowed us to generate five distinct versions of each claim in our *gold-standard* dataset, thereby significantly enriching its linguistic diversity.

Expanding relevant literature. By using PubMed's "Similar Articles" option, we were able to add more relevant articles to our collection for each health claim c and the corresponding academic paper e. High relevance and thematic consistency are ensured by this feature, which finds related articles by comparing titles, abstracts, and MeSH terms using an advanced word-weighted algorithm[5]. In effect, we enhanced the coverage of our information retrieval approach by methodically expanding the set of pertinent publications for every health claim by utilizing PubMed's similarity data.

Our final *silver-standard* dataset comprises 8,647 health claims and references a total of 33,622 biomedical research articles.

[5] pubmed.ncbi.nlm.nih.gov/help/#similar-articles

3 Experiments

We present a bilingual information retrieval (IR) system to provide an effective and scalable solution to the problem of matching health claims to research articles. In this system, the titles and abstracts of biomedical research articles serve as the documents, while health claims function as the queries. The expected output of the system is a ranked list of research papers for each health claim, where the most relevant and supportive papers are positioned at the top. This approach addresses the critical need for accurate health claim verification by linking claims to credible scientific evidence. Drawing inspiration from established *ad hoc* IR techniques [6,19], our experimental setup consists of two stages to ensure both efficiency and precision. First, we retrieve a large candidate list of potential matches. Then, we refine this list by re-ranking the candidates using a transformer-based cross-encoder.

Two different IR datasets were constructed and utilized in this work: a *gold-standard* dataset D_{gold} with 1,439 queries and a *silver-standard* dataset D_{silver} with 8,647 queries. The original Chinese versions of the claims are retained, and translated to English via the Google Cloud Translation API[6] for future research. In addition, we include 730,334 biomedical research articles sourced from the non-commercial use open-access subset of PubMed Central[7] in order to replicate a realistic scenario where a human reader or fact-checker must find the correct publication from a large collection. The final datasets contain 763,956 biomedical research articles as documents, making them a rich resource for thorough testing and assessment of information retrieval tasks. Subsequently, for experiments, we divide our datasets as follows: 739 claims are assigned to training and 700 to testing for D_{gold}; likewise, the distribution for D_{silver} is 4,439 for training and 4,208 for testing.

For evaluation on the D_{gold} dataset, we measure Recall@k (R@k, $k = 1, 10$) to determine whether the hyperlinked research paper is included in the top-K selection of the IR system. Since each claim in this dataset cites only one biomedical literature, R@k effectively captures the system's ability to retrieve the correct document within the top results. Additionally, we use Mean Reciprocal Rank (MRR) to evaluate the ranking quality of the retrieved documents, offering insights into how prominently the correct document is ranked. For evaluation on the D_{silver} dataset, where the number of relevant papers for each claim has been expanded, we employ a more comprehensive set of metrics. These include Precision@k (P@k), R@k, MRR, and Mean Average Precision (MAP). Precision@k evaluates the proportion of relevant documents in the top-k results, while Recall@k measures the ability of the system to retrieve all relevant documents within the top-K. MAP offers a single-figure measure of quality across recall levels, summarizing both precision and recall aspects. These metrics together provide a comprehensive picture of the system's performance.

[6] https://cloud.google.com/translate/docs/reference/rest/

[7] www.ncbi.nlm.nih.gov/pmc

3.1 Candidate Retrieval

In the initial phase of our process, we narrow the search domain using several algorithms and models. We utilize the BM25 algorithm [25] and its variant, BM25+[17], which are well-established token-based retrieval models. We also implement transformer-based bi-encoders, namely DistilBERT [26] and MiniLM[38], which are pretrained via Sentence-BERT on the MS-MARCO dataset [21] to enable semantic search by vectorizing health claims and matching document embeddings based on cosine similarity. To address our specific needs in the biomedical domain, we employ PubMedBERT [12], a model pretrained on 14 million PubMed abstracts, and PubMedBert-MS-MARCO [8], which is pretrained on PubMed abstracts and fine-tuned on MS-MARCO. We employ the English translation of the queries for each of these algorithms and models. Lastly, we test with the Chinese version of the queries and integrate LaBSE [10], the state-of-the-art model for cross-lingual sentence retrieval tasks.

3.2 Re-Ranking

During the re-ranking stage, we utilize a cross-encoder technique to create training pairings by matching health claims with relevant documents. If the news article is pertinent to the document, these pairings are labeled as 1; otherwise, they are labeled as 0 if they are mismatched pairs from the BM25 candidate list [33,41]. Transformer models that have been pretrained can be refined with this configuration. This setup enables the fine-tuning of pretrained transformer models. We start the process with BM25+, producing a first draft list of 100 candidates for every claim and making sure that relevant literature is included. We use both MiniLM and PubMedBert-MS-MARCO models for training, with a special focus on a MiniLM variant pretrained as a cross-encoder on the MS-MARCO dataset.

3.3 Results

The results for the D_{gold} and D_{silver} datasets are shown in Table 2. We observe that PubMedBERT-MS-MARCO performs exceptionally well during the candidate selection phase. This model, pretrained on MS-MARCO, achieves better results than the original PubMedBERT, highlighting the effectiveness of pretraining on IR datasets. In contrast, models like MiniLM and DistilBERT, while performing moderately well, are outperformed by PubMedBERT-MS-MARCO.

On the other hand, LaBSE did not meet expectations. We conjecture that this is likely due to the cross-lingual nature of this task and its lack of specific training in biomedical literature, which adds complexity and affects its ability to effectively match health claims with relevant research papers.

Traditional models such as BM25 and BM25+ show reasonable performance. In particular, they excel on the D_{silver} dataset. However, since these are token-based methods, there are challenges in the presence of a terminological mismatch

Table 2: **Information retrieval results.** Models with † are pretrained on the MS-MARCO dataset. Models with * are pretrained as cross-encoders. All models are trained for 2 and 3 epochs, with maximum sequence length of 256 and 512 tokens. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

	D_{gold}			D_{silver}					
Model	**R@1**	**R@10**	**MRR**	**P@1**	**P@10**	**R@10**	**R@50**	**MRR**	**MAP**
Candidate selection									
BM25	12.9	27.7	18.4	21.9	12.8	4.2	9.8	29.9	15.6
BM25+	13.0	27.6	18.6	21.8	12.9	4.2	10.0	29.9	15.3
DistilBERT†	14.0	31.3	19.9	15.1	6.5	2.3	4.9	20.6	10.9
MiniLM†	11.3	30.4	17.9	19.5	10.2	3.4	7.2	27.0	14.1
PubMedBERT	9.1	24.7	14.4	14.5	8.5	2.8	7.0	22.0	10.9
PubMedBERT†	**17.6**	**39.1**	**25.2**	**27.1**	**13.7**	**4.6**	**10.4**	**34.9**	**17.9**
LaBSE	3.0	11.6	6.0	4.4	2.3	0.8	2.1	8.3	4.6
Re-ranking									
MiniLM*	32.7	71.9	46.2	95.0	88.9	30.2	66.1	97.0	**87.6**
MiniLM†	34.7	90.0	52.6	91.0	87.7	29.9	**66.2**	94.6	86.6
PubMedBERT†	**58.3**	**98.0**	**72.2**	**95.5**	**90.2**	**31.0**	64.0	**97.3**	**87.6**

between the claim and the research document. In such instances, transformer-based cross-encoders demonstrate their advantages. Notably, in the re-ranking phase, PubMedBERT-MS-MARCO achieves the highest metrics, with an R@1 of 58.3% and an R@10 of 98.0% on D_{gold}, clearly demonstrating its superior ability to bridge the gap between health claims and biomedical research abstracts. These findings underscore the efficacy of advanced, domain-specific knowledge in information retrieval tasks, particularly when that task concerns the specialized language of biomedical research literature.

4 Related Work

4.1 Health Claim Verification

Within the journalistic industry, fact-checking and disinformation detection are well-established procedures. Over the years, several researches have focused on health claim verification. The method was initially based on evaluating inherent features of news items, including language patterns and the reliability of the source [5,39]. However, finding a reliable source of relevant information is necessary for readers to validate a health claim. Recent developments have brought forth methods based on outside data.

By selecting study abstracts that either confirm or contradict certain scientific claims, Wadden et al. [37] achieved important progress in scientific claim verification. Their study includes a dataset particularly created for scientific claim verification utilizing evidence-rich abstracts, and it also explains the reasoning behind each pick. Since 2020, the emergence of datasets containing biomedical and health-related claims has been significantly driven by the proliferation of online content related to the COVID-19 pandemic. The rising interest in textual claim fact-checking was brought to light by Sarrouti et al. [27], who introduced a dataset based on actual health claims with an emphasis on COVID-19. Carefully chosen scientific literature is cross-referenced to support these statements. In order to investigate medical misinformation, Srba et al. [31] released a dataset that included around 317k medical publications and 3.5k statements that had been fact-checked. This dataset includes over 51k connections that have been algorithmically tagged and 573 human-labeled links that represent the article and position of a claim. Recently, a novel dataset called HealthFC was introduced[35]. HealthFC includes 750 health-related claims in German and English, labeled for veracity by medical experts and backed by evidence from systematic reviews and clinical trials. However, there are very few datasets that focus on verifying Chinese health news claims.

4.2 Information Retrieval Methods and Datasets.

The majority of term-weighting techniques used in modern *ad hoc* information retrieval (IR) systems come from bag-of-words representations and the BM25 algorithm and its variations [18,25]. A popular ranking algorithm for text document retrieval tasks, BM25 evaluates a document's relevance to a query by looking at a number of different criteria, including how frequently the query phrases appear in the document. Historically, these techniques have been used for query-based news retrievals that target certain sections of an article [3]. But with the introduction and success of BERT [9] and its offspring [16,26] in NLP, these sophisticated language models have been increasingly used in recent IR research on evidence retrieval [30]. The use of transformer-based encoders, which show the importance of contextual information in re-ranking candidate papers that were first obtained by BM25-based models [22,42], is particularly pertinent to this work.

Large-scale IR datasets are useful benchmarks for assessing and contrasting different IR models. Among the notable instances are the TREC collections[8] and MS-MARCO [21]. These databases are mostly limited to specific document types and genres. The availability of cross-genre IR datasets is restricted, especially for bilingual retrieval tasks, despite the existence of several cross-lingual IR (CLIR) and a small number of cross-genre IR datasets [32,41]. In order to close this gap, we offer a dataset that consists of two different linguistic genres and languages. This dataset will be a useful tool in reducing the negative consequences of Chinese health misinformation.

[8] trec.nist.gov/data.html

5 Conclusion

In this work, we present a novel dataset for verifying Chinese health claims by linking health news articles to biomedical research literature. We utilize this dataset in our comprehensive pipeline, demonstrating that cross-genre and cross-lingual information retrieval is feasible for validating health-related information. Our findings emphasize the critical role of incorporating domain-specific knowledge in such an information retrieval process. Furthermore, our study underscores the need for responsible media consumption and careful fact-checking to not only identify misinformation in health-related claims, but also to promote the spread of accurate health information.

Acknowledgment. This work was supported by the Fundamental Research Funds for the Central Universities (No. 63232115).

References

1. Black, S., Gao, L., Wang, P., Leahy, C., Biderman, S.: GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow (2021), https://doi.org/10.5281/zenodo.5297715, version 1.0
2. Brown, J., Chapman, S., Lupton, D.: Infinitesimal risk as public health crisis: News media coverage of a doctor-patient HIV contact tracing investigation. Social Science & Medicine **43**(12), 1685–1695 (1996)
3. Catena, M., Frieder, O., Muntean, C.I., Nardini, F.M., Perego, R., Tonellotto, N.: Enhanced news retrieval: Passages lead the way! In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019. pp. 1269–1272. ACM (2019). https://doi.org/10.1145/3331184.3331373, https://doi.org/10.1145/3331184.3331373
4. Caulfield, T.: The Commercialisation of Medical and Scientific Reporting. PLoS Medicine **1**(3), e38 (2004)
5. Clarke, C.L.A., Rizvi, S., Smucker, M.D., Maistro, M., Zuccon, G.: Overview of the TREC 2020 health misinformation track. In: Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC. NIST Special Publication, vol. 1266 (2020), https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.HM.pdf
6. Dai, Z., Callan, J.: Deeper Text Understanding for IR with Contextual Neural Language Modeling. In: SIGIR 2019. pp. 985–988. ACM (2019), https://doi.org/10.1145/3331184.3331303
7. Damodaran, P.: Parrot: Paraphrase generation for NLU (2021), https://github.com/PrithivirajDamodaran/Parrot_Paraphraser, version 1.0
8. Deka, P., Jurek-Loughrey, A., Deepak, P.: Improved Methods To Aid Unsupervised Evidence-Based Fact Checking For Online Health News. Journal of Data Intelligence **3**(4), 474–504 (2022), https://doi.org/10.26421/JDI3.4-5
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: 2019 NAACL). pp. 4171–4186. ACL (2019). https://doi.org/10.18653/v1/N19-1423

10. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic BERT sentence embedding. In: Proceedings of the 60th Annual Meeting of the ACM (Volume 1: Long Papers). pp. 878–891. Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.62, https://aclanthology.org/2022.acl-long.62
11. Fox, S., Duggan, M.: Health Online 2013. Internet & Technology, Pew Research Center (January 2013), https://www.pewresearch.org/internet/2013/01/15/health-online-2013/, last accessed: May 31, 2020
12. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Trans. Comput. Heal. **3**(1), 2:1–2:23 (2022). https://doi.org/10.1145/3458754, https://doi.org/10.1145/3458754
13. Kotonya, N., Toni, F.: Explainable automated fact-checking: A survey. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 5430–5443. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.474
14. Lebow, M.A.: The pill and the press: Reporting risk. Obstetrics & Gynecology **93**(3), 453–456 (1999)
15. Liu, J., Liu, T., Yu, C.: NewsEmbed: Modeling News through Pre-trained Document Representations. In: KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1076–1086. ACM (2021). https://doi.org/10.1145/3447548.3467392
16. 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. arXiv preprint arXiv:1907.11692 (2019)
17. Lv, Y., Zhai, C.: Adaptive term frequency normalization for BM25. In: Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011. pp. 1985–1988 (2011). https://doi.org/10.1145/2063576.2063871
18. Lv, Y., Zhai, C.: When Documents Are Very Long, BM25 Fails! In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1103–1104. SIGIR '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/2009916.2010070, https://doi.org/10.1145/2009916.2010070
19. MacAvaney, S., Yates, A., Cohan, A., Goharian, N.: CEDR: Contextualized Embeddings for Document Ranking. In: Proceedings of the 42nd International ACM SIGIR Conference Research and Development in Information Retrieval. pp. 1101—-1104 (2019), https://doi.org/10.1145/3331184.3331317
20. Medlock, S., Eslami, S., Askari, M., Arts, D.L., Sent, D., de Rooij, S.E., Abu-Hanna, A.: Health Information–Seeking Behavior of Seniors Who Use the Internet: A Survey. Journal of Medical Internet Research **17**, e10 (2015). https://doi.org/10.2196/jmir.3749
21. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A human generated machine reading comprehension dataset. In: Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches. CEUR Workshop Proceedings, vol. 1773 (2016), https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf
22. Nogueira, R., Cho, K.: Passage Re-ranking with BERT (2020), https://doi.org/10.48550/arXiv.1901.04085
23. Ransohoff, D.F., Ransohoff, R.M.: Sensationalism in the Media: When Scientists and Journalists May Be Complicit Collaborators. Effective Clinical Practice **4**(4), 185 (2001)

24. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: EMNLP-IJCNLP 2019. pp. 3980–3990 (2019). https://doi.org/10.18653/v1/D19-1410
25. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval pp. 333–389 (2009). https://doi.org/10.1561/1500000019
26. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter (2020), https://doi.org/10.48550/arXiv.1910.01108
27. Sarrouti, M., Ben Abacha, A., Mrabet, Y., Demner-Fushman, D.: Evidence-based fact-checking of health-related claims. In: Findings of the ACM: EMNLP 2021. pp. 3499–3512. ACM, Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.findings-emnlp.297
28. Sbaffi, L., Rowley, J.: Trust and Credibility in Web-Based Health Information: A Review and Agenda for Future Research. Journal of Medical Internet Research **19**(6), e218 (2017). https://doi.org/10.2196/jmir.7579
29. Shuchman, M., Wilkes, M.S.: Medical scientists and health news reporting: A case of miscommunication. Annals of Internal Medicine **126**(12), 976–982 (1997)
30. Soleimani, A., Monz, C., Worring, M.: BERT for Evidence Retrieval and Claim Verification. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in Information Retrieval. pp. 359–366. Springer International Publishing, Cham (2020), https://doi.org/10.1007/978-3-030-45442-5_45
31. Srba, I., Pecher, B., Tomlein, M., Moro, R., Stefancova, E., Simko, J., Bielikova, M.: Monant Medical Misinformation Dataset: Mapping Articles to Fact-Checked Claims. In: Proceedings of the 45th International ACM SIGIR. pp. 2949–2959. SIGIR '22, ACM (2022). https://doi.org/10.1145/3477495.3531726
32. Sun, S., Duh, K.: CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4160–4170. ACM, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.340, https://aclanthology.org/2020.emnlp-main.340
33. Thakur, N., Reimers, N., Daxenberger, J., Gurevych, I.: Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In: Proceedings of the 2021 Conference of the North American Chapter of the ACM: Human Language Technologies. pp. 296–310. ACM, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.28
34. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for fact extraction and VERification. In: Walker, M., Ji, H., Stent, A. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 809–819. Association for Computational Linguistics, New Orleans, Louisiana (2018). https://doi.org/10.18653/v1/N18-1074
35. Vladika, J., Schneider, P., Matthes, F.: Healthfc: Verifying health claims with evidence-based medical fact-checking. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy. pp. 8095–8107. ELRA and ICCL (2024), https://aclanthology.org/2024.lrec-main.709
36. Vorobev, V., Kuznetsov, M.: A paraphrasing model based on ChatGPT paraphrases (2023), https://huggingface.co/humarin/chatgpt_paraphraser_on_T5_base

37. Wadden, D., Lin, S., Lo, K., Wang, L.L., van Zuylen, M., Cohan, A., Hajishirzi, H.: Fact or fiction: Verifying scientific claims. In: EMNLP 2020). pp. 7534–7550. ACM, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.609
38. 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. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 5776–5788. Curran Associates, Inc. (2020)
39. Wang, W.Y.: "liar, liar pants on fire": A new benchmark dataset for fake news detection. In: Proceedings of the 55th Annual Meeting of the ACM (Volume 2: Short Papers). pp. 422–426. ACM, Vancouver, Canada (Jul 2017). https://doi.org/10.18653/v1/P17-2067, https://aclanthology.org/P17-2067
40. Yavchitz, A., Boutron, I., Bafeta, A., Marroun, I., Charles, P., Mantz, J., Ravaud, P.: Misrepresentation of Randomized Controlled Trials in Press Releases and News Coverage: A Cohort Study. PLOS Medicine **9**(9), 1–11 (2012). https://doi.org/10.1371/journal.pmed.1001308
41. Zuo, C., Acharya, N., Banerjee, R.: Querying across genres for medical claims in news. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1783–1789. ACM, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.139
42. Zuo, C., Wang, C., Banerjee, R.: Cross-genre retrieval for information integrity: A COVID-19 case study. In: Advanced Data Mining and Applications - 19th International Conference, ADMA 2023, Shenyang, China, August 21-23, 2023, Proceedings, Part V. Lecture Notes in Computer Science, vol. 14180, pp. 495–509. Springer (2023). https://doi.org/10.1007/978-3-031-46677-9_34, https://doi.org/10.1007/978-3-031-46677-9_34