MARTIN-LUTHER-UNIVERSITÄT
HALLE-WITTENBERG

Martin-Luther-Universität Halle-Wittenberg
Institute of Computer Science
Degree Programme Informatik M.Sc.

Identifying and Answering Health-Related Questions

Master's Thesis

Jan Heinrich Reimer

1st Referee: Prof. Dr. Matthias Hagen
2nd Referee: Alexander Bondarenko, M.Sc.

Submission date: April 11, 2023

Acknowledgments

I'd like to thank the students who helped me in annotating questions: Marika Bauer, Sven Fibelkorn, Max Henze, Christine Kahle, Sophie Krull, Ferdinand Schlatt, and Andreas Zimmermann. A big thanks to Alexander Bondarenko and Maik Fröbe for not only investing a lot of time in annotating questions and judging abstracts, but also for their valuable feedback on this thesis. Alexander Bondarenko also provided the conversational model outputs in Section 5.5.4. Thanks to Max Henze and Johannes Reimer for taking the time to proofread this thesis. Last but not least, I want to thank Lena Merker for annotating, judging, proofreading, and for supporting me throughout the whole process.

Abstract

People frequently use web search engines to find answers to health-related questions, but often receive incorrect or biased answers. Because users trust the top-ranked search results, such misinformation can cause severe harm, especially if medical questions are answered incorrectly despite lacking evidence. General search engines lack effective measures to prevent misinformation and rarely display warnings for health-related questions. A first step towards preventing misinformation is therefore to identify health-related or medical questions. Countermeasures against misinformation can then be applied to answer these health-related questions correctly. Both identifying and answering health-related questions are open problems due to limited datasets, ineffective question classifiers or a lack of misinformation prevention in retrieval systems.

In this thesis, we explore three directions to advance the identification and answering of health-related questions: First, we collect a large, realistic dataset of medical, health-related, and non-health-related questions from various sources. We automatically label health-related and medical questions based on their source and vocabulary, by applying labeling heuristics and weak supervision. Second, we optimize feature-based and transformer-based classifiers to identify health-related and medical questions with high effectiveness. Third, we investigate a modular, evidence-based question answering and retrieval architecture for health-related yes-no questions. Our approach first infers the correct answer to the question based on trusted evidence from relevant biomedical literature, then retrieves topically relevant web documents, and finally re-ranks the documents to favor documents that support the correct answer.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Web search engines are nowadays commonly used as universal question answering systems to find answers to questions from a wide range of topics [Bro02; CTS+21; KPR+19]. Many people search for health-related information (e.g., nutrition, symptoms, or medication) online [BWS+03; CMW14], and consequentially, a large part (5–24 %) of the queries in web search are health-related [EK03; JS06; SYJ+04]. Even though health-related questions can be relatively safe and non-medical questions (e.g., "What is the best diet for weight loss?"), people also search for more severe, medical topics (e.g., "Does garlic help with thrush?") that require professional advice. The COVID-19 pandemic has highlighted the risks of spreading incorrect or biased information in search results and summarized answers, especially in health-related and medical questions [BSD+21; CRS+20]. Generic search engines are often unable to prevent such misinformation and rarely display warnings or disclaimers for health-related questions or queries. An example is shown in Figure 1.1 where the question "Does garlic help with thrush?" is incorrectly answered by Bing,[1] but no warning or suggestion to seek professional medical advice is displayed. Because people often trust major search engines, this misinformation can severely bias the searcher's choice of treatment [ND22; WH15]. Following the harmful advice of such direct answers can also be dangerous, e.g., the use of garlic to treat thrush can lead to face burns [BCI07].

Health search engines often apply countermeasures against misinformation, such as retrieving from trusted sources (e.g., biomedical literature) or verifying claims [PMN+21; ZTA+22], to answer health-related questions correctly. In a general search engine, however, most questions and queries are not health-related [EK03; Eys04; JS06]. Because misinformation is most harmful in health-related and medical questions (e.g., alternative treatments can be toxic or even hinder conventional treatment [JPG+18; PMM+17]), we want to focus countermeasures primarily to the health-related questions. It is therefore necessary to first identify which questions are health-related or medical. Training feature-based machine learning models or fine-tuning language models is a common approach for text classification that requires labeled training data [MKC+22]. But existing

[1] `https://bing.com`

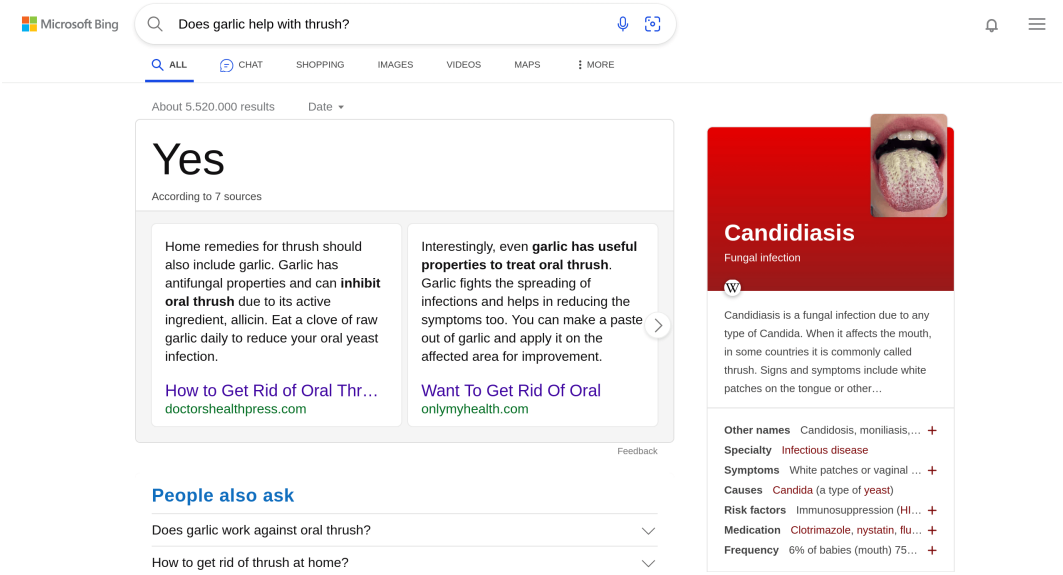


Figure 1.1: Screenshot of Bing showing an incorrect featured snippet for the query `Does garlic help with thrush?` The snippet claims that garlic is a helpful treatment for thrush, but according to Bondarenko et al. [BSD+21], there is not enough evidence for that claim. Garlic might even cause allergy, bleeding, or burns [BCI07].

health-related question answering datasets are either unlabeled, relatively small, or do not cover the wide range of health-related questions that people ask in web search (see Table 2.1). Consequently, there is a need for a large, labeled dataset of medical, health-related and non-health-related questions.

The example in Figure 1.1 also demonstrates that yes-no questions about the effectiveness of treatments or about diagnoses are particularly prone to misinformation and biases [BSD+21; WA14]. Such yes-no questions have therefore been the focus of recent research on misinformation prevention in search, e.g., at the TREC Health Misinformation tracks [ASL+19; CMS+22; CMS21; CRS+20]. After identifying a health-related yes-no question, first the correct answer to the question needs to be inferred from trusted evidence. This predicted "true" answer should then be supported by web search results that agree with the evidence. In previous TREC Health Misinformation tracks, often large language models were applied to answer health-related questions, but such models are not explainable [PMN+21; WMR+21]. It is therefore questionable whether they should be used in real search engines to answer medical questions that can cause serious harm if answered incorrectly.

Our contribution in this thesis is threefold: (1) We collect questions from various sources to build a new, automatically labeled dataset of medical, health-related, and non-health-related questions, (2) we propose new methods to identify health-related and medical questions, and (3) we build a modular search and question answering

system using biomedical literature to infer the correct answer to a health-related yes-no question and to find web documents that support this answer. We also highlight a number of open research questions that remain to be addressed in the future. In the remainder of this chapter, we first discuss the challenges that motivate our work in more detail. We then formulate research questions and finally describe the structure and contributions of this thesis.

1.1 Health-Related and Medical Queries and Questions

People use web search engines every day to find a diet for their fitness program, to search for symptoms of a disease, or even to ask for dosage of medication, that is, to seek health-related information [BMS+19; PHM+16]. In a survey, Baker et al. [BWS+03] showed that 40 % of 4,764 U.S. Internet users aged 21 years or older have looked for health-related information online, and up to 32 % reported that health-related information on the Web influenced their decisions about health or health care. A more recent study finds that more than 50 % of U.S. adults used web resources to find health-related information, most often to find medical treatments or help diagnose a health condition [CMW14]. Users also often ask for recommendations on diets or recovery [Zha10]. The Web is increasingly used as the first source of information when searching for health-related information, but results are often not satisfactory [FBG+19]. This development is concerning because the Web contains increasingly large amounts of misinformation (e.g., 55 % of tweets about cures of cancer were identified as misinformation by Bal et al. [BSD+20]).

Nonetheless, people rely on search engines to find information about health-related topics [JYX+23]. Questions such as "Can smoking prevent COVID-19?", or "What are the side effects of methadone?" are common in web search. Eysenbach and Köhler [EK03] and Eysenbach [Eys04] looked at the prevalence of such health-related searches and report that 4.5 % of the queries issued to web search engines are health-related. Similar amounts were reported by Jansen and Spink [JS06] who found that 7.5–9.5 % of search queries are about health or science. For question answering systems, a study by Spink et al. [SYJ+04] reports that as much as 24 % of the questions can be health-related, substantially more than in web search. Together, these studies show that a substantial amount of information needs is health-related and could thus benefit from improving health-related information retrieval.

White, Richardson, and Yih [WRY15] found that in real-world search engines people often formulate information needs in the form of natural language questions. 3.2 % of the English queries in a query log from 2010 to 2011 were formulated as questions. Recently, search engines like Bing and You.com have integrated conversational language models that do not only process questions in natural

language but also answer in fluent sentences.[2] This development suggests that in the future, increasingly more information needs will likely be expressed in the form of natural questions because asking questions is a more intuitive way for the public and medical professionals to access health-related information [JYX+23]. Cambazoglu et al. [CTS+21] found that a large proportion of questions in web search are yes-no questions, i.e., questions that can be answered with either "yes" or "no" like shown in our example in Figure 1.1. Yes-no questions that are asked in web search engines to support hypothesis-directed inference [CWH11] such as verifying the effectiveness of a treatment or a (self-)diagnosis are particularly prone to misinformation and biases [BSD+21; WA14]. Hence, they are of particular interest for this thesis. To automatically identify health-related questions is therefore the first step towards improving health-related information retrieval. Previous health-related question answering datasets usually do not contain non-health-related questions to contrast health questions with (see Table 2.1), which makes it difficult to use them for training classifiers to identify health-related questions.

A similar way to ask health-related questions online are community question answering platforms like Stack Exchange[3] or Yahoo! Answers,[4] where users can ask or answer questions in various categories, including medicine and other health-related topics such as fitness and wellness. Zhang [Zha10] found that people often turn to community question answering platforms because they either could not see a doctor yet or, more concerning, because they felt that their questions were not addressed completely by professionals they had previously consulted. Most questions also express negative emotions such as worry and anxiety [Zha10]. Their findings emphasize the need to actively encourage users to seek professional advice on medical questions besides providing correct and explainable answers on all health-related questions. However, most prior studies do not distinguish between medical or non-medical questions [EK03]. Others distinguish questions by the types of users, e.g., clinical questions from professionals and consumer health questions from the public [JYX+23]. But for search engines it is more important to consider the potential harm caused for questions that are either medical or not. Clearly, also consumer health questions like "Does garlic help with thrush?" (see Figure 1.1) can be medical and hence require a high level of verification. In this thesis, we therefore propose definitions of health-related and medical questions, create a dataset of questions labeled according to these definitions, and build models to identify health-related and medical questions. We experiment with various feature-based machine learning models and language model fine-tuning to approach this text classification task [Bre01; CG16; CL11; DCL+19; PNL21].

[2] Bing Chat: `https://bing.com/chat`, You Chat: `https://you.com/chat`

[3] `https://stackexchange.com/`

[4] `https://answers.yahoo.com/` (discontinued in 2021)

1.2 Health Misinformation and Biases

In a study on frequent questions about alternative remedies for common diseases, Bondarenko et al. [BSD+21] found that despite many people using search engines for personal health issues, 32 % of Google's top-10 snippets on popular medical questions promote misbeliefs. That is, Google's snippets contradict the scientific consensus as annotated by a medical expert. Figure 1.1 shows an even more concerning example where Bing displays a direct answer claiming that garlic helps with thrush. But in medical literature there is not enough evidence to support that claim. Worse yet, garlic can cause allergies, bleeding, or burns [BCI07; BSD+21]. As claimed by Potthast, Hagen, and Stein [PHS20], such direct answers pose a dilemma to search engines (and their users): Either show users direct answers with the risk of wrong answers or show users a list of documents with the risk of users not finding the right answer. A particularly concerning effect of misinformation is that people often change their beliefs based on such featured snippets and over-estimate the snippets' credibility at the same time [BZE22]. Effectively, featured snippets and other highly ranked search results are treated as if they were testimony, on which users then base their decisions [AS19; ND22]. The same behavior can be observed on question answering platforms where people over-estimate the quality of health-related answers even compared to answers from domain experts [OYW12]. Great care should therefore be taken when presenting direct answers to health-related questions. Because many people (42 %) have difficulties to assess the reliability of health-related information [DOM+21], such direct answers should therefore not only be as correct as possible but also be explainable, e.g., by citing trusted evidence. We use the PubMed corpus of biomedical literature[5] as the basis for our evidence-based answer prediction approach.

Furthermore, White and Horvitz [WH15; Whi14] also found a confirmation bias in search results for medical questions: High ranked documents often report treatments as helpful. Bondarenko et al. [BSD+21] found the same confirmation bias in featured snippets. And because biased search results significantly influence our beliefs and decisions about medical treatments [PGS+17], users tend to believe that the treatment is helpful. White and Horvitz [WH15; Whi14] also found that search results which claim that a treatment helps have higher dwell times, that is, search engine users take more time to read affirming documents. A possible explanation can be that queries are often positively framed [Azz21], e.g., "Does garlic help with thrush?" instead of "Can treating thrush with garlic be harmful?" Abualsaud and Smucker [AS19] and Azzopardi [Azz21] also found that people are more likely to infer a correct answer if more correct search results are returned. Conflicting search results, on the other hand, can lead to confusion and uncertainty [BFC22]. Besides answering the question correctly, search engines should therefore also

[5] `https://pubmed.gov`

care about the ranking position of the individual documents to support the correct answer [WH15; Whi14] and penalize incorrect beliefs, not only by treating them as irrelevant but by assigning incorrect answers a negative gain [CRS+20; PGS+17]. Our re-ranking approaches consider the predicted correctness of web documents to actively reduce the retrieval score of documents that are predicted to be incorrect.

1.3 Research Questions and Outline

Based on the problems described above and the related work, we formulate the following seven research questions that should be answered in this thesis. Finally, we outline the structure of the thesis.

RQ1 *Can we build a dataset of millions of medical, health-related, and non-health-related questions that is representative for real-world question answering and search systems?* A realistic dataset is necessary to study all health-related information needs, not only questions regarding specific topics. Such a dataset should also contain questions that are not health-related to pose as negative examples to compare against health-related questions. Research Question 1 is addressed in Chapter 3.

RQ2 *Can we automatically label medical and health-related questions with close to human-level accuracy?* Manual annotations are expensive or even infeasible for large datasets. Therefore, we need to investigate whether automatic labeling approaches can be used to label millions of medical and health-related questions. Research Question 2 is also addressed in Chapter 3.

RQ3 *How are health-related or medical questions different from other questions?* Finding differences in health-related or medical questions aims at a better understanding of how health-related questions can be identified but also can provide insights into the information needs typically associated with health-related questions. We address Research Question 3 in Chapter 3.

RQ4 *Can we build a classifier that can effectively and efficiently identify health-related and medical questions?* In a search engine, health-related questions should be identified as early as possible to show warnings (at least for medical questions) and to apply countermeasures against misinformation commonly found on the Web. Research Question 4 is addressed in Chapter 4.

RQ5 *Does increasing the training dataset size with automatic labeling benefit the classification of health-related and medical questions?* Automatic labeling allows us to substantially increase the size of the training dataset at the cost of a lower annotation quality. Research Question 5 probes the benefit of

automatic labeling approaches to build more effective classifiers by enlarging the training dataset, which we address in Chapter 4.

RQ6 *Can we answer health-related yes-no questions correctly, i.e., without spreading misinformation?* Yes-no questions about the effectiveness of treatments are common in health-related search queries. Answering such questions correctly is important to avoid amplifying cognitive biases commonly associated with health-related questions. We address Research Question 6 in Chapter 5.

RQ7 *Can we effectively retrieve trusted biomedical abstracts as evidence for answering health-related questions?* Biomedical literature is a form of evidence that is trusted by many users because scientific articles are often peer-reviewed and therefore considered to be more reliable than other sources. Research Question 7 aims at finding the best way to retrieve biomedical literature to find suitable evidence for correctly answering health-related questions, and is addressed in Chapter 5.

RQ8 *Can we effectively retrieve web documents that support the correct answer to a health-related question?* In a web search engine, users expect to find web documents that are topically relevant to their query. For health-related questions, we also have to ensure that the retrieved documents do not contradict the correct answer to the question as this might induce biases. Research Question 8 is addressed in Chapter 5.

To this extent, we first present a new dataset of 8.5 million automatically labeled questions that contains 2.0 million health-related and 1.3 million medical questions. We then propose new methods to identify health-related questions with an F_1-score [Rij79, p. 134] of up to 0.80 and medical questions with an F_1-score of up to 0.69. Finally, we build a modular search and question answering system that can be used to answer health-related questions based on evidence. Our system is able to answer health-related questions with an AUC score of 0.83 and finds web documents supporting the predicted answer with a compatibility difference of 0.19 between helpful and harmful results [CSV20; CVS20].

The structure of this thesis roughly follows the steps to extend a general purpose web search engine to identify and answer health-related natural language questions. After we have introduced the requirements and potential biases of a health-related search engine in Chapter 1, we review prior work on the labeling and classification of health-related questions, existing components of health search engines as well as misinformation detection in Chapter 2. In Chapter 3 we describe how we build a large-scale, diverse dataset of health-related and non-health-related questions and apply weakly supervised automatic labeling to tag each question as medical, health-related, or non-health. Chapter 4 focuses on the design and training of feature-based and neural classifiers for the task of identifying health-related and

medical questions. Then, in Chapter 5, we build a modular search pipeline for answering health-related yes-no questions and evaluate the effectiveness of our retrieval system based on the TREC 2022 Health Misinformation track. Finally, we conclude our findings with respect to the formulated research questions and discuss future work in Chapter 6.

Chapter 2

Related Work

The advances discussed in the subsequent chapters of this thesis build upon previous work in the fields of health-related information retrieval, question answering, claim verification, and misinformation detection. In this chapter, we introduce the background and related work in these fields. Section 2.1 continues the discussion of findings regarding health-related questions that were introduced in Chapter 1 and summarizes approaches for identifying health-related or medical questions. We then turn to recent work in health-related information and question answering in Section 2.2. After introducing the early envisioned systems and tools to build such systems, we compare current systems based on two benchmark collections. Section 2.3 discusses the state of the art in claim verification and misinformation detection. Finally, in Section 2.4, we discuss previous work that tackles misinformation when retrieving health-related web documents in the context of the TREC Health Misinformation tracks.

2.1 Health-Related Questions and Classification

In Chapter 1, we have discussed that health-related questions require task-specific processing such as misinformation detection in search engines. But only 5–24 % of all questions online are health-related [Eys04; SYJ+04], so a majority does not require misinformation detection or specialized corpora. In the ad hoc retrieval setting this means we first need to identify whether a query is a health-related question. The second important distinction to be made is between non-medical questions (e.g., questions centered around nutrition or fitness) on the one hand and medical questions (i.e., where professional expertise is needed) on the other hand. Medical questions often require answers from experts and can cause harm if answered incorrectly. Being able to give greater care when answering medical questions automatically requires methods to classify questions as medical, otherwise health-related, or not health-related at all.

In previous studies of query logs, health-related or medical questions were mostly identified by manual annotation [EK03; SYJ+04]. But manual annotation is expensive and time-consuming. It is therefore often not feasible to manually

Table 2.1: Health-related question answering datasets in comparison to the health-related questions of our dataset.

Dataset	Task	Size	Year	Reference
QA4MRE Alzheimer	reading comprehension	80	2012–'13	[PHF+13]
QALD-4 Biomedical	linked data	50	2014	[UFL+14]
BioASQ	medical professionals	4,249	2014–'21	[BKK+15; BPN+14; KNP+16; NBK+17; NBK+19; NKB+18; NKB+20; NKV+21]
HEAD-QA	job application exams	6,765	2019	[VG19]
MeQSum	summarization	1,000	2019	[BD19b]
MEDIQA 2019 RQE	question entailment	9,120	2019	[BSD19]
Medication QA	medication	674	2019	[BMS+19]
MedQuAD	consumer health	47,455	2019	[BD19a]
PubMedQA	biomedical literature	273,518	2019	[JDL+19]
TREC Health Misinfo.	consumer health	201	2019–'22	[ASL+19; CMS+22; CMS21; CRS+20]
BiQA	consumer health	7,234	2020	[LSC20]
COVID-QA	COVID-19 pandemic	2,019	2020	[MRJ+20]
Health Misbeliefs	alternative remedies	15	2021	[BSD+21]
EPIC-QA	COVID-19 pandemic	60	2022	[GDL+22]
Medical Safety	medical risk assessment	3,777	2022	[AR22]
Ours	only questions	1,990,406	2023	—

annotate larger datasets and impossible in an ad hoc retrieval system. Eysenbach and Köhler [EK03] complemented their manual annotations with a co-ocurrence threshold of search results found for the query and the same query with the added term `health`. They achieve a comparably low F_1-score of 0.50 at a co-occurrence threshold of 0.65 [EK03; Rij79, p. 134]. Liu, Antieau, and Yu [LAY11] developed classifiers for separating professional medical questions from consumer-health questions using bag-of-words, statistical, and linguistic features. Their best classifier achieves an F_1-score of 0.89 on their test set. Though, their models rather decide on the questioner's background (professional or consumer) than on the required expertise of a person or machine answering the question's topic (medical or not medical). Because consumer questions like "Can smoking prevent COVID-19?" are still considered medical questions that should be answered carefully, we cannot use the distinction between consumer and professional questions from Liu, Antieau, and Yu [LAY11] to label questions as medical or not. Their work also does not consider questions that are not health-related.

The aforementioned vocabulary-based approaches [EK03; LAY11] are also prone to errors because most information needs are unique and sometimes use different

vocabulary [BTD+12; DDH07]. Our goal is therefore to generalize classification to a long tail of unseen health-related questions, e.g., using neural classifiers. Kilicoglu et al. [KBM+18] collected a small sample of 2,614 health-related questions from an online question answering system and tag the entities in each of the collected questions. Their entity tags could serve as a good feature for machine learning approaches to generalize to previously unseen questions. But due to the lack of automatic entity tagging, it is unclear how to use entity-based features to identify health-related questions in ad hoc applications.

We see the rise of neural networks and language models [DCL+19; RSR+20; RWC+19; VSP+17] as a promising opportunity to generalize classification of health-related questions to be used for web search and question answering systems. For example, Schlatt et al. [SBH+22] used BERT models [BLC19; DCL+19; GTC+22] to identify health-related cause-effect statements in the CauseNet [HSW+20] corpus, with their best approach yielding an F_1-score of 0.92. Even though their classifiers cannot be directly applied to question classification due to the different syntax of cause-effect statements, their work is motivation to use language models fine-tuning for sentence classification. But training deep neural networks requires large labeled datasets. Table 2.1 compiles an overview of existing datasets with health-related questions released to date. Three reasons speak against directly using these existing datasets to train classifiers for identifying health-related or medical questions: (1) Most of the existing datasets are relatively small (all datasets except for PubMedQA [JDL+19] consist of only a few thousand questions), (2) they often purely focus on medical questions and do not include non-medical but health-related questions, and (3) they entirely lack examples of non-health-related questions. Consequentially, the datasets from Table 2.1 themselves are not representative for the actual distribution of health-related questions in real-world applications [Eys04; SYJ+04]. Because of the lack of negative examples for classification, training on the datasets would yield biased classifiers.

Community question answering platforms like Stack Exchange[1] or Yahoo! Answers[2] contain large numbers of real user questions and have previously been used to analyze how users interact with question answering systems [PWD+12; SCZ08]. Because these platforms feature diverse questions from real users, question answering communities are a good way to complement the smaller medical or health-related datasets from Table 2.1. Similarly, web-scraped general-purpose question answering datasets like GooAQ [KNK+21], SQuaD [RZL+16], or MS MARCO QA [NRS+16] can be supplemented. Combining questions from curated medical datasets, community question answering platforms, and general-purpose question answering datasets, can represent a realistic distribution of health-related (and medical) questions, large enough to train neural models.

[1] https://stackexchange.com/

[2] https://answers.yahoo.com/

For supervised learning, however, we need to label the questions for training. Due to the large dataset size, hand-labeling all questions quickly becomes infeasible. Ratner et al. [RBE+17] proposed the Snorkel framework for labeling large datasets using weak supervision. In Snorkel, label preferences are expressed in the form of simple heuristics, called labeling functions. By combining the output of several labeling functions, the framework estimates each individual labeling function's accuracy and correspondingly returns a weighted combination of the labels [RHD+19]. Training a classifier on the resulting labeled dataset was shown to be similarly effective as using manual annotations [RBE+17]. Yet, because manual annotation of each instance in the dataset is no longer required, the manual labor required to label larger datasets can be substantially decreased. The Snorkel framework has already been successfully applied by Alexander, Kusa, and Vries [AKV22] for large-scale labeling in information retrieval. They use Snorkel to label a sample of 2 million queries from the ORCAS dataset [CCM+20] with the query intent. The resulting labeled dataset was used to train neural classifiers for query intent prediction. The task of labeling health-related questions is conceptionally similar. With metadata about each question's source and by applying simple text matching of health-related terms, we can efficiently infer labels for millions of questions.

Another important aspect when training neural classifiers, is tuning the model's hyperparameters which often have adverse effects on model effectiveness [GBC16, p. 420]. The Weights & Biases framework[3] allows running large hyperparameter optimization experiments on distributed clusters, and to track the results in a web-based dashboard. Because neural models often feature a wide range of hyperparameters, running all configurations in a grid search is often not feasible. Falkner, Klein, and Hutter [FKH18] combined Bayesian optimization and bandit-based methods as a goal-directed parameter search to achieve strong anytime effectiveness and fast convergence to optimal configurations. The efficiency of Bayesian hyperparameter optimization can further be improved by employing the Hyperband algorithm for selecting which hyperparameter configuration should continue to be trained [LJD+17]. Because both aforementioned advances are available in Weights & Biases, Bayesian optimization with Hyperband is a promising tool for hyperparameter optimization.

To summarize the related work on the classification of health-related questions, previous approaches are mostly limited by vocabulary, use manual annotation, or do not consider questions that are not health-related. Properly trained neural classifiers can overcome these limitations but require a diverse dataset for training. After combining many existing datasets, weak supervision can be used to label the resulting dataset. For hyperparameter tuning, Bayesian optimization techniques and the Weights & Biases platform seem promising.

[3] `https://wandb.ai`

2.2 Health-Related Information Retrieval and Question Answering

Once a health-related question has been identified, question answering and information retrieval approaches are used to answer the question and find relevant documents supporting the answer. Health-related question answering systems have already been envisioned since the early 2000s. For example, in 2000, Baorto and Cimino [BC00] proposed a patient health information system to link to relevant web documents. And in 2006, Lee et al. [LCZ+06] suggested question answering systems to be used by medical professionals and highlight the need of more effective and efficient question answering and search systems for medical professionals.

Today, there is a wide range of health or medical search engines. For example, OpenMD[4] searches many government websites, journals and provides consumers with definitions for over 12,000 medical terms. The Trip Database[5] is a popular clinical search engine for finding evidence-based clinical content such as systematic reviews and allows for filtering results based on their quality. Rekabsaz et al. [RLS+21] released the TripClick click log consisting of 5 million user interactions collected in 2013–2020. The click log is complemented with an information retrieval benchmark collection of 692,000 queries that refer to documents from MEDLINE.[6] The TripClick log is the largest health-related benchmark collection publically available to information retrieval researchers. Hofstätter et al. [HAS+22] applied BERT-based re-rankers and dense ColBERT [KZ20] re-ranking to establish stronger baselines for this benchmark. Their results improved upon the original BM25 [RWJ+94] and ConvKNRM [DXC+18] baselines by a large margin. Zerveas et al. [ZRC+22] further improved upon the state of the art for the TripClick benchmark with their CODER framework, using contrastive learning for transforming a query to account for a list-wise context over multiple retrieved (i.e., non-random) hard negative candidate documents. The TREC COVID challenge was a smaller shared task focused on COVID-19-related literature search [VAB+20; WLC+20]. For this benchmark, MacAvaney, Cohan, and Goharian [MCG20] proposed a two-stage zero-shot pipeline using BM25 candidate retrieval and a SciBERT-based neural re-ranker fine-tuned on MS MARCO [NRS+16]. Pradeep, Nogueira, and Lin [PNL21] applied pointwise and pairwise re-ranking with text-to-text language models and achieve the best effectiveness across all automatic submissions for the shared task.

For health-related question answering, the PubMedQA dataset has established as the most influential benchmark [JDL+19]. Here, the goal is to correctly answer yes-no questions mined from titles of medical abstracts from PubMed.[7] On

[4] https://openmd.com

[5] https://tripdatabase.com/

[6] https://nlm.nih.gov/medline/

[7] https://pubmed.gov

the PubMedQA benchmark, large causal language models even outperform human experts in answering health-related questions [JDL+19]. Liévin, Hother, and Winther [LHW22] employed the 175 B parameter GPT-3.5 Codex model in a few-shot chain-of-thought prompting setting, slightly outperforming the human expert answers. With the much larger instruction fine-tuned Flan-PaLM model (540 B parameters), Singhal et al. [SAT+22] achieved the highest accuracy on the benchmark (accuracy: 0.79) in a few-shot prompting setting. Other models used on the PubMedQA benchmark include Galactica, a zero-shot language model trained on multi-modal scientific prompts [TKC+22], and most recently BioMedLM, which has been trained exclusively on the PubMed papers and abstracts from The Pile dataset [BHY+22; GBB+21]. Khashabi et al. [KMK+20] built the pre-trained UnifiedQA model to show that, fine-tuned text-to-text language models can effectively answer questions from diverse datasets. Even though, their model has not been evaluated on the PubMedQA dataset, their smaller model is a good baseline to use for health-related question answering.

The prior work on health-related information retrieval and question answering demonstrates that large language models are often used to answer health-related questions. For information retrieval, a multi-stage architecture with text-to-text models pre-trained for pointwise and pairwise re-ranking is the most effective approach. Question answering is dominated by large language models that outperform smaller language models pre-trained on scientific literature. However, the prior work on health-related information retrieval and question answering is limited to only a few datasets and models. Key challenges in the development of health-related QA systems are the lack of large-scale benchmark datasets, underutilization of domain knowledge such as found in biomedical literature, and answer explainability [JYX+23]. We aim to tackle those challenges by answering questions based on evidence from the PubMed, a large collection of biomedical literature. Modular information retrieval experimentation frameworks such as PyTerrier to compose retrieval pipelines [MTM+21], `ir_datasets` to load information retrieval benchmarks [MYF+21], and `ir_measures` to evaluate approaches [MMO22] nowadays allow us to combine existing retrieval models or datasets as well as to develop new approaches, in order to come closer to the goal of a general-purpose health information retrieval system.

2.3 Claim Verification and Misinformation Detection

With the rising spread of misinformation, especially in health-related topics (see Chapter 1), there is a growing need for systems that detect misinformation and reveal false claims. For claim verification, the goal is to determine whether a claim

can be supported or refuted by some evidence (e.g., news articles or scientific papers) [VR14]. Misinformation detection emerged as the task to detect fake news and disinformation on social media [SSW+17]. Srba et al. [SPT+22] pointed out that misinformation can also include unintentional false information, not only deliberate disinformation. Misinformation detection can thus be summarized as the task of classifying sources as either correct/reliable or incorrect/unreliable. Hence, the tasks of claim verification and misinformation detection are closely related. In the following, we discuss how claim verification can be used to enhance misinformation detection for health-related or medical questions.

The FEVER dataset is the first large scale dataset for claim verification [TVC+18]. It consists of 185,445 claims extracted from Wikipedia that are manually annotated with a label indicating whether the claim is supported, refuted, or not enough evidence is available. Wadden et al. [WLL+20] proposed the specialized task of scientific claim verification that focuses on finding and using scientific literature as evidence to support or refute a claim. Their baseline approach for this task, VeriSci, retrieves literature with TF-IDF [Jon72] and then predicts the claim verification label using BERT sentence embeddings. To facilitate the evaluation of scientific claim verification, Wadden et al. [WLL+20] also released the SciFact dataset with 1,409 expert-written scientific claims. Recently, Wadden et al. [WLW+22] released the stronger MultiVerS model to verify claims based on the shared encoding of the claim and full document context. The model internally uses the LongFormer model to avoid truncating longer scientific articles [BPC20]. The MultiVerS model was trained and evaluated on various scientific claim verification datasets. Wadden et al. [WLW+22] provide fine-tuned model checkpoints based on a model trained on FEVER [TVC+18], PubMedQA [JDL+19], and an evidence inference dataset [LDB+19]: (1) One checkpoint fine-tuned on the COVID-Fact dataset [SCM21] focusing on COVID-19-related claims, (2) a checkpoint fine-tuned on HealthVer [SBM+21] that also focuses on COVID-19, and (3) a checkpoint fine-tuned on the SciFact dataset [WLL+20]. All variants include health-related questions in the training data. Yes-no questions can also easily be reformulated as claims. We therefore experiment with the three fine-tuned checkpoints and the base MultiVerS model to answer health-related yes-no questions.

Misinformation detection has gained more attention recently due to the rapid spread of unreliable news in the early phase of the COVID-19 pandemic [YZK+22]. Assessing the reliability and/or correctness of information sources has been addressed by several datasets and approaches, e.g., by creating health news datasets with real and fake news [DSW20] or by mapping health-related claims to medical articles from reliable and unreliable news sites [SPT+22]. Two directions have emerged from this problem to favor reliable over unreliable sources: Estimating the reliability or correctness of text or finding reliable sources. The former is addressed, e.g., by Fernández-Pichel, Losada, and Pichel [FLP22] who fine-tuned T5

models [RSR+20] and apply sentence embeddings [RG19] similarity to estimate the reliability of health-related passages. For the latter, Zhang [Zha10] employed a logistic regression model to estimate a web domain's trustworthiness, i.e., to find reliable sources similar to Przybyla, Borkowski, and Kaczynski [PBK22]. In our thesis, we propose a combined approach: After determining the true answer based on evidence from PubMed, a trusted dataset of biomedical literature, we indirectly assess the reliability of retrieved web documents by aligning the web documents' answers with the true answer. We apply claim verification models to determine the answers based on web documents and medical articles from PubMed. Recent work also indicates that effective evidence retrieval (e.g., from PubMed) is one key component for correctly answering health-related questions.

2.4 Reducing Misinformation in Health-Related Information Retrieval and Question Answering

With advances in effective health search engines and question answering systems (Section 2.2) as well as in claim verification and misinformation detection (Section 2.3), building information retrieval and question answering systems that actively reduce misinformation is the consequential next step towards improving the access to health-related information online.

The TREC 2019–2022 Health Misinformation tracks [ASL+19; CMS+22; CMS21; CRS+20] provide a platform for evaluating information retrieval and question answering systems with respect to robustness against misinformation for health-related yes-no questions, e.g., about the efficacy of a treatment or medication. The latest track [CMS+22] featured two tasks: (1) Inferring the correct yes/no answer to a health-related question, and (2) retrieving documents that support the correct answer, while preventing documents that support the incorrect answer. Submissions to the answer prediction task were evaluated by their area under the receiver operating characteristic curve (AUC). Retrieval submissions were evaluated by their compatibility [CSV20; CVS20] to "helpful" (documents that give a relevant and correct answer) and "harmful" (incorrect) retrieved results. The objective is to maximize the compatibility score with helpful documents and at the same time minimize the compatibility score with harmful documents [CMS+22].

Several approaches have been proposed to retrieve relevant web documents and not spread misinformation. Abualsaud et al. [ACG+21] simply filtered the document collection to include only health-related content using text classifiers and website quality certifications, significantly improving the effectiveness over using the whole test collection. Fröbe et al. [FGB+22] formulated keyquery-enhanced query expansions based on simulated feedback from medical experts, in order to retrieve most of the helpful documents at high ranks and at the same time reducing

the number of harmful results. Their approach is limited by the availability of (expensive) expert feedback, but they plan to extend the approach with automatic feedback from semantic annotations. Applying argument mining techniques and axiomatic re-ranking to health-related information retrieval has been unsuccessful in the past [BFK+19]. But axioms can describe constraints on preferences between documents [BFR+22] and, due to their inherent explainability, remain an important direction to follow in the future.

Recent work on misinformation-preventing information retrieval at the TREC Health Misinformation tracks has focused on re-ranking retrieved documents to reduce misinformation. Following the same pattern as with their monoT5 and duoT5 re-rankers [PNL21], Pradeep et al. [PMN+21] fine-tuned a T5 text-to-text model [RSR+20] to predict a treatment's effectiveness as either helpful, harmful, or inconclusive given a question about the treatment and a retrieved document. This re-ranking approach sets the state-of-the-art on the TREC 2020 Health Misinformation track (compatibility difference: 0.51) [CRS+20]. In a similar setting, Fernández-Pichel, Losada, and Pichel [FLP22] re-ranked an initial set of passages retrieved using BM25 [RWJ+94] by combining the scores of a monoT5 re-ranker [PNL21] and the estimated document reliability, achieving a maximum compatibility difference of 0.35 on TREC 2020 Health Misinformation topics. Zhang et al. [ZTA+22] also fine-tuned a T5 model [RSR+20] to predict a yes/no answer and employ a logistic regression model to estimate a web domain's trustworthiness. To allow for longer passages to be used for answer prediction, they use a heuristic to select sentences that express a stance towards a yes/no answer. Both their approaches were combined with a BM25 [RWJ+94] retrieval score to re-rank retrieved documents. This trust- and answer-based re-ranking approach achieves a compatibility score difference of 0.13 on TREC 2021 Health Misinformation topics [CMS21]. Comparing the effectiveness across different editions of the shared task is problematic because the tasks used very different queries, e.g., the TREC 2020 Health Misinformation track focused solely on COVID-19-related questions.

Our participation at the TREC 2022 Health Misinformation track [CMS+22] used pre-trained question answering and claim verification models to answer the yes-no questions of the 50 health-related topics, based on top-ranked abstracts from PubMed [BFG+22], we then retrieved relevant documents that support the predicted answer. Pugachev et al. [PAB+23] continued this approach and include Wikipedia as a source of information for the question answering systems. They fine-tune a RoBERTa [LOG+19] model on BoolQ [CLC+19] and BioLinkBERT [YLL22] on PubMedQA [JDL+19] and BioASQ [TSP+12]. Their best approach using Wikipedia articles as context and their RoBERTa-large BoolQ to infer an answer achieves an AUC of 0.82 on 113 health-related yes-no questions [ASL+19; BSD+21; BTS12; CMS21], though we cannot directly compare their results to the TREC Health Misinformation tracks because they use different topics.

2.5 Summary

We have summarized the prior work on three important directions in health-related information retrieval and question answering that we build upon in this thesis. First, we have looked at different approaches to identify health-related questions and their limitations. After reviewing the development and state-of-the-art in health-related information retrieval and question answering systems based on two well-known benchmarks, we then described the challenges of misinformation. Here, we have found similarities in claim verification and misinformation detection and reviewed recent advances in both fields. Finally, we have discussed how misinformation can be prevented in health-related information retrieval and question answering systems by applying claim verification and misinformation detection techniques.

Chapter 3

Data Collection and Labeling

In this chapter, we describe how we combine existing question answering datasets and questions from community question answering platforms to build a large dataset of health-related and medical questions (Section 3.1). Simple yet effective heuristics are proposed to filter non-question texts from this collection. We then propose an approach to automatically label the questions as either being health-related or not, and further distinguish health-related questions on whether they are medical or not (Section 3.2). The overall purpose of this large dataset of questions labeled as non-health, health-related, or medical is to train classifiers to distinguish between these three classes of questions (see Chapter 4). It is therefore essential to approach a label distribution similar to the distribution of questions in online searches [Eys04; SYJ+04]. To facilitate the evaluation of our automatic labeling approach, we therefore manually label a subset of the questions (Section 3.3). Based on the manual and automatic labels, we contribute a dataset of 7,444 manually labeled questions (17 % health-related, 7 % medical) and a larger dataset of 8,544,089 automatically labeled questions (23 % health-related, 15 % medical) and create predefined dataset splits to be used for classification (Section 3.4). Evaluation of the quality of automatic labels compared to manual annotations reveals that our approach achieves a high recall but often falsely labels questions as medical (Section 3.5).

3.1 Collecting Health-Related Questions from Existing Datasets

Three sources of questions are used for our dataset: (1) Curated health-related or medical question answering datasets, (2) general-purpose question answering datasets, and (3) archived posts from community question answering platforms. The curated datasets serve as a source of high-quality questions that are guaranteed to be health-related or medical as the datasets have been created by experts. To approach a realistic distribution of health-related and medical questions in online searches, the curated datasets are complemented with general purpose questions from larger datasets and community platforms which represent a mix of health-related and non-health-related questions. In total, we collect 9,717,648 questions

from 15 curated medical question answering datasets, 4 general-purpose question answering datasets, and 2 community question answering platforms.

3.1.1 Health-Related Question Answering Datasets

A plethora of health-related question answering datasets has been published in recent years. Refer to Table 2.1 for an overview of the 15 datasets we use in this work. These datasets are usually created by manually extracting questions from a specific source, such as frequently asked questions [BSD19] or medical exams [VG19]. Many of the manually curated datasets contain only a few thousand examples due to the high cost of sourcing and labeling the questions. The only larger medical question answering datasets, MedQuAD and PubMedQA, are based on titles of biomedical websites or literature [BD19a; JDL+19]. In the following, we describe each dataset's characteristics in more detail and explain how we collect 348,973 questions from health-related question answering datasets.

QA4MRE Alzheimer's The QA4MRE challenge was a shared task focussing on question answering and machine reading comprehension [PHF+13]. Their 2012 and 2013 editions featured a task on machine reading of biomedical texts with questions about the Alzheimer's disease. Peñas et al. [PHF+13] created the question dataset by first selecting 4 English documents on Alzheimer's disease. Then up to 15 multiple-choice questions were manually created for each document and simplified versions of the full questions are added for some questions. We load the QA4MRE 2012–2013 dataset from the Hugging Face Hub[1] and remove all questions not from the Alzheimer's category, leaving 80 medical professional questions like "What regulates the production of neprilysin?".

QALD-4 Biomedical QALD is an open challenge focusing on question answering over linked data where participants aim to correctly answer a natural language question given a structured web dataset in RDF format.[2] The challenge has been held annually since 2011 [UFL+14]. In 2014, Unger et al. [UFL+14] introduced a task on biomedical question answering with 50 manually curated questions like "Which are possible drugs against rickets?", that were designed to be answerable only by linking information from two pages in the given structured dataset. We download the questions of the QALD-4 challenge's task on biomedical question answering from their website[3] and include all 50 questions.

[1] https://huggingface.co/datasets/qa4mre

[2] https://w3.org/RDF/

[3] https://qald.aksw.org/index.php?x=task2&q=4

BioASQ The first series of shared tasks primarily focusing on biomedical question answering is BioASQ [TBM+15; TSP+12], held annually at CLEF and other conferences since 2013. Each year, the organizers collaborate with biomedical domain experts to build a closed domain corpus of English questions and reference answers. The 20,200 question-answer pairs from the shared task's 2014–2021 editions [BKK+15; BPN+14; KNP+16; NBK+17; NBK+19; NKB+18; NKB+20; NKV+21] can be categorized as either boolean, factoid, list, or summarization questions. They are mostly targeted at medical professionals, for example: "Which CYP gene polymorphism is a well-known predictor of efavirenz disposition?" We download the questions from the task's website[4] and deduplicate the questions because they are often included in multiple editions of the task. From the 4,249 deduplicated questions, we remove sentences that are not formulated as questions (see Section 3.2.1), leaving a total of 3,646 medical questions.

HEAD-QA Vilares and Gómez-Rodríguez [VG19] created the HEAD-QA dataset by manually extracting 6,765 multiple-choice questions from Spanish job application exams. The exams used were designed to test the knowledge of highly trained professionals in the areas of medicine, pharmacology, psychology, nursing, biology, and chemistry. Thus, the questions are of high quality and likely to be asked online by medical professionals. We use the HEAD-QA dataset's English translated version, which is available on the Hugging Face Hub.[5] Sentences not formulated as questions are removed (see Section 3.2.1), leaving 2,522 professional medical questions like "What is estimated with the measurement of skin folds?", that often require professional medical knowledge to answer.

MeQSum Ben Abacha and Demner-Fushman [BD19b] found that while users often formulate long, more complex questions, question answering systems are often better at answering shorter summarized questions that do not include peripheral information. The MeQSum dataset is aimed at the summarization of medical questions which could then be used to improve question answering systems. To create the dataset, three medical experts were asked to summarize 1,000 longer, semantically annotated email requests received by the U.S. National Library of Medicine (NLM) customer service [KBM+18] into shorter questions like "What are the side effects of methadone?" We download the dataset from GitHub[6] and only use the 1,000 summarized questions.

MEDIQA 2019 RQE Another medical question answering dataset by Ben Abacha, Shivade, and Demner-Fushman [BSD19], from the MEDIQA 2019 shared task on

[4] http://participants-area.bioasq.org/datasets
[5] https://huggingface.co/datasets/head_qa
[6] https://github.com/abachaa/MeQSum

recognizing question entailment (RQE), consists of 9,120 question pairs. For each pair, the second question either does or does not entail the first question. The dataset's test set was created by using questions from previous NLM datasets and by mapping frequently asked questions like "What does the prenatal care checklist include?" from NLM websites to manually retrieved short questions from the NLM website [BD16]. Subsequently, the question pairs were manually validated by medical experts. We use only the 9,120 frequently asked questions from the pairs in the test set because the remaining questions are very long and consist of multiple sentences. One empty question was removed, leaving 9,119 questions in total.

Medication QA To address health questions about medications, Ben Abacha et al. [BMS+19] proposed the Medication QA corpus, consisting of 674 question-answer pairs. The dataset is based on anonymized questions from the query log of the MedlinePlus, a U.S. government-maintained health information website targeted at consumers.[7] From the query log, Ben Abacha et al. [BMS+19] selected questions that focus on a drug name as identified using named entity recognition. Questions were also filtered for understandability and answerability. Thus, the dataset contains natural language questions like "how long are you protected after taking the hep b vaccine" alongside the question's annotated reference answer, type, and focus. We download and parse the 690 questions from the author's GitHub repository,[8] that contains 16 questions more than reported in the paper [BMS+19].

MedQuAD Following their previous work on medical question answering datasets, Ben Abacha and Demner-Fushman [BD19a] constructed the larger MedQuAD dataset by crawling topic overview websites from the U.S. National Institutes of Health [BD19a]. They applied manually created patterns to extract questions and answers from the website content, structure, and titles. In total, they extracted 47,455 questions and answers from 12 trusted websites. Most questions were parsed from the MedlinePlus Medical Encyclopedia (17,348 questions, e.g., "What is the outlook for Cobalt poisoning?") and the MedlinePlus Drugs database (12,889 questions, e.g., "What to do in case of emergency or overdose of Acarbose?"), the remaining questions stem from sites that focus on specialized medical fields. We download the dataset from GitHub[9] which contains 47,441 questions, 14 less than reported by Ben Abacha and Demner-Fushman [BD19a].

PubMedQA PubMedQA is a dataset containing 273,518 biomedical yes-no questions derived from titles of PubMed abstracts [JDL+19]. Jin et al. have extracted

[7] `https://medlineplus.gov`

[8] `https://github.com/abachaa/Medication_QA_MedInfo2019`

[9] `https://github.com/abachaa/MedQuAD`

62,249 questions from structured abstracts which contained a question mark in the title and a conclusion in the abstract's text. The remaining 211,269 questions were created based on abstracts with a conclusion, where the title follows a part-of-speech tagging structure of NP-(VBP/VBZ),[10] by moving or prepending auxiliary verbs. By analyzing the MeSH[11] topic distribution of a sample of labeled questions from the PubMedQA dataset, Jin et al. [JDL+19] found that the questions mainly cover studies about human adults in variety of topics. Due to being sourced from biomedical literature, the questions mainly target medical professionals (e.g., "Does network correlate of the cognitive response to levodopa in Parkinson disease?") but some questions could also be asked by consumers (e.g., "Is Friday the 13th bad for your health?"). We download all 273,518 questions of the PubMedQA dataset from the Hugging Face Hub.[12]

TREC Health Misinformation The TREC Health Misinformation tracks focus on answering health-related yes-no questions for which online misinformation is prevalent [ASL+19; CMS+22; CMS21; CRS+20]. For the shared task's four editions from 2019 to 2021, the organizers formulated a total of 201 topics, each consisting of a natural language question (in the topic's `description` field), a keyword query, and the correct answer. The topics cover questions about the efficacy of health treatments that consumers could ask online, e.g., "Can vegan diets be healthy?" The 2020 edition has specialized on the COVID-19 pandemic with 50 topics about treatments and vaccines (e.g., "Can smoking prevent COVID-19?") collected from the World Health Organization's and Harvard Medical School's fact-checking websites [CRS+20]. We download all 201 topics from the task website[13] and use the `description` field of each topic as the question for our dataset.

BiQA Lamurias, Sousa, and Couto [LSC20] automatically extracted health-related questions in the fields of biology, medical sciences, and nutrition from community question answering platforms (Stack Exchange and Reddit). They first selected the top-voted posts as questions from both platforms (only considering post titles that contain a question mark for Reddit). Answers to the selected questions were extracted from the community answers that (directly or indirectly) contained links to PubMed abstracts. We download their first dataset version from GitHub,[14] which contains a total of 7,234 questions and 13,794 question-answer pairs. We subsequently remove sentences that are not formulated as questions (see Section 3.2.1) and use the remaining 4,835 questions in our dataset. Of these filtered questions,

[10] Penn treebank notation [MSM93].
[11] `https://nlm.nih.gov/mesh`
[12] `https://huggingface.co/datasets/pubmed_qa`
[13] `https://trec-health-misinfo.github.io`
[14] `https://github.com/lasigeBioTM/BiQA`

2,766 are about biology (e.g., "What exactly is a centimorgan?"), 1,104 are about medical sciences (e.g., "Does fasting improve your immune system even when you are already having some infection?"), and 965 are about nutrition (e.g., "What cooking oil do you use?").

COVID-QA Following the outbreak of the global COVID-19 pandemic, Möller et al. [MRJ+20] created a dataset of 2,019 question-answer pairs that were manually formulated by 15 biomedical experts based on 147 COVID-19-related scientific articles. As an additional layer of quality assurance, the questions were also subsequently verified by a medical doctor. We download the dataset from GitHub[15] and use the full 2,019 questions, which are mostly about current developments regarding the COVID-19 pandemic at the time of the dataset's creation (e.g., "What are potential vaccines based on?").

Health Misbeliefs In their study of common health-related misbeliefs and how they are answered by major search engines, Bondarenko et al. [BSD+21] derived 15 questions from the Yandex query log. For their study, they selected the 15 most frequent yes-no questions that (1) contained pairs of medical conditions and treatments, and (2) mentioned medicinal plants or alternative remedies (as crawled from Wikidata). After having selected the questions from the Russian query log, Bondarenko et al. [BSD+21] manually translated them into English, and let a medical professional annotate the true answer based on evidence from three medical literature databases, e.g. "yes" for the question "Can green tea reduce blood pressure?". Access to the English questions alongside their expert answer was given by the authors of the study, and we use all 15 questions.

EPIC-QA Goodwin et al. [GDL+22] also motivate their Epidemic QA (EPIC-QA) dataset with the problems that emerged during the COVID-19 pandemic: (1) The fast pace at which new information be generated and (2) the rapidly changing information needs. The EPIC-QA dataset contains 30 expert-annotated questions based on discussions with government officials and clinicians (e.g., "When should an employee suspected or confirmed to have COVID-19 return to work?"). The expert questions were complemented with 30 questions that were extracted from user interactions on MedlinePlus and represent consumer questions (e.g. "How long after I feel better from COVID-19 can I go back to work?"). As the examples show, many of the 60 questions appear similarly in both the expert and consumer subset, leaving the opportunity of comparing the question style across both groups. We download the full dataset from the dataset's official website.[16]

[15] https://github.com/deepset-ai/COVID-QA

[16] https://bionlp.nlm.nih.gov/epic_qa#questions

Medical Safety The most recent health-related question dataset in our collection is an automatically extracted dataset of 3,777 questions with a focus on medical risk assessment [AR22]. Abercrombie and Rieser [AR22] used post titles from the `r/AskDocs` community on Reddit to extract questions about medical decisions, excluding posts with multimedia content. The answers given by conversational agents to the extracted questions were then assessed by crowd workers (861 questions) or domain experts (1,417 questions) for the medical risk of the answer. The dataset was complemented with 1,499 non-medical questions randomly sampled from all Reddit posts. These non-medical questions can be useful as negative examples for training classifiers for health-related questions. We include all 3,777 questions in our dataset, as downloaded from GitHub.[17]

3.1.2 General Purpose Question Answering Datasets

To include negative examples in our dataset, we also include questions from 4 general purpose question answering datasets. We select 4 large datasets that are commonly used for question answering research and that are publicly available. In total, the datasets contain 4,182,463 questions and can therefore fill in missing non-health-related questions for our dataset. Additionally, 2 of the 4 general question answering datasets are tagged with question type or source, which we can use as a signal for labeling questions as health-related or medical (see Section 3.2).

SQuAD Rajpurkar et al. [RZL+16] created a large question answering dataset by first selecting passages from 536 articles sampled from the top-10k of the English Wikipedia, then asking crowd workers to formulate questions based on the passages, and again recruiting crowd workers to answer questions given the corresponding paragraph from Wikipedia. The question-answer pairs were tagged with the article title they stem from. For example, the source of the question: "When was the Suez Canal nationalized?", is tagged as `British_Empire`. The source tags allow for efficient automatic labeling of the questions. We download the SQuAD v1.1 dataset from the Hugging Face Hub[18] and include all 98,169 questions in our dataset.

MS MARCO The MS MARCO dataset [NRS+16] is a popular information retrieval benchmark collection often used in current research, e.g., at the TREC Deep Learning tracks [CMY+21]. It contains 1,010,916 anonymized questions from Bing query logs, associated with 1,026,758 answers curated by crowd workers from passages retrieved from Bing's web index. Because the questions were derived from a search engine query log, they are often not phrased as natural language

[17] `https://github.com/GavinAbercrombie/medical-safety`
[18] `https://huggingface.co/datasets/squad`

questions (e.g., "what are the two types of wind"), but rather as search queries (e.g., "radico share price"). The questions from MS MARCO were also annotated with a label describing the query type: description, entity, numeric, person, or location. For our dataset, we download the questions from the MS MARCO question answering dataset using the `ir_datasets` library [MYF+21],[19] and use only the 651,412 questions that we classify as natural language questions (see Section 3.2.1).

Natural Questions Kwiatkowski et al. [KPR+19] mined a dataset of 315,203 questions from anonymized Google search queries. Questions were collected from queries with a minimum length of 8 words that were searched by multiple users during the crawling, and are subsequently filtered heuristically to remove queries that are not questions. Kwiatkowski et al. [KPR+19] then associate each question with a Wikipedia article in the top-5 results of the Google search engine. Questions for which no Wikipedia article was found were discarded. We download all 315,203 questions using the `ir_datasets` library [MYF+21].[20]

GooAQ By collecting questions from Google's auto-completion log, Khashabi et al. [KNK+21] created a large dataset of popular questions from Google users. They automatically extracted answers from Google's quick-answer box and tagged the questions based on the answer box type as explanatory (e.g., "do fistulas always require surgery?"), list (e.g., "how to get marriage license ontario?"), knowledge (e.g., "what age is mark zuckerberg?"), or conversion (e.g., "50ml is how many ounces?", sub-categories: unit, time, or currency). We download all 3,117,679 questions from the Hugging Face Hub[21] for our dataset. The question types are retained to facilitate automatic labeling of the questions (see Section 3.2), e.g., because time conversion questions are unlikely to be health-related.

3.1.3 Community Question Answering Platforms

We complement our collection of questions from health-related and general question answering datasets with 5,186,212 questions mined from the 2 largest community question answering platforms, Stack Exchange[22] and the now discontinued Yahoo! Answers.[23] Both platforms categorize questions into topic-specific subcommunities, that we can use to automatically label the questions (see Section 3.2).

[19] `https://ir-datasets.com/msmarco-qna`
[20] `https://ir-datasets.com/natural-questions`
[21] `https://huggingface.co/datasets/gooaq`
[22] `https://stackexchange.com`
[23] `https://web.archive.org/web/20210419164027/https://answers.yahoo.com`

Stack Exchange The Stack Exchange platform[24] is an active community question answering platform that, at the time of writing, hosts 173 domain-specific communities on various topics such as technology, sciences, social life, and finance, but also health and medicine.[25] The maintainers openly release a regularly updated, anonymized database dump of all posts on the Internet Archive.[26] We use the dump from May 11, 2022, which contains 8,109,135 posts from 177 community question answering sites (3 were added, 2 removed, and 6 renamed since our download). After downloading the dump, we extract the 2,567,261 post titles that were identified as natural language questions (see Section 3.2.1). For most of the network's communities, the web domain can be used to identify the community's main topic, making post titles from Stack Exchange a valuable addition to our dataset of health-related and non-health-related questions.

Yahoo! Answers Before its dissolution in 2021, Yahoo! Answers[27] has been one of the largest question answering communities. Users could ask questions in 27 different categories to be answered by other users. In 2009, Yahoo! released the Yahoo! Answers Comprehensive Questions and Answers dataset to be used by researchers via their WebScope Program.[28] This large corpus contains all 4,483,032 questions and corresponding answers that were posted on Yahoo! Answers as of October 2007. It has been used to improve and evaluate question answering systems and text classification approaches [SCZ08; ZZL15]. We extract 2,618,951 questions from the Yahoo! Answers dataset by removing posts that were not natural language questions. Similarly to the Stack Exchange dataset, we can use the category in which a question was posted for automatic labeling.

3.1.4 Data Cleaning

After collecting 348,973 questions from health-related question answering datasets, 4,182,463 questions from general question answering datasets, and 5,186,212 questions from community question answering platforms, we merge all collected questions to form a large dataset of 9,717,648 questions (see Table 3.1). Heuristic filters are applied to only retain English, query-like, and anonymized questions.

First, we remove questions that are likely to be spam: Questions with excessive punctuation (i.e., at least 4 punctuation characters in a row) and "yelling" text (i.e., more than 50 % of all characters are uppercase). The remaining questions are then split into sentences and words using tokenizers from the NLTK library [BL04; Por80].

[24] `https://stackexchange.com`

[25] For example, these communities: `health.stackexchange.com`, `fitness.stackexchange.com`, `vegetarianism.stackexchange.com`, `cogsci.stackexchange.com`

[26] `https://archive.org/details/stackexchange`

[27] `https://web.archive.org/web/20210419164027/https://answers.yahoo.com`

[28] `https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=11`

Table 3.1: Source datasets used in our question dataset, original size, questions used, and remaining size after dataset cleaning. The last two column groups denote distributions of positive (✔) and negative labels (✖) from automatic labels of the cleaned corpus.

Dataset	Original (see § 3.1)	Used (see § 3.1)	Cleaned (see § 3.1.4)	Health-rel. (see § 3.2.2)		Medical (see § 3.2.3)	
				✔	✖	✔	✖
QA4MRE Alzheimer's	80	80	74	100 %	0 %	80 %	20 %
QALD-4 Biomedical	50	50	49	100 %	0 %	98 %	2 %
BioASQ	4,249	3,646	3,510	100 %	0 %	75 %	25 %
HEAD-QA	6,765	2,522	1,213	92 %	8 %	55 %	45 %
MeQSum	1,000	1,000	930	100 %	0 %	93 %	7 %
MEDIQA 2019 RQE	9,120	9,119	7,704	100 %	0 %	78 %	22 %
Medication QA	674	690	678	100 %	0 %	82 %	18 %
MedQuAD	47,455	47,441	46,550	100 %	0 %	90 %	10 %
PubMedQA	273,518	273,518	204,692	100 %	0 %	94 %	6 %
TREC Health Misinfo.	201	201	199	100 %	0 %	83 %	17 %
BiQA	7,234	4,835	4,224	85 %	15 %	60 %	40 %
COVID-QA	2,019	2,019	1,912	100 %	0 %	59 %	41 %
Health Misbeliefs	15	15	15	100 %	0 %	80 %	20 %
EPIC-QA	60	60	52	100 %	0 %	83 %	17 %
Medical Safety	3,777	3,777	2,988	66 %	34 %	42 %	58 %
SQuAD	98,169	98,169	93,729	15 %	85 %	4 %	96 %
MS MARCO	1,010,916	651,412	647,081	25 %	75 %	19 %	81 %
Natural Questions	315,203	315,203	312,816	12 %	88 %	8 %	92 %
GooAQ	3,117,679	3,117,679	3,109,801	28 %	72 %	23 %	77 %
Stack Exchange	8,109,135	2,567,261	2,132,629	9 %	91 %	2 %	98 %
Yahoo! Answers	4,483,032	2,618,951	2,086,576	20 %	80 %	7 %	93 %
∑ *Total*	*17,490,351*	*9,717,648*	*8,657,422*	*23 %*	*77 %*	*15 %*	*85 %*

We remove questions that consist of more than 2 sentences (i.e., more sentences than 99 % of all unfiltered questions), fewer than 3 words (to still allow short definition questions like "What is cancer?"), or more than 17 words (i.e., more words than 90 % of all unfiltered questions). Our choices for the limits on the question length are also motivated by commonly reported average lengths for question-like queries in web search engines [BKK+03; PBW07; WRY15; YOA+19]. Furthermore, because mathematical (LaTeX-style) equations and multi-word parentheses are very unlikely to occur in real user queries, we also remove questions that contain such expressions.

We then apply a fastText language identification model [JGB+16; JGB+17] that was trained on Wikipedia, Tatoeba, and SETimes to recognize 176 languages.[29]

[29] `https://pypi.org/project/fasttext-langdetect`

Table 3.2: Types of personally identifiable information removed from our dataset.

Country	Types
Global	credit card numbers, crypto wallets, email addresses, IBAN codes, IP addresses, phone numbers, medical licenses
USA	bank numbers, driver licenses, tax IDs, passports, social security numbers
UK	national health service numbers
Australia	business numbers, company numbers, tax file numbers, medicare numbers

Questions that are tagged with a language other than English with a probability score higher than 50 % are removed from our dataset.

Lastly, because health-related questions often contain privacy-sensitive information, we anonymize the dataset by filtering out questions that contain personally identifiable information. The questions are analyzed for personally identifiable information using Microsoft's Presidio library.[30] Presidio first identifies personally identifiable information using regular expressions and named entity recognition and then validates the identified matches using the surrounding context. The library can be customized for various types of personally identifiable information and has already been used to anonymize medical records with high precision [KSF+22]. We remove questions where the types of personally identifiable information listed in Table 3.2 have been found with a confidence score of at least 0.75.

The resulting dataset contains 8,657,422 anonymized, English, and query-like questions of which 274,790 originate from health-related question answering datasets, 4,163,427 from general question answering datasets, and 4,219,205 from community question answering platforms, as shown in Table 3.1. Even though the two community question answering platforms were the largest source datasets, they contain only a limited amount of questions and large amounts of spam or non-English questions. Consequently, the GooAQ dataset is the largest contributor to our final, cleaned dataset.

3.2 Automatic Labeling using Weak Supervision

We have described how we collected a large-scale dataset of health-related and non-health-related questions from various sources. In this section, we first explain our approach for automatically labeling whether a text is formulated as a (closed) question or not. This heuristic is used to filter out non-questions from our dataset as explained in Section 3.1. Then, we describe how we create heuristic labeling functions and apply weak supervision to label all questions in our dataset as medical

[30] https://github.com/microsoft/presidio

Table 3.3: Interrogative words. Each can be combined with suffixes: –ever, –so, –soever, or be preceded by helper words: after, as, at, for, if, in, on, with.

what	where	which	whom	why
when	whether	who	whose	how

Table 3.4: Auxiliary verbs. Each can be preceded by helper words: after, as, at, for, if, in, on, with.

am	wasn't	doesn't	couldn't	hadn't	needn't	won't
are	were	did	have	may	shall	would
aren't	weren't	didn't	haven't	might	shan't	wouldn't
is	do	can	has	must	should	
isn't	don't	can't	hasn't	mustn't	shouldn't	
was	does	could	had	need	will	

(i.e., requiring answers from medical professionals), otherwise health-related (e.g., about fitness or nutrition), or not related to health. The resulting labeled dataset of questions can serve as training data to identify health-related questions from queries in a search engine (see Chapter 4).

3.2.1 Natural Language Questions and Yes-No Questions

Not every post title on question answering platforms such as Stack Exchange or Yahoo! Answers is formulated as a question. Instead, users often title their posts with a claim, to start a discussion (e.g., the post "On similar concepts in mathematics whose similarity is a non-trivial fact", from Stack Exchange[31]), or with a subject, to ask for inspiration (e.g., the question "Appetizers or desserts…I need your ideas […]", from Yahoo! Answers). Our work focuses only on natural language questions that can appear as search queries. From some included question answering datasets, we therefore filter out sentences that are not formulated as questions, if it is not clear whether the dataset might contain non-question sentences.

Another important question type to label are yes-no questions, that is, questions that can be answered only with "yes" or "no". In health-related contexts, such yes-no questions are often used to ask for the effectiveness of a treatment, e.g., "Does aspirin prevent heart attacks?" Because in Chapter 5 we evaluate our approach on the TREC Health Misinformation track's questions which are yes-no questions, we also develop a labeling approach for yes-no questions. In future work, we plan to build an integrated health search engine that detects such effectiveness-centered

[31] https://mathoverflow.net/questions/116649

```
function is_question(sentence S) is
    if S does not contain a verb then
        return false;
    else if S starts with an interrogative word from Table 3.3 then
        return true;
    else if S does not end with question mark then
        return false;
    else if S starts with an auxiliary verb from Table 3.4 then
        return true;
    else
        return false;
    end
end
```

Algorithm 3.1: Rules for labeling natural language questions.

yes-no questions and then applies customized health-related retrieval approaches to give the correct answer and reduce misinformation in the search results.

Both our labeling approaches for questions are based on simple syntactic rules applied to each sentence in the question text, as split using NLTK [BL04]. Algorithm 3.1 describes how English questions can be labeled using four simple conditions. Our intuition is that questions always start with an interrogative word (Table 3.3) or with an auxiliary verb (Table 3.4). Furthermore, we assume that a valid sentence should at least contain one verb (part of speech tags from NLTK [BL04]). Most questions end with a question mark, and we therefore require the sentence to end with a question mark. But because our dataset includes query-like questions where the question mark is often missing, we also allow questions that do not end with a question mark, as long as they start with one of the interrogative words from Table 3.3. Yes-no questions can similarly be labeled as shown in Algorithm 3.2. Here we flip the condition for interrogative words, because questions that start with an interrogative word are open-ended and thus cannot be yes-no questions.

Table 3.5: Proportion of natural language questions in filtered question answering datasets.

Dataset	Size	Natural Questions	
BioASQ	4,249	3,646	86 %
Yahoo! Answers	3,895,407	2,619,092	67 %
MS MARCO	1,010,916	651,412	64 &
HEAD-QA	6,765	2,522	37 %
Stack Exchange	8,123,683	2,579,633	32 %

```
function is_yes_no_question(sentence S) is
    if S does not contain a verb then
        return false;
    else if S starts with an interrogative word from Table 3.3 then
        ▷ Notice that this condition's output is flipped compared to Algorithm 3.1.
        return false;
    else if S does not end with question mark then
        return false;
    else if S starts with an auxiliary verb from Table 3.4 then
        return true;
    else
        return false;
    end
end
```

Algorithm 3.2: Rules for labeling yes-no questions.

Filtering datasets for natural language questions reveals that only 32 % of the post titles on Stack Exchange are formulated as questions but most of the posts on Yahoo! Answers (67 %) are. A comparison of all 5 datasets that were filtered for questions is given in Table 3.5. Other datasets were not filtered for natural language questions because either the dataset's description or manual inspection revealed that all questions are formulated in natural language. In our current dataset, we currently do not filter for yes-no questions, but we plan to facilitate our labeling approach for yes-no questions in the future to mine health-related yes-no questions such as used in Chapter 5.

3.2.2 Health-Related Questions

After cleaning our dataset, we automatically label whether each question is health-related or not. The resulting labels serve as training data to tune and evaluate health-related question classifiers (see Chapter 4). First, a definition for health-related questions is proposed. Then, we describe how we apply weak supervision for automatic labeling, and describe the heuristic labeling functions that we use as the input to our label model.

Following the World Health Organization's definition of health [46] and common topics in health-related literature (see Chapter 2), we define a question as health-related if it covers any of the following topics:

- physical, mental, and social well-being;
- diseases, illnesses, disorders, or other medical conditions;
- physical or mental states;

- diagnosis, prevention, or risk factors thereof;
- treatments, medication or drugs, or exercises;
- healthcare service or social measures;
- anatomy or biochemical processes; or
- fitness, sports, lifestyle, sex, and nutrition.

We consider health-related questions regarding both humans and animals. But a question is not considered health-related if it is purely navigational (e.g., asking for a website or phone number) or if it is a factual biological question without a clear reference to health (e.g., asking for the structure of a protein but not for its biochemical interactions with the body). Our definition of health-related questions is kept broad to be able to capture diverse health-related information needs that users of a search engine or question answering system might have. Note also, that here we intentionally include non-medical health questions, for example fitness or nutrition tips, where a professional opinion is not always needed (see Section 3.2.3 for a distinction of medical and non-medical questions).

Due to the large number of questions in our dataset, manually labeling all question we collected becomes infeasible. The large amount of data to label is a common bottleneck in machine learning [RHD+19]. Ratner et al. [RBE+17] propose the Snorkel framework to address this problem. In Snorkel, heuristic labels from several labeling functions are combined into a probabilistic label set. A label model then first estimates each labeling function's accuracy. Weighted voting is applied to combine the labels from all labeling functions into a set of probabilistic aggregated labels. The combined labels can be used to train discriminative neural models that generalize beyond the labeling functions' coverage [RBE+17; RHD+19].

Many datasets or subcategories of the datasets we collected only contain health-related questions, e.g., the PubMedQA dataset or the Health category on Yahoo! Answers. Additionally, names of drugs (e.g., Ibuprofen) or conditions (e.g., headache) are almost only used in health-related questions. Other health-related terms like "hospital" indicate a similar signal. Following the aforementioned assumptions, we derive four heuristic labeling functions that mark a question as health-related or not, based on: (1) The question's data source and category, (2) the occurrence of drug names in the question, (3) the occurrence of condition names, and (4) the occurrence of other health-related terms.

Data Source and Category From most of the dataset descriptions, we can directly infer whether a dataset contains health-related questions or not. Many of the datasets also contain metadata about the category or subset of a question. Hence, we implement the first labeling function as a simple lookup table of the question's source and (if applicable) category or subset. This lookup table is shown in the second last column of Table 3.6. If the dataset description does not provide

Table 3.6: Lookup table for labeling a question as health-related (HR) and medical (Med.) based on its source dataset and category or subset. Positive labels are denoted by ✔, negative by ✖, and abstain by ?.

Dataset	Category/Subset	HR	Med.
BioASQ, EPIC-QA, Health Misbeliefs, Medication QA, MEDIQA 2019 RQE, MedQuAD, MeQSum, PubMedQA, QA4MRE Alzheimer's, QALD-4 Biomedical, TREC Health Misinformation		✔	✔
BiQA	medical	✔	✔
	biology	?	?
	nutrition	✔	✖
COVID-QA	—	✔	?
GooAQ	explanatory, knowledge, list	?	?
	conversion	✖	✖
HEAD-QA	medicine, psychology, pharmacology	✔	✔
	nursery	✔	?
	biology, chemistry	✖	✖
Medical Safety	expert-annotated, crowd-annotated	✔	?
	negative examples	✖	✖
MS MARCO	description, entity, numeric	?	?
	person	?	✖
	location	✖	✖
Natural Questions	—	?	?
SQuAD	genome, immunology, infection, pharmacy	✔	✔
	brain, diarrhea, digestion, gene, immune system, mammal, memory, pain, pharma industry, tuberculosis	✔	?
	bacteria, pesticide, symbiosis	?	?
	nutrition	✔	✖
	adolescence, animal, beer, bird, education, emotion, gymnastics, identity, sexual orientation, uranium	?	✖
	others	✖	✖
Stack Exchange	health	✔	?
	biology, cognitive science	?	?
	beer, coffee, fitness, lifehacks, martial arts, open data, outdoor, parenting, sustainability, vegetarianism	?	✖
	others	✖	✖
Yahoo! Answers	health	✔	?
	pregnancy, parenting	?	?
	pets, travel	?	✖
	others	✖	✖

clear information about the contained questions being health-related, we randomly sample 10 questions from each category or subset to infer whether the questions in that dataset, category or subset are mostly health-related. For the questions from Stack Exchange, we also inspect the community homepage which contains a short self-description of the community, to infer a label for the lookup table. An exception is the SQuAD dataset: Because the question categories directly correspond to titles of Wikipedia articles, we infer whether the category's questions are health-related by inspecting the Wikipedia article's content. If we only find very few health-related questions in the examined samples, or if it is otherwise unclear whether the dataset, category, or subset mainly contains health-related questions, we abstain from voting a label for the question. Table 3.7 shows the distribution of labels contributed by the source-based labeling function (see row "Source") in comparison to the other labeling functions.

Health-Related Terms Because health-related questions often use a specific vocabulary with health terms, drugs, and medical conditions, we compile lists of such terms. We create three heuristic labeling functions, one for each term list.

First, we extract 15,425 general health-related terms from OpenMD dictionary[32] and from curated lists of health topics, definitions, and health services provided by the MedlinePlus.[33] Based on an online outline of health sciences[34] and the World Health Organization's list of health topics,[35] we propose a list of 133 additional health terms that were not included in either the OpenMD's or the MedlinePlus's vocabularies and merge all terms to form a vocabulary of 15,558 general health-related terms, e.g., "family therapy" or "pacifier".

A similar approach is used to extract drug names from five drug databases: (1) The European Medicines Agency's public assessment reports database of EU-authorized drugs,[36] (2) the U.S. Food and Drug Administration's National Drug Code Directory containing drug products for sale in the U.S.,[37] (3) the DrugBank's open vocabulary,[38] (4) the SIDER side effects database's list of drug names [KCL+10],[39] and the list of drugs on WikiData.[40] Only commercial brand names (e.g., "Omidria") and synonym generic names (e.g., "ketorolac phenylephrine") are used, resulting in a list of 404,726 drug names.

[32] https://openmd.com/dictionary
[33] https://medlineplus.gov/xml.html
[34] https://en.wikipedia.org/wiki/Outline_of_health_sciences
[35] https://who.int/health-topics
[36] https://ema.europa.eu/en/medicines/download-medicine-data
[37] https://open.fda.gov/data/downloads
[38] https://go.drugbank.com/releases/latest
[39] http://sideeffects.embl.de/download
[40] Instances of the "drug" class: https://wikidata.org/wiki/Q8386

Finally, medical conditions are mined from the WHO's ICD-11, the most recent international classification of diseases [HWJ+21],[41] from the SIDER side effects database [KCL+10], and from the list of medical conditions on WikiData.[42] From the ICD-11, we only include the names of conditions within the dataset's "primary tabulation". In total, 182,664 medical conditions are extracted, that mostly use a medical vocabulary, e.g., "implant site pain" or "hyperkaluria".

Our lists of general health terms, drug names, or condition names are subsequently filtered by removing terms that are too short (i.e., fewer than 3 characters), too long (i.e., more than 5 words or more than 1 sentence; split with NLTK [BL04]), too general (i.e., more than 50 % words are stop words; from NLTK), contain less than 50 % alphabetic characters, or contain parentheses or chemical prefixes. We normalize the remaining terms by lower-casing them, removing accents and non-alphanumeric characters, and stemming each word of the term with the English Snowball stemmer [Por80]. After deduplication, 14,681 general health terms, 74,127 drug names, and 109,117 medical conditions remain.

A first analysis of the filtered terms shows that our lists still contain terms that are only health-related in some contexts, but in others might not be. Examples include names of fruits (e.g., "peach"), chemicals (e.g., "sulfur"), or locations (e.g., "america"). Two annotators therefore inspect terms with shorter than or equal 10 characters and manually collect a list of 2,532 terms that are often not health-related. We remove these terms from our term lists. Our final lists contain 14,125 general health terms, 70,825 drug names, and 104,397 medical conditions.

For each of the three lists, we build a regular expression to search for occurrences of any term from the list at a word boundary in the question text. We use the occurrence count to derive three labeling functions that check whether the question text contains (1) at least one general health term, (2) at least one drug name, or (3) at least one medical condition. Each labeling function abstains from labeling if no term is found in the question text. A comparison of the label distributions as contributed by each labeling function is shown in Table 3.7.

Label Model The four labeling functions for health-related questions are used to train a Snorkel label model [RBE+17] to automatically label our 8,657,422 collected questions as health-related or not health-related. We do not preset the class balance, and train the model for 500 epochs with stochastic gradient descent (constant learning rate of 0.01) and no regularization. Questions where the label model does not return an aggregated label (e.g., due to a tie between the labeling functions) are labeled as not health-related. As shown in Table 3.7, the aggregated label distribution is very similar to the label distributions observed in prior work (cf. Chapter 1) with 23 % of the questions labeled as health-related. Some source

[41] `https://icd.who.int`

[42] Instances of the "health problem" class: `https://wikidata.org/wiki/Q2057971`

Table 3.7: Labels contributed by labeling functions for health-related questions and medical questions. Aggregated labels are returned by Snorkel label models [RBE+17] trained on the labeling functions. Ties of the label model resolve to the negative label.

Function	Positive		Negative		Abstain	
Health-related labeling functions						
Source	426,476	5 %	4,077,916	47 %	4,153,030	48 %
Health terms	1,529,398	18 %	0	0 %	7,128,024	82 %
Drugs	317,350	4 %	0	0 %	8,340,072	96 %
Conditions	528,015	6 %	0	0 %	8,129,407	94 %
⇝ Aggregated (Snorkel)	1,990,406	23 %	6,667,016	77 %	—	—
Medical labeling functions (only health-related questions)						
Source	266,409	13 %	453,987	23 %	1,270,010	64 %
Health terms	0	0 %	461,008	23 %	1,529,398	77 %
Drugs	0	0 %	1,673,056	84 %	317,350	16 %
Conditions	0	0 %	1,462,391	73 %	528,015	27 %
Medical topics	553,506	28 %	99,025	5 %	1,337,875	67 %
⇝ Aggregated (Snorkel)	1,317,873	66 %	672,533	34 %	—	—

datasets contribute more health-related questions than others. Table 3.1 contains a comparison of label distributions per source dataset in the second last column group. Most of the health-related question answering datasets contain close to 100 % health-related questions, indicating that the label model puts a lot of weight on the source-based labeling function. Hence, the questions from general question answering datasets or community platforms follow the global distribution of health-related questions more closely.

3.2.3 Medical Questions

We further distinguish health-related questions into medical questions like "What are the side effects of methadone?" and non-medical questions like "how many litres of water daily?". We define medical questions as health-related questions that require additional, professional expertise, e.g., from a doctor, nurse, pharmacist, or therapist. Our definition includes research questions from clinical studies (i.e., asked from professional to professional) as well as questions from consumers to professionals (e.g., for diagnosis). We exclude questions where the answer would depend on personal preference or that are common sense, even for laymen.

To label questions as medical or not, we apply the same framework as for health-related questions (Section 3.2.2) but slightly tune the labeling functions to account for the different vocabulary used in medical questions. The last column of Table 3.6

shows the labels that are returned by the source-based labeling function that just uses an adjusted (stricter) lookup table. We use the same lists of general health terms, drugs and medical conditions as for health-related questions for counting term occurrences but our term-based labeling functions for labeling medical questions are stricter: Questions that do not contain any general health term, drug name, or condition name, respectively, are labeled as not medical, and we abstain from labeling if a term is found.

Initial experiments showed that with using the four labeling functions alone, many medical questions were not labeled as such. We therefore add a fifth labeling function based on MedlinePlus health topics.[43] The topics are tagged with a category which we use to group topic names and aliases into 227 non-medical terms (categories: "Food and Nutrition", "Wellness and Lifestyle", "Fitness and Exercise", "Health System", "Disasters", and "Safety Issues") and 1,684 medical terms (all other categories). Our labeling function then labels a question as medical if the question text contains a medical term but does not contain a non-medical term. If the question contains a non-medical term but does not contain a medical term, the question is labeled as not medical. The labeling function abstains from labeling questions that do not contain any of our medical or non-medical terms.

Using the five labeling functions for medical questions, we train a Snorkel label model [RBE+17] to label our collected questions as medical or not medical. Because non-health-related questions are, by definition, non-medical, we only consider health-related questions for training the label model. The model is trained for 500 epochs with stochastic gradient descent (constant learning rate of 0.01). No regularization is applied, and we do not preset the class balance. Ties between the labeling functions are resolved by labeling the question as not medical. The aggregated label distribution is shown in Table 3.7. From the health-related questions in our dataset, a relatively high proportion (66 %) is labeled as medical, which we further investigate in Section 3.5. When considering all questions, 15 % are labeled as medical, higher than most proportions reported in prior work (cf. Chapter 1). Comparing the label distributions of questions from each source dataset in Table 3.1, shows that besides the health-related question answering datasets, the GooAQ and MS MARCO datasets contribute many medical questions.

3.3 Manual Annotation

We complement the automatic question labels from Section 3.2 with manual annotations to allow for subsequent evaluation of our labeling approaches (Section 5.5). Post titles from Stack Exchange and Yahoo! Answers are used for manual annotation of natural language questions and yes-no questions. For the manual annotation

[43] `https://medlineplus.gov/xml.html`

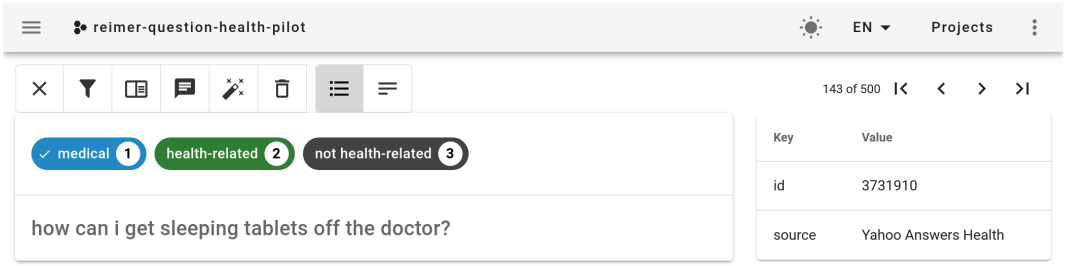


Figure 3.1: Screenshot of the Doccano annotation interface.

of health-related and medical questions, we consider questions from our whole data collection. We recruit eleven volunteer annotators: three PhD students, four Master's students, and one Bachelor's student in computer science, one Bachelor's student in mathematics, one Master's graduate in media and communication science, and one student in human medicine. Even though the annotators are not native English speakers, all are fluent in English. We provide the annotators with instructions for each task and use the Doccano annotation tool[44] shown in Figure 3.1 to facilitate the annotation process. To distinguish the two types of labels, manual and automatic, we refer to the subset of manually annotated questions as "gold" label dataset in the following. Automatically labeled questions are of lower methodical quality than manually labeled questions, and we therefore refer to them as our "silver" label dataset.

3.3.1 Natural Language Questions and Yes-No Questions

For evaluating the automatic labeling of natural language questions and yes-no questions, we randomly sample 2,750 post titles from Stack Exchange and Yahoo! Answers. We create annotator instructions for two sub-tasks: (1) Annotating whether a text is formulated as a natural language question, and (2) annotating whether a text is a yes-no question. The full (revised[45]) instructions are given in Appendix A.1. For annotation, each of our eleven annotators is presented with the texts to annotate in random order on Doccano. We do not allow a text to be annotated as not formulated as a natural language but being a yes-no question at the same time, as explained in the annotator instructions (Appendix A.1). On the annotation platform, we therefore use a single nominal label that encodes the two binary labels of each sub-task. This nominal label is converted to the two binary labels (health-related or not, and medical or not) after annotation.

After giving an initial short explanation of the task at hand, we first conduct a pilot study with 500 randomly sampled post titles. Each question of this first pilot

[44] `https://github.com/doccano/doccano`
[45] That is, the version after revision based on annotator feedback.

study is annotated by each annotator. Two annotators did not participate in the pilot study. It was noted during the pilot study that many post titles contained spam or were not written in English language. We therefore filter the post titles in the same way as described in Section 3.1.4 and only consider the 366 cleaned post titles for the annotation. We measure inter-annotator agreement using Fleiss' κ [Fle71], a measure to assess the degree of annotation agreement over the agreement expected by chance. An agreement of 0.93 is measured for the sub-task of annotating texts as natural language questions, and an agreement of 0.80 is measured for the sub-task of annotating questions as yes-no questions. This reflects almost perfect agreement for identifying questions and substantial agreement for yes-no questions, according to Landis and Koch [LK77]. We mainly find inconsistencies in questions without a question mark (e.g., "What is covid", should be labeled as question), questions with minor grammatical errors (e.g., "Any use for pencils with broken lead?"), and in questions that are formulated like yes-no questions but instead ask for an open answer (e.g., "Can you show me the way to the nearest hospital?", should not be labeled as yes-no question).

Following the pilot study, we organize an online meeting where the inconsistencies are explained to and discussed with all annotators, including one of the two annotators who did not participate in the pilot study. The annotator instructions were revised to account for the discussion. We then sample 2,400 questions for the main annotation round (sampled from cleaned post titles, as described in Section 3.1.4). Each annotator is given 400 questions, of which 200 are the same given to all annotators and 200 are questions only annotated by one annotator. The 200 questions that were annotated by all 11 annotators are used to evaluate the inter-annotator agreement (Fleiss' κ) in the main study. Due to presenting the questions in random order, our annotators were unaware which of their 400 questions were used for measuring agreement. Agreement for annotating natural language questions was measured at 0.92 and annotating yes-no questions yielded an agreement of 0.70. These results confirm that our annotators identify questions with almost perfect agreement [LK77]. The agreement for yes-no question annotation slightly decreased, even though it still represents substantial agreement. Measuring pairwise agreement using Cohen's κ [Coh60] reveals that no agreement ($\kappa < 0$) is found between the two annotators who did not participate in the pilot study. For pairs where at least one annotator participated in the pilot study, we mostly observe fair agreement. Pairwise agreement with the annotator who also missed the discussion was consistently worse than with the second annotator not participating in the pilot study, suggesting that the discussion was helpful for the annotators.

We conduct majority voting to determine final labels for the questions that were annotated by multiple annotators. The author's vote was used to break ties in the vote. Annotations of the remaining questions are included as is to form a gold standard of 2,766 questions for evaluation.

3.3.2 Health-Related and Medical Questions

To facilitate evaluations of our automatic labeling approach for health-related and medical questions, we sample 7,500 questions from our collected dataset (see Section 3.1). From our definition of health-related and medical questions (see Section 3.2), we derive annotator instructions for two sub-tasks of the annotation: (1) Annotating whether a question is health-related, and (2) annotating whether a question is medical. Because a non-health-related question can, by definition, not be medical, we use three nominal labels to encode the two binary labels of each sub-task: Medical, otherwise health-related, or not health-related. The nominal label is converted to the two binary labels of each sub-task after annotation. The full (revised) annotation guidelines are given in Appendix A.2. We use the Doccano annotation tool to present each question to the annotators in random order.

We give the annotators a short explanation of the scope of the annotation task. A sample of 500 questions is then used to conduct a pilot study where each of the 11 annotators is asked to annotate the same questions. We measure an inter-annotator agreement (Fleiss' κ [Fle71]) of 0.77 for the sub-task of annotating questions as health-related, and an agreement of 0.69 for identifying medical questions. According to Landis and Koch [LK77], this reflects substantial agreement for both sub-tasks. For health-related questions, we find inconsistencies for questions centered around housekeeping (e.g., "how long to keep live lobsters?", only concerns the storage of food and should not be labeled as health-related). Questions about social well-being were often wrongly labeled as not health-related (e.g., "at what age does personality stabilize"). Some unobtrusive medical questions were missed (e.g., "Is a 2 pound baby healthy?", severe underweight) and academic questions about specific biochemical processes were sometimes not labeled as medical (e.g., "Are glucose impairment and ghrelin gene variants associated to cognitive dysfunction?").

An online meeting was organized to discuss the inconsistencies with our annotators and to agree on the revised annotator instructions as included in Appendix A.2. Another 7,000 questions were sampled from our dataset for the main annotation round. We assign each annotator to annotate a slice of the questions and include 250 questions to measure inter-annotator agreement after the clarifying discussion. The annotators are again unaware which of their assigned questions are used for measuring agreement. For the main annotation round, we measure an inter-annotator agreement (Fleiss' κ) of 0.78 for health-related question annotation and 0.66 for medical question annotation, indicating that the discussion neither improved nor worsened the agreement. Inconsistencies were mainly observed for questions about nutrition (e.g., "how many litres of water daily?").

Majority voting was applied to determine the final labels for questions that were used for measuring agreement. We use the author's vote in case of a tie. Including the remaining questions as is, we obtain a gold standard dataset of 7,500 questions.

Table 3.8: Comparison of the label distributions of the gold and silver label datasets, and automatic labeling effectiveness. Positive labels are denoted by ✔, and negative by ✖. Labeling effectiveness is measured as (binary) precision (P), recall (R), F_1-score and accuracy (Acc.) using gold labels as ground truth.

Label	Gold		Silver					
	✔	✖	✔	✖	P	R	F_1	Acc.
Question	51 %	49 %	46 %	54 %	0.99	0.89	0.94	0.94
Yes-no question	12 %	88 %	13 %	87 %	0.80	0.87	0.83	0.96
Health-related	17 %	83 %	23 %	77 %	0.66	0.84	0.74	0.90
– training split	16 %	84 %	23 %	77 %	—	—	—	—
– validation split	18 %	82 %	23 %	77 %	—	—	—	—
– test split	17 %	83 %	—	—	—	—	—	—
Medical	7 %	93 %	15 %	85 %	0.40	0.84	0.54	0.90
– training split	7 %	93 %	15 %	85 %	—	—	—	—
– validation split	8 %	92 %	15 %	85 %	—	—	—	—
– test split	8 %	92 %	—	—	—	—	—	—

3.4 Deduplication and Dataset Splits

After collecting the automatically and manually labeled questions, we assign a unique ID[46] to each question based on the question text. We find that 56 questions from our manual annotations for health-related and medical questions were duplicated (0.7 %). From the automatic labels, we find 113,333 duplicates (1.3 %). Duplicates are removed from the datasets, leaving 7,444 manually labeled questions and 8,544,089 automatically labeled questions.

Our annotations for health-related and medical questions are used to create two datasets: A gold label dataset with manually annotated questions, and a silver label dataset with automatically labeled questions. We derive predefined splits to be used for training, validation, and testing of classifiers. Providing predefined dataset splits is a common practice to facilitate comparable evaluation in a leaderboard setting, e.g., with the MS MARCO retrieval tasks [NRS+16].[47] We randomly split our gold label dataset into 4,466 questions for training, 1,489 questions for validation, and 1,489 questions for testing (60 %/20 %/20 %). The silver label dataset is split into 6,835,271 questions for training and 1,708,818 questions for validation (80 %/20 %). We intentionally do not create a test set from the automatically labeled questions because we want to evaluate all classifiers on the same test set. Label distributions

[46] Name-based SHA-1 UUID according to RFC 4122: `https://rfc-editor.org/rfc/rfc4122`

[47] `https://microsoft.github.io/msmarco`

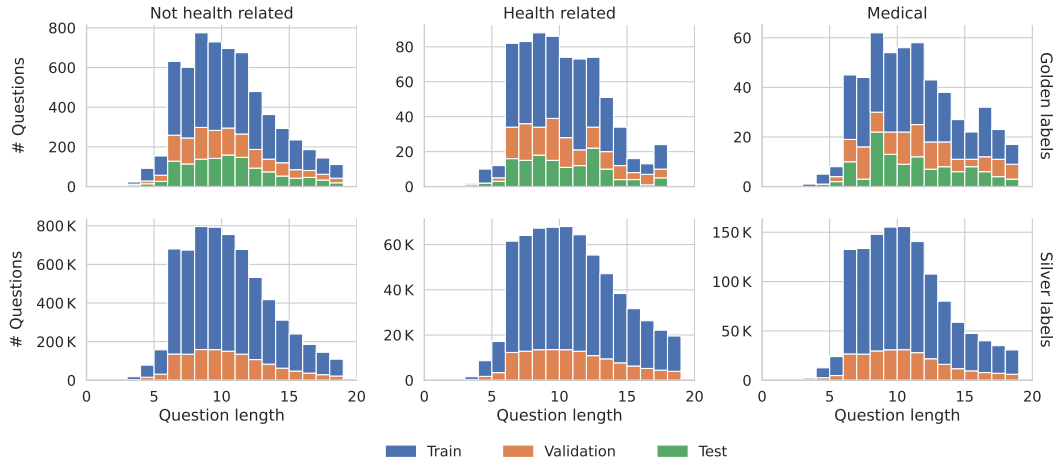


Figure 3.2: Histograms of the question length in our dataset splits for medical, health-related or non-health-related questions. Length is measured as number of words as split by the NLTK tokenizer [BL04; Por80]. Punctuation is counted as a word.

for manual (gold) and automatic (silver) labels are given in Table 3.8. Because the splits were sampled randomly, the label distributions of all splits of either the gold label dataset or the silver label dataset are similar and therefore models trained on the training split should be applicable on the test split.

3.5 Evaluation

To get a better understanding of the questions in our manually annotated (gold) dataset and the automatically labeled (silver) dataset, we first conduct an exploratory data analysis on the questions at hand, and then evaluate the effectiveness of our automatic labeling approaches for health-related and medical questions.

Question Length Previous research has found that health-related queries are longer (5.9 words on average) than other queries (4.2 words on average) [BKK+03]. It was also found that question-like queries are longer than keyword queries, with an average length of 7.4 words [WRY15], which can be attributed to their higher specificity [PBW07]. Consequentially, we expect the average length of health-related questions to be longer than 7 words. Histograms of the question length observed in our dataset are given in Figure 3.2. The histograms indicate that health-related questions are only slightly longer (10.2 words on average on gold labels) than questions that are not health-related (10.0 words). Medical questions contain 10.7 words on average, which is slightly longer than health-related questions. All three distributions are skewed towards shorter questions. Due to our previous

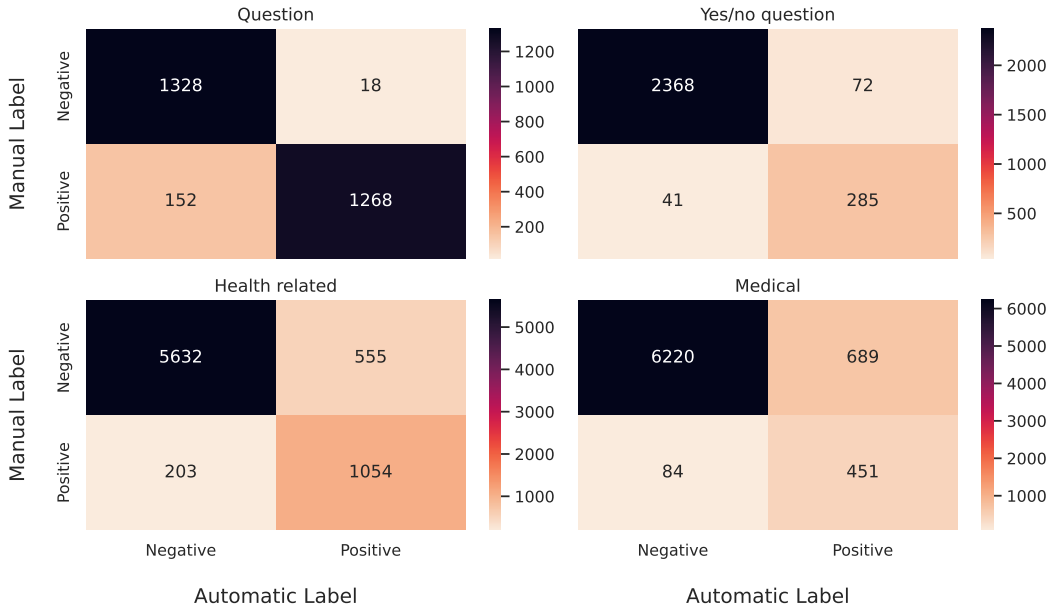


Figure 3.3: Confusion matrices of the automatic labels compared to manual labels.

filtering steps (see Section 3.1.4), we cannot evaluate the characteristics of the questions from the long tail. Slight differences can also be observed between the length distributions of our gold and silver labels. Questions labeled as medical with our automatic labeling approach are shorter (10.1 words on average) than questions that were manually labeled as medical (10.7 words). This discrepancy suggests that the question length could be another useful signal for automatic labeling.

Automatic Labeling Effectiveness We evaluate our automatic labeling approaches by measuring the accuracy of automatically re-labeling the manually annotated questions from the gold label dataset and the most common error types. As a second indicator of effectiveness, we also compare the closeness of the label distributions of our manual and automatic labels for (1) natural language questions and yes-no questions, and (2) health-related and medical questions.

As Table 3.8 demonstrates, the silver label distributions for well-formed questions are very similar to the corresponding gold label distributions, with half of the post titles from Stack Exchange and Yahoo! Answers being labeled as natural language questions by human annotators and our simple labeling heuristic. The heuristic can label questions with a very high precision of 0.99 and still achieve a high recall of 0.89. Consequentially, both the F_1-score [Rij79, p. 134] and the accuracy are high. The confusion matrix in Figure 3.3 supports our findings and illustrates the slightly decreased recall (152 post titles being falsely labeled as not a question). Similar results are found for labeling yes-no questions. With a slightly lower precision

of 0.80, the heuristic still achieves an F_1-score of 0.83 and an accuracy of 0.96. Silver labels for yes-no questions also resemble the gold label distribution more closely, so the confusion matrix is nearly symmetric.

For labeling health-related questions, Table 3.8 shows that only 17 % of the questions were health-related in our gold label dataset, but 23 % were automatically labeled as health-related (silver labels). A similar discrepancy can be observed for medical labels (8 % of the gold labels, 15 % of the silver labels). Even though the Snorkel label models achieve high accuracy for both health-related and medical question labeling, the precision is low, and hence also is the F_1-score. A possible explanation might be that our labeling functions are too broad (include too many false positives) or that the label model did not learn the right label distribution. In practice, however, identifying health-related questions is a recall oriented task, therefore still allowing our automatic labeling approach to be used to train classifiers. The confusion matrices in Figure 3.3 show that our automatic labeling approach only misses 203 questions that are actually health-related and 84 questions that are actually medical. But as the confusion matrix for medical questions and the precision of only 0.42 show, our automatic approach also assigns a medical label to more non-medical questions than to questions that are actually medical. Initial experiments show that for all questions automatically labeled as medical, the health term labeling function abstained from voting. The labeling function could have learned that the question must be medical if the health term labeling function abstained from voting, but further investigation is needed to confirm this hypothesis.

3.6 Summary

We have presented a novel dataset of 8.5 million questions from curated medical question answering datasets, general purpose question answering datasets, and community question answering platforms, an unprecedented scale for datasets of health-related questions. By collecting questions from multiple sources, we ensure a high diversity of health-related questions and provide a realistic distribution of health-related questions. Our manual labels of 7,500 randomly sampled questions indicate that 17 % of the questions are health-related and that 7 % are medical, which reflects a realistic distribution of health-related and medical questions compared to related studies on query logs (5-24 %, cf. Chapter 1). For Research Question 1, we can therefore confirm that a large dataset with millions of health-related or medical questions can be created by combining existing datasets. By comparison with related work, we also confirm that this dataset is realistic in terms of the distribution of health-related questions. We cannot give a clear conclusion to Research Question 2. Even though, we show that automatic labeling using weak supervision can be applied to our large question dataset with decent recall, our

heuristic labeling functions are often too optimistic and therefore lead to a low precision, especially for medical questions. Consequentially, 23 % of the questions from our dataset were automatically labeled as health-related (6 p.p. higher than the real proportion) and 15 % were labeled as medical (8 p.p. higher). Moreover, the low recall can propagate to models trained on the automatic labels (see Chapter 4). In order to reach close to human-level accuracy, we suggest extending our set of labeling functions with additional heuristics that label questions as not health-related or not medical. Our exploratory data analysis of the questions in our dataset is limited to the question length. Because other important characteristics such as topical similarity or the vocabulary were not discussed, we cannot give a clear conclusion to Research Question 3. However, we find that health-related and medical questions are slightly longer than questions that are not health-related. Mainly due to its size our dataset is nonetheless a valuable resource for future research on health-related question answering.

Chapter 4

Identifying Health-Related and Medical Questions

To identify health-related questions, we train three types of classifiers: (1) Feature-based classifiers with using sentence embeddings as features, (2) transformer-based encoder models, (3) causal language models (decoder-only transformer), and (4) text-to-text language models (encoder-decoder transformer). An overview of all model variants used in our classification experiments is given in Table 4.1. We only experiment with classifiers that can be fitted to our question dataset, and use the predefined training, validation, and test splits as described in Chapter 3. Rule-based approaches are therefore not included. Separate models are fine-tuned on two tasks: (1) To distinguish between health-related and non-health-related questions and (2) to distinguish between medical and non-medical questions. This additional distinction is motivated by the need for stricter measures against misinformation when answering medical questions.

For each type of classifier and task, we tune hyperparameters on the validation set using the Weights & Biases framework [Bie20]. Bayesian hyperparameter optimization with the Hyperband algorithm is employed for faster convergence to optimal parameter configurations [FKH18; LJD+17]. After hyperparameter optimization on the validation set, the most effective configurations from each type of classifier and for each of the two classification tasks (health-related and medical) are evaluated on the test set. We also evaluate a cascading setting where questions are first classified as health-related prior to medical classification, to avoid falsely classifying non-health-related questions as medical. In the following sections, we first outline the training settings (Sections 4.1, 4.2 and 4.3). We then describe our cascading classification approach (Section 4.4) and evaluate the most effective models from each type of classifier with respect to effectiveness and efficiency (Section 4.5). Finally, the chapter is concluded with a short comparison to prior approaches (Section 4.6).

Table 4.1: Neural models used in our approaches for classifying health-related and medical questions. Variant and parameters indicate which model was used in this thesis. Links to model checkpoints are provided in Appendix B.

Type	Model	Variant	Params	Year	Ref.
Sentence embedding models (see Section 4.1)	MiniLM	all-round	33 M	2019	[RG19]
	MiniLM	paraphrase	33 M	2019	[RG19]
	MPNet	all-round	110 M	2019	[RG19]
	Instructor	base	335 M	2022	[SSK+22]
Encoder models (see Section 4.2)	BERT	base, uncased	110 M	2019	[DCL+19]
	RoBERTa	base	125 M	2019	[LOG+19]
	SciBERT	uncased, SciVocab	110 M	2019	[BLC19]
	BART	base	140 M	2020	[LLG+20]
	PubMedBERT	base, uncased, abstract	110 M	2022	[GTC+22]
	BioLinkBERT	base	110 M	2022	[YLL22]
Causal language models (see Section 4.3)	GPT-2	small	124 M	2019	[RWC+19]
	GPT-Neo	—	125 M	2021	[BLW+21]
	OPT	—	125 M	2022	[ZRG+22]
	Galactica	—	125 M	2022	[TKC+22]
	BioGPT	—	347 M	2022	[LSX+22]
	BioMedLM	—	2,700 M	2022	[BHY+22]
Text-to-text language models (see Section 4.3)	T5	base	220 M	2020	[RSR+20]
	SciFive	base, PubMed	220 M	2021	[PAT+21]
	LongT5	TGlobal, base	250 M	2022	[GAU+22]
	Flan-T5	base	220 M	2022	[CHL+22]

4.1 Training Feature-Based Classifiers with Sentence Embeddings

Our first batch of classifiers uses feature-based machine learning models: (1) Support vector machine [CL11], (2) random forest [Bre01] (both implemented in scikit-learn [PVG+11]), and (3) gradient boosting (implemented in XGBoost [CG16]). We use pre-trained sentence embeddings from the `sentence-transformers` library [RG19] and instruction-fine-tuned text embedding models [SSK+22] with custom prompts to embed each question in a vector space. The aforementioned classifiers are then trained to distinguish between health-related and non-health-related questions and between medical and non-medical questions respectively.

We use only the gold label dataset for training the feature-based classifiers. Initial experiments showed that the silver label dataset is too large to be trained with either of the three classifiers, because the implementation in scikit-learn loads full

Table 4.2: Hyperparameter prior distributions for training support vector machine, random forest, or gradient boosting classifiers to identify health-related or medical questions.

Parameter	Distribution	Values
Embeddings model	uniform cat.	`all-MiniLM-L6-v2`, `paraphrase-MiniLM-L6-v2`, `all-mpnet-base-v2`, `instructor-base`[a]
Class weights	uniform cat.	balanced, imbalanced
Support vector machine		
Regularization C	log-uniform	0.1 — 1000
Kernel	uniform cat.	linear, polynomial x^2, x^3, x^4, x^5, x^6, RBF, sigmoid
Kernel coefficient γ	log-uniform	0.01 — 100
Random forest		
Number of trees	log-uniform	1 — 100
Max tree depth	log-uniform	1 — 100
Min samples to split	log-uniform	2 — 20
Min samples in leaf	log-uniform	1 — 10
Gradient boosting		
Number of trees	log-uniform	10 — 1000
Learning rate	log-uniform	5e-3 — 1
Max tree depth	log-uniform	1 — 500
Min loss reduction γ	log-uniform	1e-3 — 100
Min child weight	log-uniform	1 — 100
Subsampling by tree	uniform	0 — 1
Subsampling by level	uniform	0 — 1
Subsampling by node	uniform	0 — 1

[a]Prompt: `Represent the question for health related classification: [question]` or `Represent the question for medical classification: [question]`

training dataset into memory multiple times (e.g., for each decision tree). Attempts to train the classifiers on sub-samples of approximately 10 % of the silver label dataset have also failed. From the gold label dataset, we use the predefined training split to train each model, and the validation split to tune hyperparameters. The test set is only used after completion of hyperparameter optimization to evaluate the most effective model configuration on the test set (see Section 4.5).

We use the Bayesian Optimization and Hyperband algorithm [FKH18] from the Weights & Biases framework to determine the best hyperparameters for each model by targeting high effectiveness with respect to the F_1-score [Rij79, p. 134]. In contrast to random search, this Bayesian hyperparameter approach samples initial hyperparameter configurations from a prior distribution and evaluates the F_1-score on the validation set. Based on the evaluation results, the hyperparameter

distributions are iteratively updated to favor configurations with higher F_1-scores. The prior distributions for our feature-based classifiers are specified in Table 4.2 and are based on the recommended hyperparameter choices for text classification in scikit-learn and XGBoost [CG16; PVG+11]. For each of the three model types, the hyperparameter optimization is performed with 25 runs, each time drawing new parameters from the hyperparameter distributions.

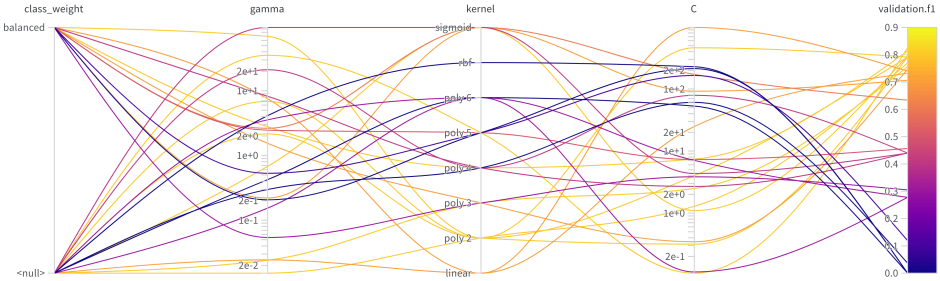
Our trained random forest classifiers always predicted the majority classes (not health-related or not medical, respectively). Consequentially, all trained random forest classifiers yield a high accuracy of 0.84 or 0.93 but an F_1-score of 0.00 on the validation set. The feature importances of the trained random forest classifiers (i.e., accumulated impurity decrease within each tree) shows that all components of the sentence embedding vectors were ignored. Consequentially, the classifier never predicts the positive classes (health-related or medical, respectively). Due to the failed training of random forest models, we do not report results for random forest classifiers in our further analyses.

Figure 4.1 shows parallel coordinates plots of all hyperparameter configurations trained for the support vector machine (SVM) and gradient boosting classifiers. To simplify the visualization, we select only the four most important parameters as axes of the parallel coordinates plot. The parameter importance is inferred from a random forest model's feature importance after training the model to predict the target metric based on the parameter values, implemented in Weights & Biases [Bie20]. The color of each line indicates the F_1-score on the validation split with better F_1 appearing in brighter colors. Using the parallel coordinates plots, we can identify optimal ranges for each hyperparameter and particularly well- or bad-performing hyperparameter configurations. For example, in Figure 4.1a and 4.1b the blue lines running through the polynomial kernels of degree 3 and the RBF kernel indicate that a more complex kernel function might harm the classification effectiveness of SVM classifiers. Figures 4.1c and 4.1d both show that balanced class weights are beneficial for gradient boosting classifiers and that a minimum loss between 0.01 and 1 is optimal for splitting nodes in the gradient boosted trees.

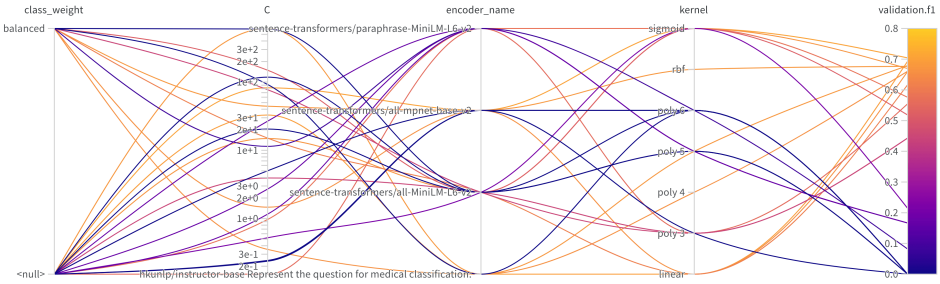
For classifying questions as health-related or not, the most effective SVM classifier[1] achieves an F_1-score of 1.00 on the training set and 0.83 on the validation set. Our most effective gradient boosting classifier[2] achieves very similar F_1-scores of 1.00 on the training set and 0.75 on the validation set. Both models are overfitting the training set as indicated by the perfect F_1 on the training set and the lower F_1-score on the validation set. From our medical question classifiers, the

[1]Embeddings: `instructor-base`; regularization: 0.11; kernel: x^5; coefficient: 35; class weight: imbalanced.

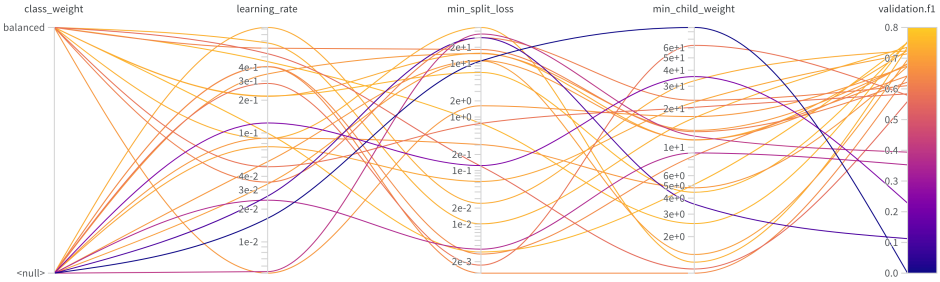
[2]Embeddings: `all-mpnet-base-v2`; trees: 89; maximum depth: 4; learning rate: 0.45; minimum child weight: 2.53; minimum split loss reduction: 0.76; feature column subsampling by level: 37 %; by node: 65 %; by tree: 34 %, class weight: balanced.



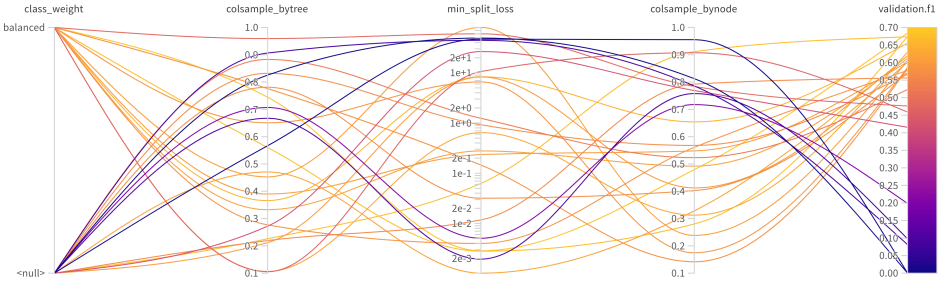
(a) SVM, health-related questions.



(b) SVM, medical questions.



(c) Gradient boosting, health-related questions.



(d) Gradient boosting, medical questions.

Figure 4.1: Parallel coordinates plots of validation F_1-scores with different hyperparameter configurations when training support vector machine or gradient boosting models to classify health-related or medical questions. Configurations with higher F_1 appear brighter.

Table 4.3: Hyperparameter prior distributions for fine-tuning encoder models to identify health-related or medical questions.

Parameter	Distribution	Values
Pre-trained model	uniform cat.	`bert-base-uncased`, `roberta-base`, `bart-base`, `scibert_scivocab_uncased`, `BioLinkBERT-base`, `BiomedNLP-PubMedBERT-base-uncased-abstract`
Weight decay	uniform	$0 - 0.5$
Epochs	uniform	$1 - 30$
Batch size	uniform	$1 - 10$
Class weights	uniform cat.	balanced, imbalanced
Optimizer	fixed	Adam $\beta = (0.9, 0.999)$ [KB15]
Learning rate	log-uniform	1e-6 — 1e-3
Scheduler	uniform cat.	linear, cosine
Warm-up steps	uniform	$0 - 5000$

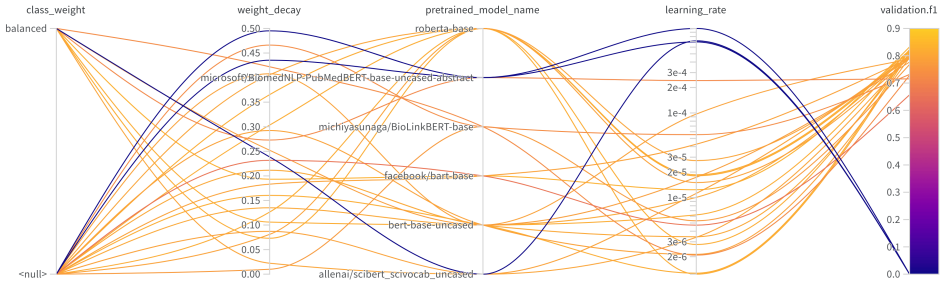
most effective SVM classifier[3] achieves an F_1-score of 0.83 on the training set and 0.70 on the validation set, whereas the most effective gradient boosting classifier[4] is overfitting slightly more with an F_1-score of 0.96 on the training set and 0.68 on the validation set. For both, SVM and gradient boosting classifiers, the best model parameters are substantially different between classifiers for health-related and medical questions, indicating that the optimal hyperparameters for feature-based classifiers are task-specific. The two most effective SVM classifiers and the two most effective gradient boosting classifiers are evaluated in more detail in Section 4.5.

4.2 Fine-tuning Transformer-Based Encoder Models

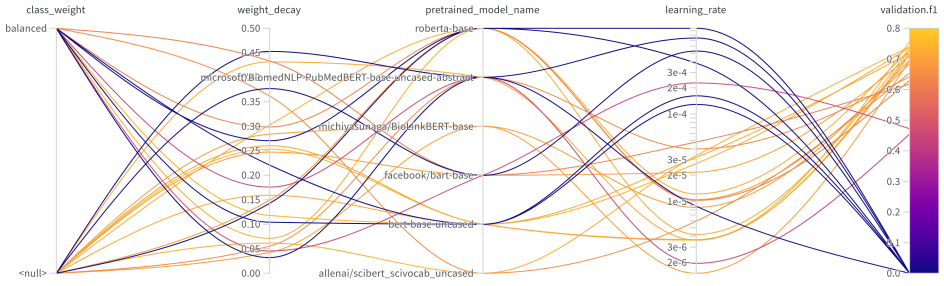
To classify texts with transformer-based encoder models [VSP+17], a linear classification layer is added on top of the encoder's pooled output [DCL+19]. By applying this pattern, we fine-tune pre-trained bidirectional encoder models on the downstream task of classifying questions using the `transformers` library [WDS+20]. The classification models are fine-tuned using gradient descent on a shared cluster of 24 Nvidia A10 GPUs. We use the same Bayesian Optimization and Hyperband framework as for the feature-based classifiers [FKH18]. This Bayesian hyperparam-

[3]Embeddings: `all-mpnet-base-v2`; regularization: 82; kernel: sigmoid; coefficient: 0.04; class weight: imbalanced.

[4]Embeddings: `all-MiniLM-L6-v2`; trees: 973; maximum depth: 73; learning rate: 0.017; minimum child weight: 17.78; minimum split loss reduction: 0.003; feature column subsampling by level: 40 %; by node: 48 %; by tree: 74 %; class weight: balanced.



(a) Health-related questions.



(b) Medical questions.

Figure 4.2: Parallel coordinates plot of validation F_1-scores with different hyperparameter configurations when fine-tuning encoder models to classify health-related or medical questions. Configurations with higher F_1 appear brighter.

eter optimization samples hyperparameters from a prior distribution and after each run updates the distributions to favor configurations with higher F_1-scores. We perform the hyperparameter optimization with 25 runs for each model and use the prior distributions specified in Table 4.3 based on recommended hyperparameter ranges from the `transformers` library [WDS+20]. All our runs use the predefined training, validation, and test splits of the gold label dataset.

Figure 4.2 shows parallel coordinates plots of the four most important parameters. Parameter importance is again inferred from a random forest model's feature importance trained to predict the target metric based on the hyperparameter values [Bie20]. By far the most influential hyperparameter for classifying health-related questions or medical questions is the learning rate. Higher learning rates consistently lead to worse effectiveness with respect to the F_1-score on the validation set, which could be explained by the learning rates being too high for the gradient descent optimization to converge. We also observe lower F_1-scores when fine-tuning the PubMedBERT model [GTC+22]. Applying balanced class weights to the models' loss tends to improve classification effectiveness for health-related questions, but a slight negative correlation was found for medical questions. Both the number of epochs and the batch size are less important for classification effectiveness.

The most effective encoder model to classify health-related questions, a fine-tuned BART model [LLG+20],[5] achieves an F_1-score of 0.88 on the training set and 0.83 on the validation set, i.e., the model nearly does not overfit. The best parameters classifying medical questions were different. Here, the most effective model, based on BERT [DCL+19],[6] overfits the training set slightly more with an F_1-score of 0.89 on the training set and 0.76 on the validation set. The different best parameters indicate that (like the feature-based classifiers), the best hyperparameter choice depends on the task (classifying health-related questions or medical questions).

We use a combination of well-performing hyperparameters of the encoder models fine-tuned on the smaller gold label dataset to fine-tune two BERT models [DCL+19] on the larger silver label dataset:[7] one model for identifying health-related questions and one model to classify medical questions. The hyperparameters were selected based on the most effective models that were trained on the gold label dataset are used, but slightly adapted to respect our findings from parameter importances and to make the fine-tuning of the two new models more reproducible. We only fine-tune the two models for three epochs, because the silver label dataset is much larger than the gold label dataset. For classifying health-related questions, this approach yields an F_1-score of 0.98 on the validation set (silver labels) and an F_1-score of 0.97 for medical questions. Due to the large size of the training set (6.8 M questions), effectiveness was only measured on the validation set (1.7 M questions). Hence, we cannot conclude if the model overfits.

The classification effectiveness of our two most effective encoder models fine-tuned on the gold label dataset and the two encoder models fine-tuned on the silver label dataset are evaluated in more detail in Section 4.5, to evaluate if the effectiveness of the models can be improved by fine-tuning on the larger, automatically labeled silver label dataset instead of the manually annotated gold label dataset.

4.3 Fine-tuning Transformer-Based Text Generation Models

Our third family of question classifiers explores using the first output token of transformer-based text generation models (i.e., causal and text-to-text language models) to predict the question's label, an approach that has previously been successfully used for text classification using the T5 text-to-text language model [PNL21;

[5] Model: `bart-base`, weight decay: 0.19, epochs: 3, batch size: 10, class weights: balanced, optimizer: Adam $\beta = (0.9, 0.999)$, learning rate: 1e-5, scheduler: linear, warm-up steps: 1,513.

[6] Model: `bert-base-uncased`, weight decay: 0.25, epochs: 2, batch size: 1, class weights: imbalanced, optimizer: Adam $\beta = (0.9, 0.999)$, learning rate: 4e-6, scheduler: cosine, warm-up steps: 921.

[7] Model: `bert-base-uncased`, weight decay: 0.2, epochs: 3, batch size: 128, class weights: balanced, optimizer: Adam $\beta = (0.9, 0.999)$, learning rate: 1e-5, scheduler: linear, warm-up steps: 1,500.

Table 4.4: Hyperparameter prior distributions for fine-tuning text-to-text language models to identify health-related or medical questions.

Parameter	Distribution	Values
Pre-trained model	uniform cat.	`t5-base`, `long-t5-tglobal-base`, `flan-t5-base`, `SciFive-base-Pubmed`
Prompt (health related)	uniform cat.	`Question: [question] Health related?`, `Question: [question] Is this question health related?`, `Question: [question] Is this question about something health related?`
Prompt (medical)	uniform cat.	`Question: [question] Medical?`, `Question: [question] Is this question medical?`, `Question: [question] Is this question about something medical?`
Labels	uniform cat.	`true`/`false`, `yes`/`no`
Weight decay	uniform	$0 - 0.5$
Epochs	uniform	$1 - 30$
Batch size	uniform	$1 - 4$
Optimizer	fixed	Adam $\beta = (0.9, 0.999)$ [KB15]
Learning rate	log-uniform	1e-6 — 1e-3
Scheduler	uniform cat.	linear, cosine
Warm-up steps	uniform	$0 - 5000$

RSR+20]. Text-to-text language models are based on a standard encoder-decoder transformer architecture [VSP+17] and are trained to decode the target text given an encoded prompt. Raffel et al. [RSR+20] suggest using the model's input to prompt a simple task that specifies how the text should be classified. The generated output is then used as the classified label. With this simple approach, they achieve an accuracy of 83.28 on the GLUE classification benchmark but also highlight a potential issue: If the model generates a token that is not a valid class label, it is not clear what label should be assigned to the input. Pradeep, Nogueira, and Lin [PNL21] address this issue by fine-tuning a T5 text-to-text model to predict either `true` or `false` as the first output token of the text-to-text model when given a prompt like `Query: [query] Document: [document] Relevant?` For inference, Pradeep, Nogueira, and Lin [PNL21] look at the decoder's first output token probability instead of directly decoding the text. The softmax function [GBC16, pp. 180–184] is applied after masking all other token IDs except for the predefined `true` and `false` tokens to constrain the model to predict one of the target tokens. The classifier then returns the token with the highest probability as class label, and consequentially always predicts a valid label. We adapt this approach for classifying questions by using different prompts (e.g., `Question: [question] Is this`

Table 4.5: Hyperparameter prior distributions for fine-tuning causal language models to identify health-related or medical questions.

Parameter	Distribution	Values
Pre-trained model	uniform cat.	`gpt2`, `opt-125m`, `gpt-neo-125M`, `galactica-125m`, `biogpt`, `BioMedLM`
Prompt (health related)	uniform cat.	`Question: [question] Is this question health related?`, `Question: [question] Is this question about something health related?`, `Question: [question] Is this question health related, yes or no?`, `Question: [question] Is this question about something health related, yes or no?`
Prompt (medical)	uniform cat.	`Question: [question] Is this question medical?`, `Question: [question] Is this question about something medical?`, `Question: [question] Is this question medical, yes or no?`, `Question: [question] Is this question about something medical, yes or no?`
Labels	uniform cat.	`true`/`false`, `yes`/`no`
Weight decay	uniform	$0 - 0.5$
Epochs	uniform	$1 - 30$
Batch size	uniform	$1 - 4$
Optimizer	fixed	Adam $\beta = (0.9, 0.999)$ [KB15]
Learning rate	log-uniform	1e-6 — 1e-3
Scheduler	uniform cat.	linear, cosine
Warm-up steps	uniform	$0 - 5000$

`question health-related?`) and target label tokens (e.g., `yes` and `no`) as listed in Table 4.4. But the same method is used to extract a prediction from the text-to-text language model output token probabilities.

We also transfer this approach to causal language models like GPT-2 [RWC+19]. Causal language models are decoder-only transformer models trained to predict the next token that follows a sequence of given input tokens. For inference, we pass the prompt to the model and then look at the output logits of the next token, i.e., the first predicted token after the prompt. The softmax function is again applied after masking all other token IDs except for the target label tokens, to predict the token with the highest probability. During fine-tuning, the target label token is appended to the prompt and the model is fine-tuned using cross-entropy loss. Hyperparameters for fine-tuning causal language models are given in Table 4.5.

The same Bayesian Optimization and Hyperband framework [FKH18] is used as for the encoder models. Optimization is performed with 25 runs for each model by

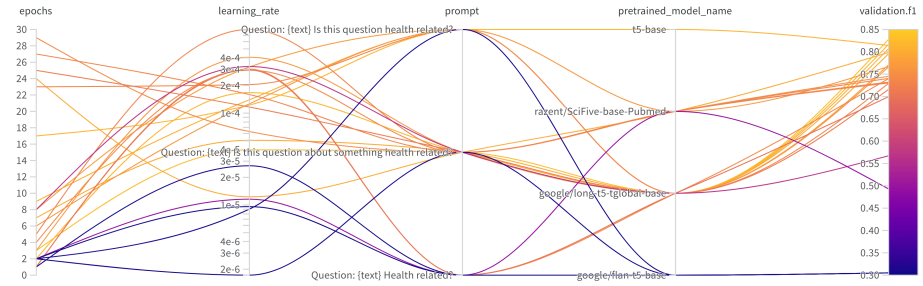
drawing parameters from the hyperparameter distributions specified in Table 4.5 and Table 4.4. The predefined training split of the gold label datasets is used to fine-tune the classifiers and the models are evaluated based on their F_1-score on the validation split. Due to time constraints we do not fine-tuned language models on the larger silver label dataset. Hyperparameter optimization was run for one week on a shared cluster of 24 Nvidia A10 GPUs. With causal language models, 23 runs were fine-tuned successfully for classifying health-related questions and 22 for medical questions. For text-to-text language models, there were 23 successful runs for classifying health-related questions and 23 for medical questions. Even though we had to stop hyperparameter tuning before all planned 25 runs per model were completed, the amount of runs is still large enough to draw conclusions about the best hyperparameters for each classification task.

We show the four most important hyperparameters for each model in Figure 4.3. Choosing the best pre-trained model checkpoint is the most important parameter to tune for text classification and depends on the classification task. Of the text-to-text language models, for example, the Long-T5 model [GAU+22] performs best to identify health-related questions, but SciFive [PAT+21] works better to distinguish between medical and non-medical questions. From the causal language models, the BioGPT model [LSX+22] achieves the best F_1-score on the validation set for both tasks. A longer, clarifying prompt is also beneficial for the text-to-text language models. Shorter prompts similar to the prompts used by Pradeep, Nogueira, and Lin [PNL21] are less effective for question classification. Most of the successful models required ten or more epochs of fine-tuning. Text-to-text models required higher learning rates for fine-tuning than causal language models. Causal language models also yield higher F_1-scores with higher weight decay. The diverse parameter settings of the most effective models show that hyperparameter optimization is important when fine-tuning transformer-based text generation models to classify health-related or medical questions.

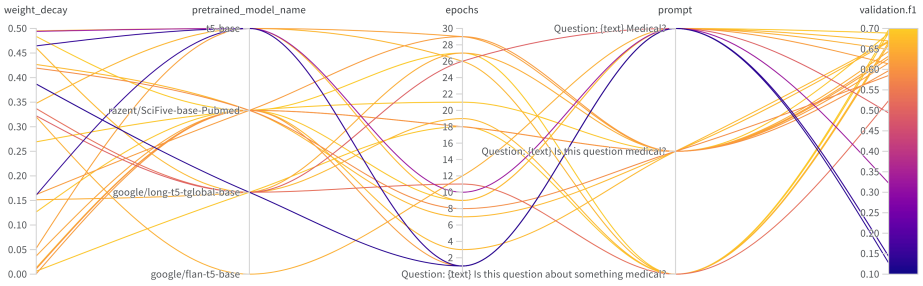
Our most effective text-to-text language model for classifying health-related questions, a fine-tuned Long-T5 model [GAU+22],[8] achieved an F_1-score of 0.88 on the training set and 0.83 on the validation set (nearly no overfitting). For classifying medical questions, the most effective text-to-text model was a fine-tuned SciFive model [PAT+21],[9] overfitting more severely with an F_1-score of 1.00 on the training set and 0.70 on the validation set. Our most effective causal language models for identifying health-related and medical questions are both based

[8] Model: `long-t5-tglobal-base`, prompt: `Question: [question] Is this question about something health-related?`, labels: `true/false`, weight decay: 0.46, epochs: 3, batch size: 2, optimizer: Adam $\beta = (0.9, 0.999)$, learning rate: 4e-5, scheduler: cosine, warm-up steps: 1,610.

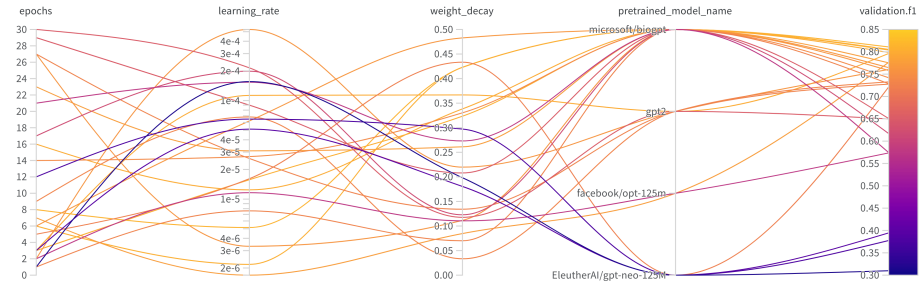
[9] Model: `SciFive-base-Pubmed`, prompt: `Question: [question] Is this question about something medical?`, labels: `yes/no`, weight decay: 0.48, epochs: 27, batch size: 2, optimizer: Adam $\beta = (0.9, 0.999)$, learning rate: 3e-5, scheduler: linear, warm-up steps: 3,895.



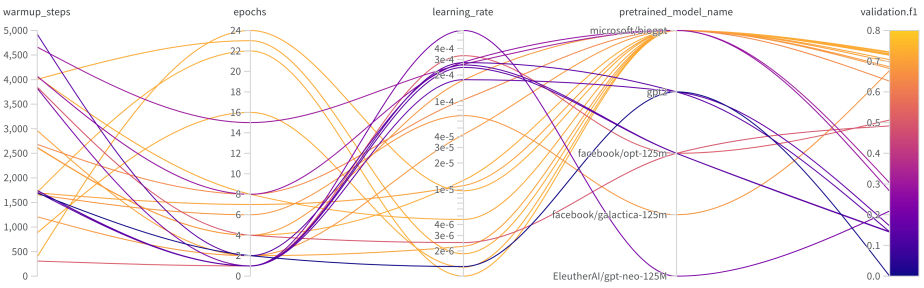
(a) Text-to-text language models, health-related questions.



(b) Text-to-text language models, medical questions.



(c) Causal language models, health-related questions.



(d) Causal language models, medical questions.

Figure 4.3: Parallel coordinates plot of validation F_1-scores with different hyperparameter configurations when fine-tuning language models to classify health-related or medical questions. Configurations with higher F_1 appear brighter.

on BioGPT [LSX+22]. Here, the best model to classify health-related questions[10] achieves an F_1-score of 0.02 on the training set and 0.81 on the validation set, indicating that the model does not learn the training data. The most effective causal language model classifier for medical questions[11] behaves similarly and yields an F_1-score of 0.01 on the training set and 0.73 on the validation set. The training logs of both models reveal that the accuracy of both models increases during training but the precision, recall, and F_1-score decrease. Due to time constraints, we were not able to investigate the cause of this behavior, but it seems likely that the models predict the majority class (not health-related or not medical, respectively) too often.

4.4 Cascading Classification Strategies

We propose two cascading classification strategies to avoid an error of our medical classifiers: The classifiers sometimes classify a question as medical even if it is not health-related. But by our definition of medical questions in Chapter 3, a question must always be health-related to be medical. All our models return prediction scores for each class that sum up to 1, and we thus interpret the scores as probabilities, i.e., $P(H)$ for classifying health-related questions and $P(M)$ for classifying medical questions. The aforementioned constraint can also be formulated in reverse: A question that is not health-related cannot be medical. Consequentially, the probability $P(M = 1,\ H = 0)$ should equal 0. However, our medical classifiers sometimes classify non-health-related questions as medical, i.e., $P(M = 1,\ H = 0) > 0$. Our cascading strategies multiply the scores of the health-related and medical question classifiers (inspired by mixture of experts models [Bis06, pp. 672 sqq.]) to correct the probability $P'(M)$ such that $P'(M = 1,\ H = 0) = 0$. We first correct the conditional probability as follows:

$$P'(M = m \mid H = h) = \begin{cases} 0 & \text{if } h = 0 \\ P(M = m) & \text{if } h = 1 \end{cases}$$

[10] Model: `biogpt`, prompt: `Question: [question] Is this question about something health-related?`, labels: `yes/no`, weight decay: 0.42, epochs: 3, batch size: 3, optimizer: Adam $\beta = (0.9, 0.999)$, learning rate: 5e-6, scheduler: cosine, warm-up steps: 1,195.

[11] Model: `biogpt`, prompt: `Question: [question] Is this question medical?`, labels: `yes/no`, weight decay: 0.31, epochs: 22, batch size: 2, optimizer: Adam $\beta = (0.9, 0.999)$, learning rate: 1e-6, scheduler: linear, warm-up steps: 1,709.

With this correction of the conditional probability, the constraint that questions which are not health-related cannot be medical $P'(M = 1,\ H = 0) = 0$ holds true:

$$\begin{aligned} P'(M = 1,\ H = 0) &= \frac{P'(M = 1 \mid H = 0)}{P(H = 0)} \\ &= \frac{0}{P'(H = 0)} = 0 \end{aligned}$$

We then use the corrected conditional probability to compute the corrected probability $P'(M)$ for medical questions:

$$\begin{aligned} P'(M = 1) &= P'(M = 1,\ H = 0) + P'(M = 1,\ H = 1) \\ &= P'(M = 1 \mid H = 0) \cdot P(H = 0) \\ &\quad + P'(M = 1 \mid H = 1) \cdot P(H = 1) \\ &= 0 \cdot P'(H = 0) + P'(M = 1) \cdot P(H = 1) \\ &= P(M = 1) \cdot P(H = 1) \end{aligned}$$

Our cascading classification framework uses two different ways to compute the probability $P(H)$ of a question being health-related: (1) Manual ground-truth labels as a way to simulate the best achievable effect of our cascading strategy, and (2) the predicted scores from our health-related question classifiers to evaluate the effect achievable in real-world applications.

4.5 Evaluation

We evaluate the most effective feature-based classifiers (Section 4.1), transformer-based encoder models (Section 4.2), and transformer based text generation models (Section 4.3) for both health-related and medical question identification. Each of the most effective models is evaluated in terms of classification effectiveness and efficiency. We measure classification effectiveness by area under the receiver operating characteristic curve, as well as accuracy, precision, recall, and F_1-score to predict the correct class (health-related vs. not health-related or medical vs. not medical). Inference efficiency is measured by throughput, i.e., the number of questions that can be classified by the model per second on a high-end laptop (Intel i7 8-core CPU, no GPU used, measured time includes data loading, embeddings, and inference). Recall that the classifiers are targeted at identifying health-related or medical questions in a search engine, in order to let the system use specialized retrieval approaches (e.g., misinformation detection). An ideal model should be able to identify most of the health-related or medical questions but should also not classify too many false positives. We therefore primarily focus on the F_1-score of the classifiers, but also consider a high throughput to be beneficial for application

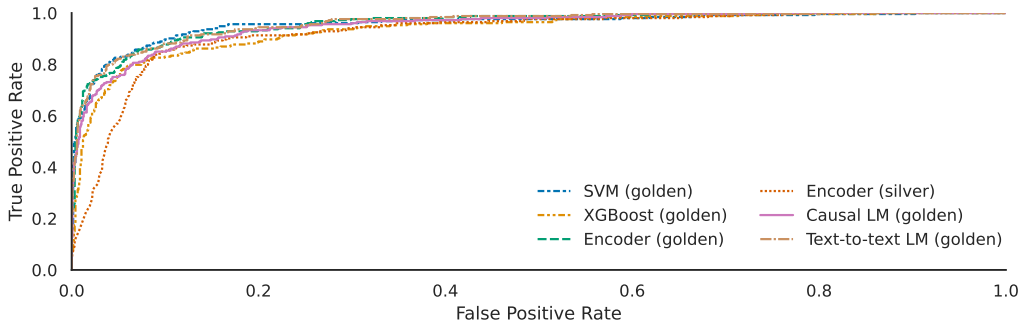
Table 4.6: Classification performance of our most effective models trained to identify health-related or medical questions. Effectiveness is reported with respect to area under the ROC curve (AUC), accuracy (Acc.), precision (P), recall (R), and F_1-score on the test set. Throughput is measured for inference on an Intel i7 8-core CPU. Best results per task are highlighted in bold.

Model	**Effectiveness**					**Efficiency**
	AUC	Acc.	P	R	F_1	Throughput
Classifiers for health-related questions						
SVM (gold labels)	0.96	0.93	0.85	0.74	0.79	41.6/s
XGBoost (gold labels)	0.93	0.92	0.80	0.70	0.75	**63.8/s**
Encoder (gold labels)	0.96	**0.94**	**0.88**	0.73	0.80	24.4/s
Encoder (silver labels)	0.92	0.90	0.67	**0.81**	0.73	29.8/s
Causal LM (gold labels)	0.95	0.92	0.82	0.72	0.76	5.5/s
Text-to-text LM (gold labels)	**0.96**	0.94	0.87	0.74	**0.80**	6.8/s
Classifiers for medical questions						
SVM (gold labels)	0.96	**0.96**	**0.74**	0.64	**0.69**	68.6/s
XGBoost (gold labels)	0.97	0.95	0.70	0.62	0.66	**408.6/s**
Encoder (gold labels)	**0.97**	0.95	0.67	0.64	0.65	29.9/s
Encoder (silver labels)	0.93	0.90	0.42	**0.81**	0.55	27.7/s
Causal LM (gold labels)	0.96	0.95	0.68	0.67	0.68	5.8/s
Text-to-text LM (gold labels)	0.94	0.94	0.63	0.60	0.61	9.7/s
Cascading classifiers for medical questions (using ground-truth health-related labels)						
SVM (gold labels)	**0.99**	**0.96**	**0.81**	0.64	**0.72**	—
XGBoost (gold labels)	0.98	0.96	0.79	0.62	0.70	—
Encoder (gold labels)	0.98	0.95	0.72	0.64	0.68	—
Encoder (silver labels)	0.97	0.93	0.52	**0.81**	0.63	—
Causal LM (gold labels)	0.98	0.96	0.72	0.67	0.69	—
Text-to-text LM (gold labels)	0.98	0.95	0.73	0.60	0.66	—
Cascading classifiers for medical questions (using health-related predictions)						
SVM (gold labels)	**0.98**	**0.96**	**0.79**	0.61	**0.69**	25.9/s
XGBoost (gold labels)	0.97	0.95	0.72	0.58	0.64	**55.2/s**
Encoder (gold labels)	0.97	0.95	0.69	0.60	0.64	13.4/s
Encoder (silver labels)	0.94	0.90	0.42	**0.80**	0.55	14.4/s
Causal LM (gold labels)	0.97	0.95	0.68	0.67	0.68	2.8/s
Text-to-text LM (gold labels)	0.96	0.95	0.69	0.60	0.64	4.0/s

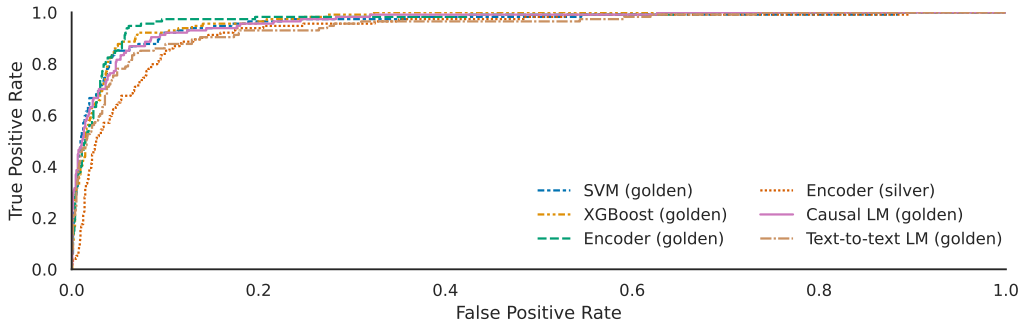
in ad hoc applications. Our random forest classifiers always predict the same class and are therefore not considered for further evaluation. Our cascading classification framework is evaluated in both variants, using manual labels and using the predictions from our health-related question classifiers.

Table 4.6 shows an overview of the classification performance of our most effective models for both classification tasks. We report three classification settings: (1) Identifying health-related questions from a set of questions that could be queried by users, (2) directly identifying medical questions, and (3) identifying medical questions only within the pool of previously identified health-related questions. A cascade of two classifiers – one for health-related questions and one for medical questions – can be beneficial because a question that is not health-related can by definition never be medical. We evaluate cascades that either use the manual ground-truth label to rule out non-health-related questions or use the predictions of the most effective health-related question classifier of the same type (e.g., best SVM classifier for health-related questions when identifying medical questions with an SVM). We also report the receiver operating characteristic (ROC) curves (Figure 4.4) and precision-recall curves (Figure 4.5) for all evaluated classifiers. In the health-related question classification setting, our text-to-text language model fine-tuned on the gold label dataset achieves the best F_1-score (0.80) and area under the ROC curve (AUC: 0.96), indicating that fine-tuned text generation models can successfully be used for binary classification of questions. But as text generation models were the slowest models in our evaluation, using them in a real-world search system might not be desirable. The most effective encoder model fine-tuned on the gold labels is comparably effective but four times faster. Fine-tuning the encoder model on the larger silver label dataset instead of the gold label dataset increases recall (0.81) at the cost of lower precision (0.67). The encoder model's ROC curve in Figure 4.4a reveals that it yields more false-positives even at lower thresholds of the true positive rate. Except for the encoder model fine-tuned on the silver label dataset, all model types have a similar precision-recall curve. Even the simpler and faster embedding-based SVM classifier performs well to identify health-related questions. It achieves an F_1-score of 0.79, only slightly worse than our most effective neural models. One explanation for the worse effectiveness of our classifiers trained on the silver label dataset is that the automatic silver labels are not very precise (labeling precision: 0.66, F_1: 0.74; cf. Table 3.8) and the encoder model trained on these labels (precision: 0.67, F_1: 0.73; cf. Table 4.6) is limited by the quality of automatic labels.

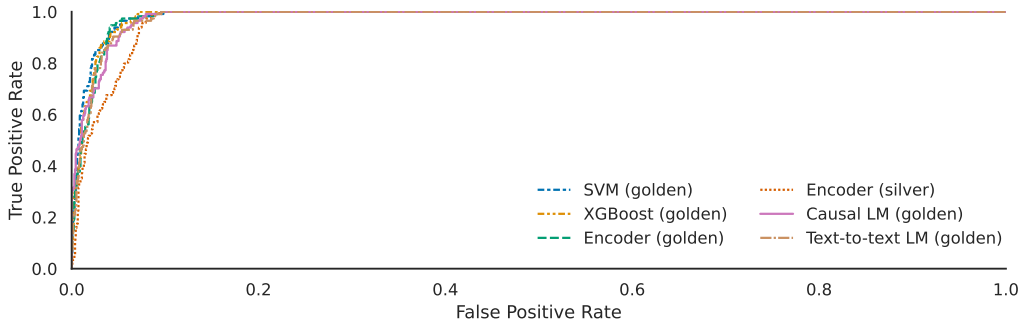
SVM classifiers using sentence embeddings are also the most effective models to identify medical questions, and yield an F_1-score of 0.69. Because medical questions are rarer and often use a complex vocabulary, classification effectiveness is generally worse than for health-related questions. This is also reflected in the precision-recall curves in Figure 4.5b, where all model types have a lower precision at high recall



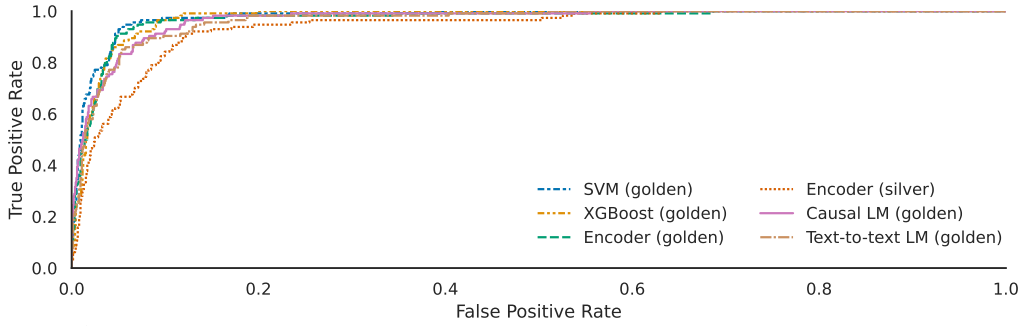
(a) Classifiers for health-related questions.



(b) Classifiers for medical questions.

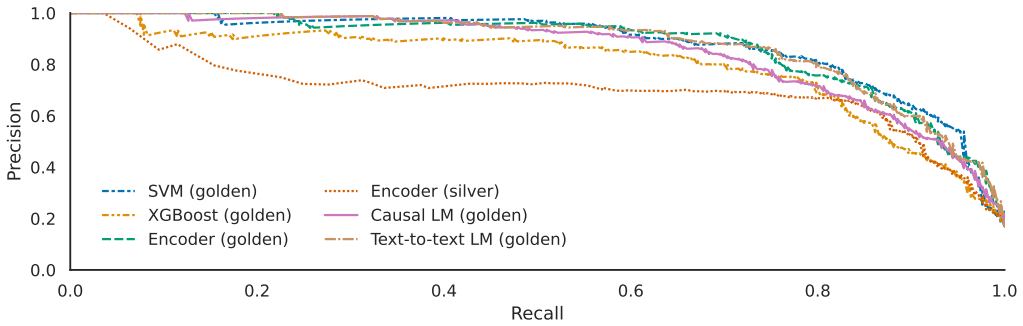


(c) Cascading classifiers for medical questions (using ground-truth health-related labels).

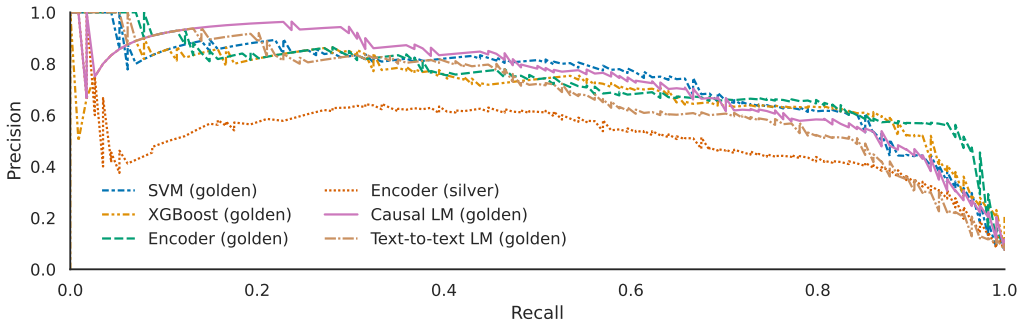


(d) Cascading classifiers for medical questions (using health-related predictions).

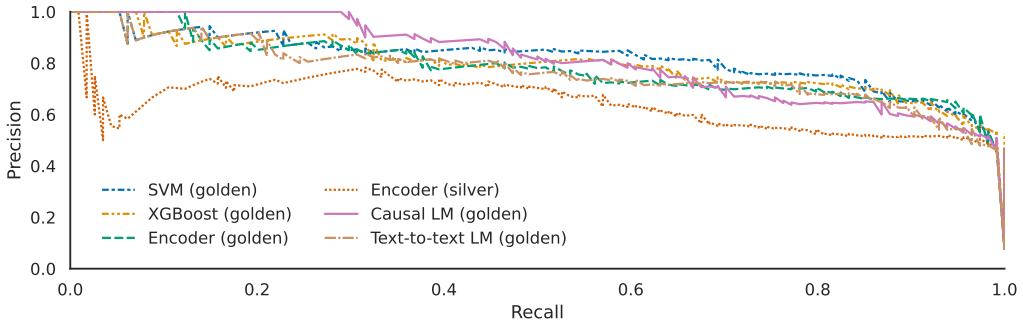
Figure 4.4: Receiver operating characteristic curves of our most effective models to identify health-related or medical questions. Better models have a larger area under the ROC curve, perfect models touch the upper left corner (i.e., FPR: 0, TPR: 1).



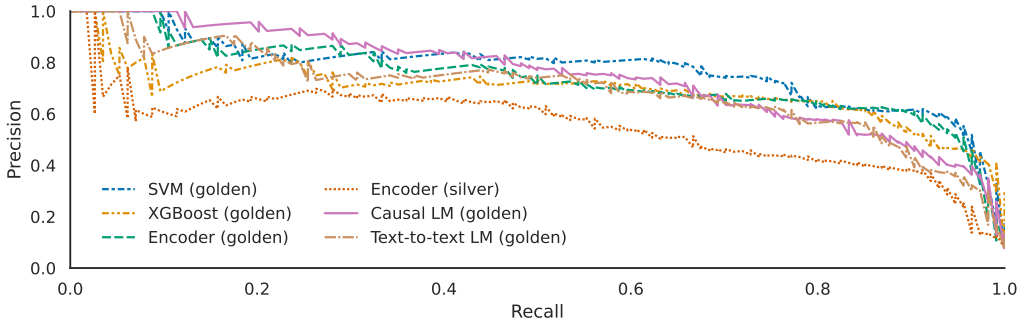
(a) Classifiers for health-related questions.



(b) Classifiers for medical questions.



(c) Cascading classifiers for medical questions (using ground-truth health-related labels).



(d) Cascading classifiers for medical questions (using health-related predictions).

Figure 4.5: Precision-recall curves of our most effective models to identify health-related or medical questions. Better models have a larger area under the precision-recall curve, perfect models touch the upper right corner (i.e., recall: 1, precision: 1).

thresholds. The best recall (0.81) is achieved by the encoder model fine-tuned on the larger silver label dataset at the cost of a low precision of only 0.42, that is likely limited by the automatic labeling quality (precision: 0.40; cf. Table 3.8). Because of this limitation, we expect future improvements in the automatic labeling process to propagate to the encoder models that are fine-tuned on these silver labels. Gradient boosting also yields good effectiveness results for classifying medical questions, and is only slightly outperformed by the SVM classifier. But with a throughput of 409 questions per second, it is by far the fastest evaluated model in our experiments. The high efficiency and decent effectiveness, again highlights the usefulness of sentence embeddings for fast feature-based classification.

Our cascading classification strategy for medical questions is evaluated in two scenarios: (1) With manual ground-truth labels and (2) with predictions from the health-related question classifiers. Using manual labels, all 1,236 non-health-related questions from the test set are marked as not medical, leaving only the 253 health-related questions to be further classified as medical or not. As Figure 4.4c shows, this pre-filtering caps the worst false positive rate that could be achieved while identifying all true positives correctly (TPR: 1) at 17 % (i.e., if all remaining questions are classified as medical). In the scenario of using predictions from the health-related question classifiers, the slightly deteriorated predictions from actual health-related question classifiers reduce the beneficial effect of cascading classification. But we still observe higher effectiveness for identifying medical questions. The most effective cascade model, our SVM classifier trained on the gold label dataset, achieves a precision of 0.79 and a recall of 0.79. Compared to the same medical question classifier without cascade, this reflects a precision improvement of 5 percent points but decline in recall of 3 percent points. Because the effects of the cascade are marginal even when using the manual ground-truth labels to pre-filter questions that are not health-related, we conclude that the cascade is not beneficial for identifying medical questions. An important consideration is that we trained our model for identifying medical questions on the full (gold or silver label) datasets, i.e., with a skewed label distribution. Another promising approach could be to limit training of medical question classifiers to only the health-related questions. This would reduce the number of negative training examples and thus balances the label distribution. Our datasets allow for such an approach, but due to time constraints, we leave further analyses for future work.

4.6 Summary

We have presented efficient and effective approaches to identify both health-related and medical questions using a variety of feature-based and transformer-based models. Hyperparameter tuning with Bayesian Optimization and Hyperband allowed

us to find the best parameters for each type of model and task in just 25 iterations. Our approaches based on sentence embeddings proved to be both effective and efficient for identifying health-related and medical questions. For classifying medical questions feature-based approaches even outperform much larger encoder models. From our models to identify health-related questions, we found that the most effective model is a text-to-text language model fine-tuned on the gold label dataset, that achieves an F_1-score of 0.80 on the test set. Hence, we improve upon the rule-based prior approaches by Eysenbach and Köhler [EK03] (F_1: 0.50). Advances by Schlatt et al. [SBH+22] who use BERT models to classify cause-effect statements as health-related are not directly comparable to our work, as they do not classify questions but rather statements. Nonetheless, their reported F_1 effectiveness of 0.92 is considerably higher than our best model. Our most effective model for identifying medical questions is an SVM using sentence embeddings as features. This SVM model is also fine-tuned on the gold label dataset and achieves an F_1-score of 0.69. Cascading approaches for identifying medical questions are not beneficial in our setting. Prior approaches, e.g. by Liu, Antieau, and Yu [LAY11] (F_1: 0.89), is often not comparable to our approach due to the different definition of medical questions (i.e., we have labeled based on the expertise needed to answer the question while others used labels based on the background of the questioner).

With respect to our Research Question 4, it is hard to compare against prior work because no benchmark dataset exists yet for the task of identifying health-related questions or sentences. Nonetheless, our best classifiers achieve a good F_1-score of 0.80 for health-related questions and a reasonable F_1-score of 0.69 for medical questions. At least for health-related questions, we can therefore conclude that our best approach (fine-tuning a text-to-text language model) is effective. It is also reasonably effective, with a throughput of multiple questions per second on a consumer-grade CPU, even though our evaluations do not consider parallelization or GPU acceleration. Our medical question classifier is even faster but not as effective. Research Question 5 concerns the effectiveness of training encoder models on larger, automatically labeled datasets. We find that our models closely approximate the label distribution of our silver label dataset, and hence, achieve a higher recall than models trained on the smaller gold label dataset (0.81 for both label types, highest of all evaluated models). This increase in recall comes at the cost of lower precision and F_1-scores and we therefore can neither conclude that automatic labeling increases nor that it decreases the effectiveness of encoder models. We leave further analyses of the effectiveness of automatic labeling for future work.

Chapter 5

Answering Health-Related Yes-No Questions

In the previous two chapters, we have demonstrated how health-related or medical questions can be identified in search engines. We now discuss approaches to answer these questions. In particular, we focus on answering health-related yes-no questions about treatments for diseases, the type of questions used in the TREC Health Misinformation tracks [ASL+19; CMS+22; CMS21; CRS+20]. Answering open questions is left as an open challenge for future work.

We first build a modular pipeline using evidence retrieved from PubMed (Section 5.1) to answer health-related yes-no questions (Section 5.2). After the topic answer is found, our approach retrieves supporting web documents to back up the answer (Section 5.3) and re-ranks the documents to favor correct information over misinformation (Section 5.4). In view of increasing calls for improving the reproducibility of information retrieval research [Lin22], our pipeline consists of self-contained modules that each serve a single purpose. The PyTerrier framework [MTM+21] is utilized to implement and compose the modules, and we use Elasticsearch[1] to build our search indices. In Figure 5.1, we give an overview of our four-stage pipeline for inferring the correct answer and ranking web documents to reduce misinformation. The stages are explained in the first sections of this chapter.

For evaluation, we then discuss how this modular retrieval pipeline has been used in our participation in the TREC 2022 Health Misinformation track, and explain the

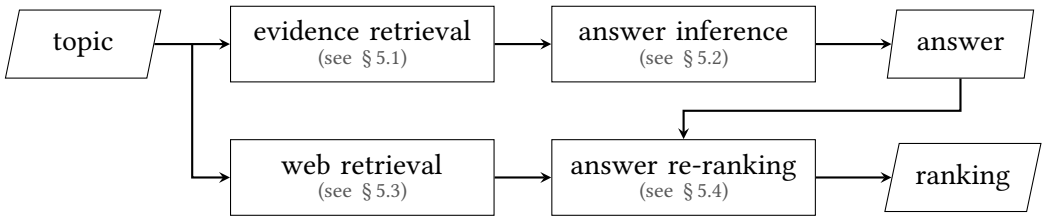


Figure 5.1: Overview of our pipeline for answering health-related yes-no questions and retrieving helpful documents that support the answer.

[1] https://elastic.co/elasticsearch

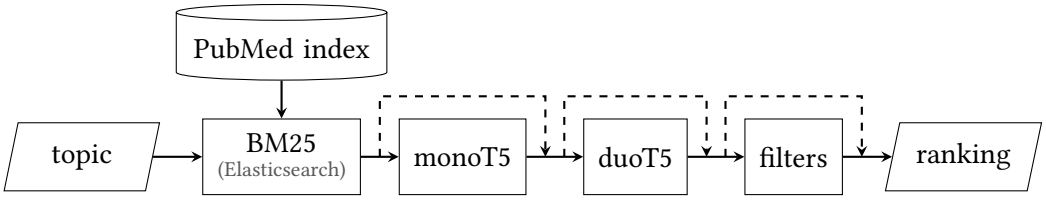


Figure 5.2: Flowchart of our evidence retrieval pipeline. Dashed arrows indicate that a step can be skipped.

five runs submitted for answer inference and ten runs submitted for web retrieval (Section 5.5.2). The runs from our TREC submissions are complemented with four runs based on a multi-stage grid search optimization to find the best parameters for evidence retrieval, answer inference, web retrieval, and answer re-ranking, based on topics from the TREC 2021 Health Misinformation track (Section 5.5.3). To facilitate tuning and evaluation of the evidence retrieval stage, we gather manual relevance and answer judgments for medical abstracts retrieved from PubMed (Section 5.5.1). As an outlook to recent advances in large conversational language models, we also build runs using two chat models for answering health-related yes-no questions (Section 5.5.4). Our runs for each stage are then evaluated with the PyTerrier and `ir_measures` libraries [MMO22; MTM+21] on the topics from the TREC 2022 Health Misinformation track. We discuss important aspects learned from the shared task results and show how grid search optimization improves our health-related question answering pipeline (Section 5.5.5). The chapter is concluded with a summary of our contributions, which also addresses the research questions that were posed in Chapter 1 (Section 5.6).

5.1 Evidence Retrieval

Recall our example from Figure 1.1 in Chapter 1: Popular search engines like Bing often incorrectly answer health-related questions because their answer can be based on any web document that was found to be relevant to the question, even including unverified sources. This approach clearly fails to prevent misinformation. We argue that answers to health-related questions must instead be grounded on solid evidence, such as found in peer-reviewed biomedical publications.

The PubMed[2] is a large government-maintained repository of biomedical literature and contains citations and abstracts of 35 million scientific articles from the fields of life sciences, behavioral sciences, chemical sciences, and bioengineering. Because the PubMed contains evidence of all grades, from relatively recent case

[2] `https://pubmed.gov`

studies, e.g., about garlic-induced face burn,[3] to systematic reviews, e.g., about the effectiveness of COVID-19 vaccines,[4] we can use it to retrieve evidence for answering both current and long-standing health-related questions. Scientific abstracts also often contain a conclusion to the articles research question. Intuitively, we can therefore use PubMed abstracts as context to infer the correct answer to a health-related question. We propose a retrieval pipeline, shown in Figure 5.2, which first retrieves abstracts from an index of PubMed abstracts and then uses pre-trained text-to-text models to re-rank the initial candidates, as proposed by Pradeep, Nogueira, and Lin [PNL21]. Filtering can be applied after re-ranking to remove articles without an abstract text or without a title.

The U.S. National Library of Medicine each year releases a complete snapshot of all articles in PubMed. We use Elasticsearch to index the abstracts and titles of the 34 million scientific articles from the snapshot of 2022,[5] resulting in an index of 73 GB. From this index, we retrieve up to 1,000 abstracts for each topic with BM25 [RWJ+94] on the title and abstract text, as implemented in Elasticsearch. Either the topic's natural language question (`description` field) or the provided keyword query (`title` field) can be used with this approach. We then re-rank up to 1,000 of the top results from BM25 using monoT5 [PNL21], a pointwise neural re-ranker based on a fine-tuned text-to-text model. Different monoT5 models are utilized in our experiments, which were all tuned on subsets of the MS MARCO dataset [NRS+16]. Up to 50 of the top-ranked results from the monoT5 ranking are then re-ranked using duoT5 [PNL21], a pairwise neural re-ranker similarly based on a text-to-text prompting. Again, we try different duoT5 models that were pre-trained on MS MARCO. The last steps of our evidence retrieval stage consist of optionally filtering out abstracts with an empty body or title, as they are not useful for answering the question.

5.2 Answer Inference

After retrieving relevant evidence for a topic, we use pre-trained claim verification models and question answering models to infer a yes/no answer to the topic's question as shown in Figure 5.3. Both claim verification models and question answering models expect an input of a text context on which the answer is based. We use the retrieved articles' abstracts and titles as context for the models. Claim verification models then predict whether a claim can be supported or refuted given the context [TVC+18]. Question answering models instead use the context to directly answer the given question [KMK+20]. Multiple models can be used in

[3] PubMed ID: 33691515
[4] PubMed ID: 35202601
[5] `https://pubmed.gov/download#annual-baseline`

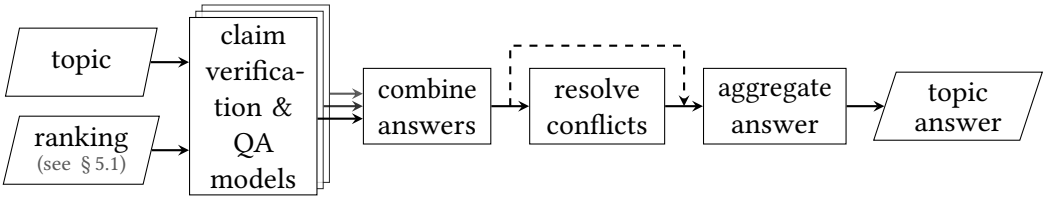


Figure 5.3: Flowchart of our answer inference pipeline. Dashed arrows indicate that a step can be skipped.

our pipeline. The article-level answers of different models are averaged to form one answer for each article. We then optionally apply axiomatic re-ranking using the publication dates to resolve conflicts where the predicted answer based on one article contradicts the answer based on another article [HVG+16]. Finally, the answers from all articles that are retrieved for a topic are aggregated to form a single answer for the topic's question.

Two claim verification models and two question answering models are considered in this thesis: (1) MultiVerS, a claim verification model with checkpoints pre-trained on different (scientific) claim verification datasets [WLW+22], (2) Vera, which was proposed to predict document-level effectiveness of treatments in the TREC 2021 Health Misinformation track [PMN+21], (3) UnifiedQA, a question answering model pre-trained on various question answering datasets [KMK+20], and (4) a RoBERTa model fine-tuned on the BoolQ dataset to predict binary yes/no answers to closed-ended questions. Different models and checkpoints are used for our TREC submissions (Section 5.5.2) and our grid search experiments (Section 5.5.3), and we experiment with using either the topic's `title` or `description` field as the model's claim or question. Answers are predicted for up to 1,000 abstracts, depending on the configuration. We normalize all models' outputs to a 0–1 normalized answer score where values near 1 denote "yes", near 0 denotes "no", and values in between denote uncertainty. Even though the claim verification models were originally trained to predict the supporting/refuting probabilities (given a claim and a text passage), we interpret their predictions as yes/no answer prediction scores. This follows our intuition that evidence which supports a claim about some treatment is analogous to giving a yes answer to a question about the effectiveness treatment based on the same evidence.

The normalized answers for each topic-article pair are combined by averaging the predicted answer scores from all used models, leaving one answer score for each topic-article pair. Sometimes, the answers inferred from different articles of the evidence ranking may contradict each other, e.g., if the answer of the first-ranked article is "no", but the answer of the second article is "yes". In medical sciences, it is not uncommon that results of older studies are disproved by newer studies. For example, it had been generally accepted that vaccines could cause

autism in children until studies in the 2000s disproved this myth [GO09]. Intuitively, we should therefore favor newer over older evidence when inferring answers to health-related questions. In our pipeline, we reflect this preference by re-ranking the predicted article-level answers using information retrieval axioms [HVG+16]. Axioms are constraints that induce a pairwise preference between two items, given that the axiom's precondition is met. We use the `ir_axioms` library [BFR+22] to formulate the following axiom: Given two articles with contradicting answers to the topic, the more recently published article's answer should be preferred. Article publication dates are retrieved using the PubMed API.[6] The axiom is applied to all pairs of articles retrieved for a topic, and the conflicting articles are re-ranked with the KwikSort algorithm [ACN08].

We propose five different answer aggregation strategies. The first four strategies aggregate the answer scores of all considered articles disregarding their ranking position: (1) Averaging, (2) strict aggregation that returns the most pessimistic (closest to 0) answer score, (3) relaxed aggregation that returns the most optimistic (closest to 1) answer score, and (4) majority voting by counting the number of "yes" and "no" answers (answer score threshold: 0.5). The drawback of the aforementioned aggregations is that the topical relevance of the retrieved evidence (estimated by the retrieval score) is not considered. But we assume that answers inferred from more relevant (and more recent, in case of conflict re-ranking) articles might be closer to the true answer, and thus, should receive a higher aggregation weight. In our fifth aggregation strategy, we therefore discount the aggregation weights of articles with lower ranking positions, similar to how cumulative gain is discounted in the nDCG relevance measure [JK02].

Discounted answer aggregation of the topic answer score consists of three steps: First, the predicted answer score score_i from the article at rank i is discounted by the logarithm of its rank, and we compute the discounted cumulative answer score DCA for the top-k articles:

$$\text{DCA}_k = \sum_{i=1}^{k} \frac{\text{score}_i}{\log_2 i + 1}$$

Second, the normalization factor for a ranking of k articles is computed as the maximum achievable (ideal) discounted cumulative answer score IDCA (i.e., if all article answers were "yes"):

$$\text{IDCA}_k = \sum_{i=1}^{k} \frac{1}{\log_2 i + 1}$$

[6] `https://ncbi.nlm.nih.gov/books/NBK25499/#chapter4.ESummary`

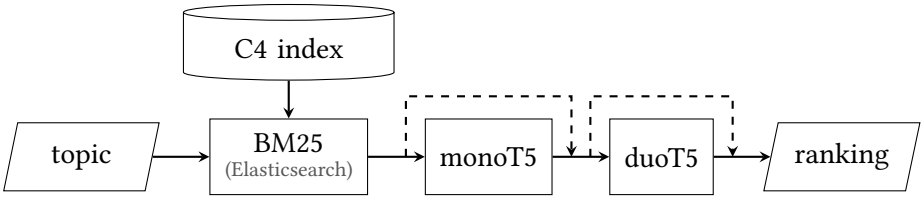


Figure 5.4: Flowchart of our web retrieval pipeline. Dashed arrows indicate that a step can be skipped.

Finally, we use the normalized discounted cumulative answer score nDCA_k as the predicted answer score to the topic:

$$\text{nDCA}_k = \frac{\text{DCA}_k}{\text{IDCA}_k}$$

This approach imitates human search behavior like nDCG does: When skimming through evidence to answer a question, people tend to trust the top-ranked articles more than the lower-ranked ones [AS19]. We experiment with different cutoff points for each of the five aggregation strategies, using up to 1,000 articles for each topic to infer the answer.

5.3 Web Retrieval

Our web retrieval stage follows a standard ranking architecture composed of lexical candidate retrieval and neural re-ranking as proposed by Pradeep, Nogueira, and Lin [PNL21]. Figure 5.4 shows a flowchart of the components of the web retrieval pipeline. The C4 corpus is used as the web collection for our web retrieval experiments [DSM+21; RSR+20]. Like with evidence retrieval (Section 5.1), we index the 1 billion documents in Elasticsearch, resulting in a distributed index size of 18 TB. Up to 1,000 documents are retrieved for each topic with Elasticsearch's BM25 scoring [RWJ+94] using the topics' question or query. Up to 1,000 of the top results from BM25 are then re-ranked with monoT5, and up to 50 of the top-ranked results from the monoT5 ranking are again re-ranked with duoT5 [PNL21]. We experiment with varying re-ranking depths and different monoT5/duoT5 models that were pre-trained on parts of the MS MARCO dataset [NRS+16]. Topic answers are not yet considered in the web retrieval stage.

5.4 Answer-Based Re-Ranking

The documents retrieved by our web retrieval pipeline are so far only ranked based on their topical relevance. But for health-related questions, it is important to also

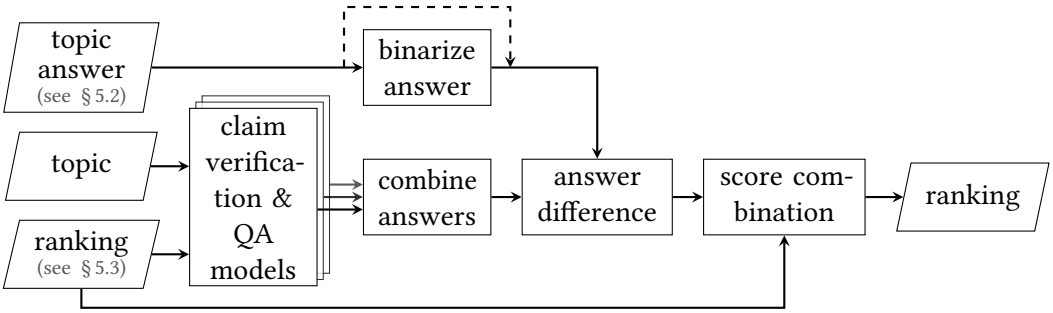


Figure 5.5: Flowchart of our answer re-ranking pipeline. Dashed arrows indicate that a step can be skipped.

consider the correctness of the information contained in the retrieved documents. Misinformation should be penalized and never appear on high ranks. To this end, we propose to first predict the answer to the topic's question for each retrieved document, and then compare the document answer to the "true" answer that was predicted for the topic (Section 5.2). We then combine the document retrieval scores with the answer prediction scores using different strategies to obtain a final answer-based ranking with less misinformation at the topmost ranking positions. The steps of this answer-based re-ranking pipeline are shown in Figure 5.5.

To predict answers from web documents, we apply the same set of claim verification and question answering models that were used to infer the topic answer (Section 5.2) and similarly combine the document-level answers if more than one model is used. We interpret the topic answer inferred from PubMed evidence as the "true" answer to compare the document answers to. To "sharpen" the topic answer, we optionally binarize the topic answer at a configurable threshold, such that the topic answer is set to 1 ("yes") if the predicted topic answer score is greater than the threshold and to 0 ("no") otherwise.

For re-ranking, we then compute the difference between the predicted (optionally binarized) answer score of the topic T (used as the "true" answer) and the answer score each ranked document D:

$$\Delta\,\text{answer} = \big|\,\text{answer}(T) - \text{answer}(D)\big|$$

This answer difference can serve as a proxy to the incorrectness of a web document. That is, if the topic answer was "yes", then Δ answer is smaller for documents where the predicted answer is also "yes". The closeness $1 - \Delta$ answer to the predicted "true" topic answer can similarly serve as a proxy for document correctness. We propose four answer-based re-ranking strategies that combine a document's retrieval score (estimating the topical relevance) with its correctness: (1) Linear score boosting multiplies the retrieval score with the closeness to the predicted topic answer:

$$\text{score}_{\text{lin}}(D) = \text{score}_{\text{BM25+T5}}(D) \cdot (1 - \Delta\,\text{answer})$$

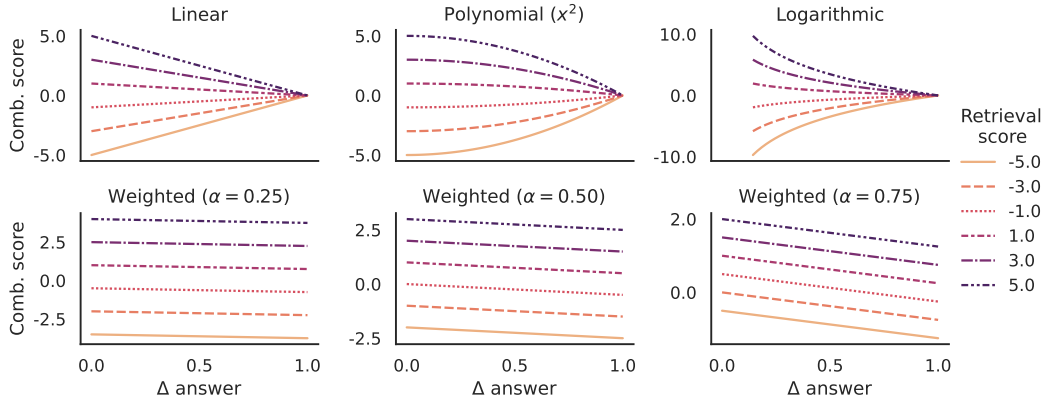


Figure 5.6: Influence of answer difference on the original retrieval score with different score combination strategies.

(2) Polynomial score boosting multiplies the original retrieval score with the softened closeness to the predicted topic answer, by squaring the answer difference. For documents answers relatively close to the topic answer, the impact on the retrieval score is not as strong as for documents with a large answer difference:

$$\text{score}_{\text{pol}}(D) = \text{score}_{\text{BM25+T5}}(D) \cdot (1 - \Delta\,\text{answer}^2)$$

(3) Logarithmic score boosting multiplies the original retrieval score with the negative logarithm of the difference to the predicted topic answer:

$$\text{score}_{\text{log}}(D) = \text{score}_{\text{BM25+T5}}(D) \cdot -\log(\Delta\,\text{answer})$$

And (4) weighted score combination combines the original retrieval score with the closeness to the predicted topic answer using a configurable weight α:

$$\text{score}_{\text{com}}(D) = \alpha \cdot \text{score}_{\text{BM25+T5}}(D) + (1 - \alpha) \cdot (1 - \Delta\,\text{answer})$$

In Figure 5.6, we compare the six answer combination strategies. The score boosting strategies (linear, polynomial, and logarithmic) in the upper row of Figure 5.6 multiply the original score. Consequentially, the combined scores approach 0 for higher answer differences. The polynomial and logarithmic variants can be seen as "softer" and "stricter" variants of the linear score boosting strategy, respectively. For the weighted score combinations in the lower row of Figure 5.6 scores are shifted down for higher answer differences. The trade-off factor α can be used to adjust the impact of the answer difference on the final score.

5.5 Evaluation

We facilitate two settings to evaluate our approaches: (1) By participating in the TREC 2022 Health Misinformation Track [CMS+22] with five answer prediction and ten retrieval runs based on our framework (Section 5.5.2), and (2) by performing a grid search parameter optimization using relevance and answer judgments from TREC 2021 and manually annotated judgments for PubMed articles (Section 5.5.3).

Because the four stages of our approach are different in their goals and outputs, we evaluate them separately using specialized measures for each stage. For relevance judgments, we adopt the notion of graded relevance proposed by Clarke et al. [CRS+20] for the TREC Health Misinformation tracks from 2020 to 2022. Graded relevance labels do not only take into account a document's relevance but also its correctness and credibility. Clarke et al. [CRS+20] claim that a relevant but incorrect document is even worse than an irrelevant document and assign the graded labels accordingly. They refrain from using normalized discounted cumulative gain (nDCG) [JK02] on graded judgments as the primary measure for misinformation-aware rankings because it does not penalize incorrect information [CRS+20]. But because incorrect yet relevant information is considered to be the most harmful, we must consider the incorrectness of the retrieved documents in evaluation. Clarke, Vtyurina, and Smucker [CSV20; CVS20] propose a new measure, the compatibility to an ideal ranking, to evaluate the effectiveness of rankings where relevance levels are hard to define, e.g., when considering incorrect documents. At TREC Health Misinformation, the compatibility is separately measured for helpful documents (i.e., relevant and correct documents) and harmful documents (i.e., relevant but incorrect), and the effectiveness of misinformation prevention is measured as the difference between the helpful and harmful compatibility. We use the compatibility measures as the primary measure to evaluate our three stages which return rankings: (1) Evidence retrieval, (2) web retrieval, and (3) answer-based re-ranking. For the answer prediction stage, we use a set of common classification measures with the primary measure being the area under the receiver operating characteristic curve (AUC) as proposed at TREC 2022 [CMS+22].

The TREC Health Misinformation tracks also provide an extensive set of relevance judgments for documents from the C4 corpus and provide verified expert-annotated true answers for each of the topics used in the tracks [CMS+22; CMS21]. The judgments were obtained by NIST assessors[7] labeling the relevance, correctness, and credibility of the documents from submitted TREC runs. Because of the high quality of TREC judgments, we use these judgments and true answers for our evaluation. Evidence retrieval effectiveness was, however, not measured nor annotated in the TREC shared tasks. We therefore create our own relevance judgments for evidence

[7]The U.S. National Institute of Standards and Technology (NIST) hosts the TREC conference series and usually recruits retired employees for relevance assessments.

retrieval by manually annotating the relevance and answer of 2,096 documents retrieved for 151 topics (Section 5.5.1).

5.5.1 Manual Judgments for Evidence Retrieval

To assess the effectiveness of evidence retrieval approaches from the PubMed corpus, we create graded relevance judgments based on the manual assessment of topical relevance and represented answer of 2,096 medical abstracts pooled for 151 topics from the TREC 2019, 2021, and 2022 Health Misinformation tracks [ASL+19; CMS+22; CMS21] as well as for the 15 consumer health questions by Bondarenko et al. [BSD+21]. Top-5 pooling is used on six retrieval systems that retrieve abstracts and titles from an Elasticsearch index of 34 M articles from PubMed: (1) The top-1000 retrieved by Elasticsearch's BM25 [RWJ+94] using the topic's query (`title` field), (2) the 1,000 documents from the BM25 ranking after first re-ranking the top-100 with monoT5 (`monot5-base-msmarco`) and then re-ranking the top-5 again with duoT5 (`duot5-base-msmarco`) both pre-trained on MS MARCO [NRS+16; PNL21], (3) the monoT5-duoT5 re-ranked abstracts additionally filtered for non-empty title and abstract text; and (4–6) the same three rankings applied using the topic's natural language question (`description` field) instead of the query. In total, 2,096 documents are included in the pool.

To build relevance judgments in the same way as the TREC Health Misinformation tracks [CRS+20], we need to label each document's relevance, answer, and credibility. Four volunteers with a computer science or media science background (1 Master's student, 1 Master's graduate, and 2 PhD students) were asked to annotate the pooled abstracts. While none of the annotators had professional medical training, they were familiar with medical texts and had previously worked on medical information retrieval annotation tasks. We derive annotator instructions similar to the guidelines used in the TREC 2022 Health Misinformation track [CMS+22]. To reduce annotation time for the volunteer annotators, we simplify the original instructions from TREC to only consider one relevance level (i.e., relevant or not). Instructions for answer assessment are used without modification. Credibility assessments are not made under the assumption that almost all abstracts on PubMed are published in peer-reviewed journals and hence, are automatically labeled as credible. The full (revised) annotation instructions are available in Appendix A.3.

We conduct a pilot study where each of the three annotators is asked to annotate the same 66 abstracts from five topics, to assess the annotation time and inter-annotator agreement. Annotating the 66 abstracts from the pilot study was measured to take 40 minutes. We therefore estimate that judging the remaining pooled documents takes 21 hours or five hours per annotator. We measure inter-annotator agreement using Fleiss' κ [Fle71] for the two labels, relevance and answer, separately. Moderate to substantial agreement is achieved with $\kappa = 0.64$ on the

relevance labels and $\kappa = 0.55$ on the answer labels [LK77]. Both the high annotation time[8] and the moderate inter-annotator agreement indicate that the annotation task is challenging and that the instructions are not clear enough. After a chat discussion with the annotators, we therefore extended the annotation instructions with examples and clarifications. Nonetheless, the annotator agreement is considered high enough to be able to use the judgments for evaluation.

The remaining abstracts are distributed to the four annotators and annotated by each annotator individually. Due to the high annotation time we refrain from conducting a second inter-annotator agreement study. Instead, we compare the label distribution of the complete pool with the label distribution from the Monant Medical Misinformation dataset [SPT+22], a dataset of 317,000 medical news articles with associated fact-checked claims. We expect both datasets to follow similar label distributions, as both datasets are based on medical texts and both datasets are annotated for effectiveness (i.e., "yes" answer) and ineffectiveness (i.e., "no" answer). Our labels of scientific abstracts yield a less severe confirmation bias than the labels by Srba et al. [SPT+22] on scientific news articles. Relevant abstracts from PubMed were confirming the question in 44 % of all cases compared to of 66 % confirmed claims in the Monant dataset. The number of inconclusive abstracts or articles is higher in our judgments (46 %, Monant dataset: 6 %) and fewer abstracts from PubMed contradict the question (10 %, Monant dataset: 28 %). Except for a higher percentage of inconclusive abstracts, the label distribution is similar to the Monant dataset and our judgments are therefore considered valid.

During the annotation process, we also observed that some abstracts are not medical, but related to other fields like material sciences.[9] We also found that some abstracts are duplicates, indicating that deduplication of evidence search results might be necessary. An example of duplicated abstracts on PubMed is the article "Oral contraceptives for functional ovarian cysts", which appeared five times[10] for the question "Will taking birth control pills treat an ovarian cyst?" Due to time constraints, we were unable to address this issue, but future work should investigate how retrieving duplicated abstracts or abstracts from other field could bias answer prediction as was previously demonstrated by Fröbe et al. [FBP+20; FBR+20].

Annotations from the pilot study are aggregated by majority vote of the four annotators, using the author's vote as a tie-breaker. We then use the annotations from all 2,096 documents to derive binary and graded judgments in the same way as Clarke et al. [CRS+20]: First, correctness labels are obtained by comparing each article's annotated answer label with the topic's true answer field. For simplicity and due to high amount of peer-reviewed articles on PubMed, we consider all articles

[8] Settles, Craven, and Friedland [SCF08] report an average annotation time of 7.6 seconds for a similar binary annotation task on PubMed abstracts, 5 times less time than our task (36 s).

[9] E.g., one abstract about the fragility of crystals; PubMed ID 23023553.

[10] PubMed IDs: 17054275, 19370628, 19701050, 21901701, and 24782304.

Table 5.1: Graded and binary relevance judgments derived from manual annotations of PubMed abstracts on 4 datasets of health-related yes-no questions. Relevance labels are generated for helpful results only (Help), harmful results only (Harm), relevant results, relevant and correct results (R & Co), and incorrect results (Incor.).

Track/Dataset	**Graded**			**Binary**	
	Help	Harm	Relevant	R & Co	Incor.
TREC 2019 Health Misinfo. [ASL+19]	438	5	652	97	5
TREC 2021 Health Misinfo. [CMS21]	287	54	625	141	54
TREC 2022 Health Misinfo. [CMS+22]	193	27	645	98	27
Health Misbeliefs [BSD+21]	43	7	166	16	7
$\sum$ *Total*	*961*	*93*	*2088*	*352*	*93*

as credible. We then use the mapping from Clarke et al. [CRS+20] to derive graded labels that combine the preference orderings of an article's relevance, correctness, and credibility. Table 5.1 shows how many relevance judgments were created for each annotated dataset. These relevance judgments are used for hyperparameter optimization (Section 5.5.3) and to evaluate the effectiveness of evidence retrieval approaches discussed in Section 5.5.5.

5.5.2 TREC Health Misinformation Track

Shared tasks such as TREC[11] enable the large-scale of advancing approaches in the field of information retrieval. With a focus on preventing misinformation in the retrieval for health-related questions, the TREC 2022 Health Misinformation track represents a good opportunity to evaluate our approach on yet unseen topics [CMS+22]. The 2022 edition of the shared task was the first to feature a second task besides the web retrieval task: Participants were not given the true topic answers in advance but instead had to predict the answer for all 50 topics. We participate in both tasks, answer prediction and web retrieval.

We participate in the TREC Health Misinformation track with an international team of researchers from the Webis group [BFG+22]. From the 20 runs submitted to the track by the Webis team, 15 runs are based on our proposed question answering and retrieval pipeline and the remaining runs were created by other team members. The five answer prediction runs use the first two stages of our pipeline (Sections 5.1–5.2) and the ten web retrieval runs use all four stages described in Sections 5.1–5.4. In the following, we shortly characterize our submissions to the two tasks. The results are discussed in Section 5.5.5.

[11] `https://trec.nist.gov/`

Answer Prediction Task To predict a correct answer to the 50 health-related yes-no questions, we employ our pipeline with different pre-trained question answering and claim verification models. All five answer prediction runs use the abstracts from 1,000 PubMed articles as evidence. After retrieving 1,000 abstracts from our PubMed index on Elasticsearch using the topic's question as the query and BM25 for scoring, we first re-rank all 1,000 results from BM25 with monoT5 [PNL21] pre-trained on medical passages from MS MARCO (`monot5-3b-med-msmarco`) and then again re-rank the top-50 results from monoT5 with duoT5 [PNL21] also pre-trained on MS MARCO (`duot5-3b-med-msmarco`).[12] Our runs use different answer prediction models: One run uses a question answering model, two runs use claim verification models, and two runs use a combination of both. In all runs, the topic answer is based on the predicted answer scores from all retrieved 1,000 question-abstract pairs. In all runs, we aggregate the topic answer score by discounting ranking positions nDCA_k (with $k = 1,000$).

Five runs that were submitted in the team are not discussed in this thesis because they have been developed mainly by other team members: *Webis-goo-boolq-abs*, *Webis-goo-lbert-abs*, *Webis-goo-lbert-title-abs*, *Webis-nlm-boolq-abs*, and *Webis-nlm-lbert-abs*. We shortly list the characteristics of the remaining runs that were developed by the author of this thesis:

Webis-uniqa-dis predicts the answers to the topics' questions with a UnifiedQA model (`unifiedqa-t5-large`) pre-trained on various question answering datasets and uses the article abstracts as context for the model [KMK+20].

Webis-longck-dis predicts the answers to the topics' questions using the MultiVerS[13] claim verification model pre-trained on the FEVER [TVC+18] dataset (`fever_sci` checkpoint) and uses each article's abstract and title as context input for the model [WLW+22].

Webis-verasent-dis predicts the answers to the topics' questions with the Vera model (Vera-3B checkpoint) pre-trained on the questions and answers from the TREC 2019 Decision track [ASL+19] and uses the article abstracts as context for the model [PMN+21]. To circumvent Vera's 512 token input limit, we select only the "most relevant" sentences from the abstracts for inclusion in the prompt by applying a heuristic using term frequencies of selected indicator words that was proposed by Zhang et al. [ZTA+22].

Webis-longck-uniqa-dis uses the average of the scores predicted with both, the UnifiedQA question answering model [KMK+20] and the MultiVerS claim verification model [WLW+22].

[12] Model links provided in Appendix B.

[13] The model was previously called LongChecker, hence the different name in the run tag.

Webis-longck-uniqa-ax-dis also uses the average of the answer scores from the UNIFIEDQA [KMK+20] and MultiVerS [WLW+22] models. But after predicting the article-level answer, the top-1,000 PubMed abstracts are axiomatically re-ranked [HVG+16] based on the abstracts' publication dates (more recently published abstracts ranked higher) to resolve potentially contradicting answers from different articles before aggregating the topic answer.

Web Retrieval Task Prior to retrieval from the C4 corpus, all our web retrieval runs use the same topic answer inference framework as above. Interpreting the predicted topic answer as the "true" answer, we then apply different answer-based re-ranking strategies to the web documents based on the closeness of answers predicted for a web document to the topic answer. For answer prediction on web documents, we use different question answering and claim verification models.

The topic answer is predicted using pre-trained question answering and claim verification models based on evidence from 1,000 abstracts retrieved from PubMed. Abstract retrieval again uses Elasticsearch's BM25 and re-ranking with monoT5 (top-1,000, `monot5-3b-med-msmarco`) and subsequently with duoT5 (top-50 from monoT5, `duot5-3b-med-msmarco`) [PNL21]. Question answering and claim verification models are then used to predict an answer on all 1,000 abstracts. After axiomatic re-ranking [HVG+16] to resolve answer conflicts based on publication dates, the article-level answer scores are aggregated by discounting ranking positions nDCA_k (with $k = 1,000$) to get the topic answer. No binarization is applied.

We then retrieve 1,000 documents from the C4 index [RSR+20], with Elasticsearch's BM25 scoring. All 1,000 retrieved documents are first re-ranked with a monoT5 [PNL21] model that was pre-trained on medical passages from MS MARCO (`monot5-3b-med-msmarco`). The top-50 documents from the monoT5 re-ranking are then again re-ranked with duoT5 [PNL21] also pre-trained on MS MARCO corpus (`duot5-3b-med-msmarco`). The document-level answer is predicted using either the MultiVerS claim verification model [WLW+22], the UNIFIEDQA question answering model [KMK+20], or the average score of both models. The closeness of the document-level answer score to the "true" topic answer is then combined with the retrieval score of each document using three different strategies: (1) Linear score boosting, (2) polynomial score boosting (squared answer difference), or (3) a weighted score combination with $\alpha = 0.75$. In each run, the same models are used for both topic and document answer prediction. The individual runs applying answer-based re-ranking are described in the following:

Webis-longck-ax-lin predicts the answer scores for abstracts and documents the MultiVerS claim verification model pre-trained on the FEVER [TVC+18], PubMedQA [JDL+19], and evidence inference datasets [LDB+19] (`fever_sci` checkpoint) and uses each article's abstract and title or each document's text

as context input, respectively [WLW+22]. Retrieval scores are then boosted linearly, based on the closeness between a re-ranked document's answer score and the predicted "true" topic answer.

Webis-uniqa-ax-lin predicts the answer scores for abstracts and documents with a UnifiedQA question answering model (`unifiedqa-t5-large`) pre-trained on various question answering datasets [KMK+20], using the abstract text or document text as context for the model. Linear score boosting is applied as the answer-based re-ranking strategy.

Webis-longck-uniqa-ax-lin uses the average of the answer scores predicted with both, the MultiVerS claim verification model [WLW+22] and the UnifiedQA question answering model [KMK+20]. Linear score boosting is applied analogous to the previous two runs.

Webis-longck-ax-pol uses the same MultiVerS claim verification model [WLW+22] to predict topic and document answers, but applies polynomial score boosting for answer-based re-ranking (answer difference is squared).

Webis-uniqa-ax-pol uses the UnifiedQA question answering model [KMK+20] to predict topic and document answers, but applies polynomial score boosting for answer-based re-ranking.

Webis-longck-uniqa-ax-pol uses the average of the answer scores predicted with both, the MultiVerS claim verification model [WLW+22] and the UnifiedQA question answering model [KMK+20]. Polynomial score boosting is applied analogous to the previous two runs.

Webis-longck-uniqa-pol predicts the topic and document answer analogous to the previous run, but does not apply axiomatic re-ranking to resolve answer conflicts before topic answer prediction.

Webis-longck-ax-com uses the same MultiVerS claim verification model [WLW+22] to predict topic and document answers, but applies a weighted combination ($\alpha = 0.75$) of the re-ranked document's retrieval score with the closeness of the document's answer score to the predicted "true" topic answer for answer-based re-ranking.

Webis-uniqa-ax-com uses the UnifiedQA question answering model [KMK+20] to predict topic and document answers, but applies a weighted combination ($\alpha = 0.75$) of retrieval score and answer closeness for answer-based re-ranking.

Webis-longck-uniqa-ax-com uses the average of the answer scores predicted with both, the MultiVerS claim verification model [WLW+22] and the UnifiedQA question answering model [KMK+20]. Weighted score combination ($\alpha = 0.75$) is applied analogous to the previous two runs.

Table 5.2: Hyperparameters for grid search optimization of our evidence retrieval stage (Section 5.1). Best parameters are indicated in bold.

Parameter	Values
Retrieval	top-1000 Elasticsearch BM25 on PubMed index
Query field	**query** (`title` field), question (`description` field)
monoT5 cutoff	0, 10, **100**, 1000
monoT5 model	`monot5-base-msmarco`, `monot5-3b-msmarco`, **`monot5-3b-med-msmarco`**
duoT5 cutoff	**0**, 5, 50
duoT5 model	`duot5-base-msmarco`, `duot5-3b-msmarco`, `duot5-3b-med-msmarco`

5.5.3 Optimizing Evidence Retrieval, Answer Prediction, and Web Retrieval

To evaluate the best achievable answer prediction and web retrieval effectiveness with our four-stage approach, we conduct a systematic grid search hyperparameter optimization using the Weights & Biases platform [Bie20]. Each stage is optimized separately to keep the number of possible hyperparameter combinations at a feasible level. To measure realistic effectiveness, we use the topics and relevance judgments from the TREC 2021 Health Misinformation track to optimize our parameters (i.e., as the validation set) and finally, evaluate the effectiveness of the best hyperparameter configurations on unseen topics from 2022. Our grid search optimization is performed in four steps, one for each stage of our pipeline as shown in Figure 5.1: (1) For the evidence retrieval stage, (2) for the answer prediction stage (using the best parameters found for the evidence retrieval stage), (3) for the web retrieval stage, and (4) for the answer-based re-ranking stage (using the best parameters of the three other stages).

First, we optimize evidence retrieval. As suggested by Clarke et al. [CRS+20], the difference between a ranking's compatibility [CSV20; CVS20] to helpful and harmful results is used as our target measure for optimization. The hyperparameter choices that are considered for the grid search are listed in Table 5.2. All configurations use the top-1,000 articles retrieved from the PubMed index with Elasticsearch's BM25 scoring as the first stage retrieval. We use either the topic's query or question for retrieval and include different monoT5 and duoT5 models [PNL21] that were pre-trained on either the full MS MARCO dataset or just medical passages from MS MARCO. We also consider different cutoff-points for the two re-rankers (monoT5 and duoT5). In total, 216 configurations are tested on a shared cluster of 24 Nvidia A10 GPUs. The best configuration yields a maximum difference between helpful and harmful compatibility of 0.11 on TREC 2021 topics. It uses the topic's query for retrieval and then re-ranks the top-100 documents from the BM25 ranking

Table 5.3: Hyperparameters for grid search optimization of our answer inference stage (Section 5.2). Best parameters are indicated in bold.

Parameter	Values
Evidence retrieval	best configuration from Table 5.2
Evidence cutoff	1, 3, 5, 10, **100**, 1000
Claim/question field	query (`title` field), **question** (`description` field)
Answer model	MultiVerS (`fever`, `fever_sci`, `healthver`, `scifact`), **Vera-3B** (relevant sentence selection, **truncation**), `roberta-large-boolq-finetuned`, `unifiedqa-t5-large`
Conflict re-ranking	publication date (most recent first), **none**
Answer aggregation	mean, strict, relaxed, majority vote, **nDCA**

with monoT5 pre-trained on medical passages from MS MARCO. The pairwise duoT5 re-ranking does to not increase the difference between helpful and harmful results and is not used in the best configuration (duoT5 re-ranking cutoff: 0).

Second, we optimize our answer inference stage and use the articles retrieved by the best evidence retrieval run as context for different claim verification and question answering models as shown in Table 5.3. We also vary the amount of evidence used for answer prediction (cutoff from 1 to 1,000). The Vera model [PMN+21] is tested in two variants, either with selecting the "most relevant" sentences to fill the model's 512-token input window (heuristic by Zhang et al. [ZTA+22]) or by truncation. Axiomatic re-ranking based on the publication date is either enabled or disabled, and we try all five aggregation strategies described in Section 5.2. From the 960 configurations, the best yielded an area under the receiver operating characteristic curve of 0.89. The configuration uses evidence from the top-100 retrieved articles and then predicts the answer with the Vera-3B (truncation to 512 tokens) using the topics yes-no question as the model input. Axiomatic re-ranking is not used in the most effective configuration, and answers are aggregated by nDCA_{100} (with $k = 100$, discounting the top-100 ranking positions).

For optimizing our web retrieval stage, we use the same hyperparameter ranges (shown in Table 5.4) as for evidence retrieval, except for retrieving from the C4 [DSM+21] index instead of the PubMed index on Elasticsearch. For re-ranking, we again use different pre-trained monoT5 and duoT5 models [PNL21]. We ran 216 configurations on our cluster. Our best configuration as evaluated on TREC 2021 topics uses the topic's yes-no question as the query for retrieval and then re-ranks all 1,000 retrieved documents from the BM25 with monoT5 pre-trained on medical passages from MS MARCO. As with evidence retrieval, duoT5 re-ranking is not used in the best configuration (re-ranking cutoff: 0). The best achieved difference between helpful and harmful compatibility on the validation topics is 0.04.

Finally, the hyperparameter optimization of our answer-based re-ranking stage

Table 5.4: Hyperparameters for grid search optimization of our web retrieval stage (Section 5.3). Best parameters are indicated in bold.

Parameter	Values
Retrieval	top-1000 Elasticsearch BM25 on C4 index
Query field	query (`title` field), **question** (`description` field)
monoT5 cutoff	0, 10, 100, **1000**
monoT5 model	`monot5-base-msmarco`, `monot5-3b-msmarco`, **`monot5-3b-med-msmarco`**
duoT5 cutoff	**0**, 5, 50
duoT5 model	`duot5-base-msmarco`, `duot5-3b-msmarco`, `duot5-3b-med-msmarco`

Table 5.5: Hyperparameters for grid search optimization of our answer-based re-ranking stage (Section 5.4). Best parameters are indicated in bold.

Parameter	Values
Topic answer	best configuration from Table 5.3
Binarization threshold	0.40, 0.45, 0.50, **0.55**, 0.60
Web retrieval	best configuration from Table 5.4
Claim/question field	query (`title` field), **question** (`description` field)
Answer model	MultiVerS (`fever`, `fever_sci`, `healthver`, `scifact`), **Vera-3B** (**relevant sentence selection**, truncation), `roberta-large-boolq-finetuned`, `unifiedqa-t5-large`
Score combination	linear score boosting, polynomial score boosting (x^2), logarithmic score boosting, **weighted score combination** ($\boldsymbol{\alpha = 0.25}$, $\alpha = 0.50$, $\alpha = 0.75$)
Re-ranking cutoff	0, 10, 100, **1000**

builds on the best configurations of the three aforementioned stages. Hyperparameter ranges are shown in Table 5.5. Because the aggregated topic answer scores are often near 0.5 (inconclusive), we binarize the topic answer scores with five different thresholds. We then use different claim verification and question answering models to predict a document-level answer for the topic's query or question. All six score combination strategies for combining the closeness of the document answers to the predicted "true" topic answer with the original retrieval score from the web retrieval stage are tested. Answer-based re-ranking is applied with varying cutoff points. The best configuration on TREC 2021 topics uses the same Vera-3B claim verification model [PMN+21] as for evidence retrieval and the topic's question to predict an answer for each retrieved document (with relevant sentences selection to avoid truncation [ZTA+22]). The weighted combination ($\alpha = 0.25$) of the document answer's difference to the binarized topic answer (threshold: 55) is then used to

re-rank all 1,000 documents. Answer-based re-ranking improves the web retrieval stage's compatibility difference evaluated on TREC 2021 topics from 0.04 to 0.06.

The grid-search optimized configurations for our question answering and retrieval pipeline are then used to create four additional runs complementing our runs from the TREC Health Misinformation track, one for each stage of our retrieval pipeline. All runs are evaluated in Section 5.5.5.

5.5.4 Outlook: Answering Health-Related Questions with Large Conversational Models

In the last few months, large language models have attained remarkable effectiveness on a variety of natural language processing tasks, including medical question answering [SAT+22]. Large conversational models like ChatGPT[14] and You Chat[15] make it easy to ask questions like "Are squats bad for knees?", hence, they are a promising candidate for answering health-related yes-no questions. As an outlook to future work in the health domain, we therefore briefly discuss the potential of these models for answering health-related yes/no questions. We use the topic's question (`description` field) to build the following prompt: `[question] Answer in one word: "yes" or "no".` For example, the prompt for the example above would become: `Are squats bad for knees? Answer in one word: "yes" or "no".` We then use both ChatGPT and You Chat with the same prompt to generate the answer. Due to usage quotas, we were only able to obtain answers for a random sample of 15 yes/no questions from the TREC 2022 Health Misinformation track on both models. The results are discussed in Section 5.5.5.

5.5.5 Results

To assess the effectiveness of our approaches for evidence retrieval (Section 5.1), answer inference (Section 5.2), web retrieval (Section 5.3), and answer-based re-ranking (Section 5.4), we apply our pipeline to the 50 topics of the TREC 2022 Health Misinformation track [CMS+22]. We compare our runs submitted to TREC and our grid search optimized runs with the current state of the art.

Evidence Retrieval For evidence retrieval, in Table 5.6, we report the compatibility to helpful (i.e., correct) and harmful (i.e., incorrect) results and the normalized discounted cumulative gain (nDCG) with respect to graded relevance judgments [CSV20; CVS20; JK02]. We use our own relevance judgments for the evidence retrieval stage, as described in Section 5.5.1. Because all our TREC runs

[14] `https://chat.openai.com/`
[15] `https://you.com/chat`

Table 5.6: Retrieval effectiveness on TREC 2022 Health Misinformation topics, reported as compatibility to an ideal ranking and nDCG using graded relevance labels. Compatibility is measured for helpful results (Help), harmful results (Harm), or the difference thereof (Δ). Results sorted by the compatibility difference or by nDCG in case of ties. Superscripts mark significant changes to other runs in the same group (Student's t-test, $p < 0.05$, Bonferroni correction). Third-party runs are greyed out and excluded from significance tests because no run files were available. Best results per group are highlighted in bold.

Run	Compatibility			nDCG
	Help	Harm	Δ	Graded
Evidence retrieval (see Section 5.1)				
(a) Grid search	**0.63**^{b}	0.53^{b}	**0.10**	**0.63**^{b}
(b) Webis-…-dis	0.15^{a}	**0.13**^{a}	0.03	0.31^{a}
Web retrieval (see Section 5.3)				
(c) Grid search	**0.27**^{d}	0.16^{d}	**0.10**	**0.66**^{d}
(d) Webis-…-dis	0.17^{c}	**0.08**^{c}	0.09	0.48^{c}
(e) h2oloo-bm25 (best baseline at TREC 2022 [CMS+22])	0.20	0.15	0.05	—
Answer-based re-ranking (see Section 5.4)				
(f) h2oloo-hm22-ref-comb.vera-mt5 (best run at TREC 2022 [CMS+22])	**0.35**	0.09	**0.26**	—
(g) Grid search	0.26^{p}	0.08	0.19	0.63^{knp}
(h) Webis-longck-ax-com	0.27^{p}	0.15	0.12	**0.66**^{ijknpq}
(i) Webis-longck-uniqa-pol	0.17	0.08	0.09	0.57^{h}
(j) Webis-longck-uniqa-ax-pol	0.17	0.08	0.09	0.57^{h}
(k) Webis-longck-uniqa-ax-lin	0.14	0.07	0.08	0.52^{ghlm}
(l) Webis-uniqa-ax-com	0.26^{p}	0.17	0.07	0.66^{knpq}
(m) Webis-longck-uniqa-ax-com	0.25^{p}	0.17	0.06	0.65^{knpq}
(n) Webis-longck-ax-pol	0.15	0.09	0.05	0.54^{ghlm}
(o) Webis-uniqa-ax-pol	0.18	0.14	0.05	0.58^{p}
(p) Webis-longck-ax-lin	0.11^{ghlm}	**0.07**	0.04	0.49^{ghlmo}
(q) Webis-uniqa-ax-lin	0.15	0.12	0.03	0.56^{hlm}

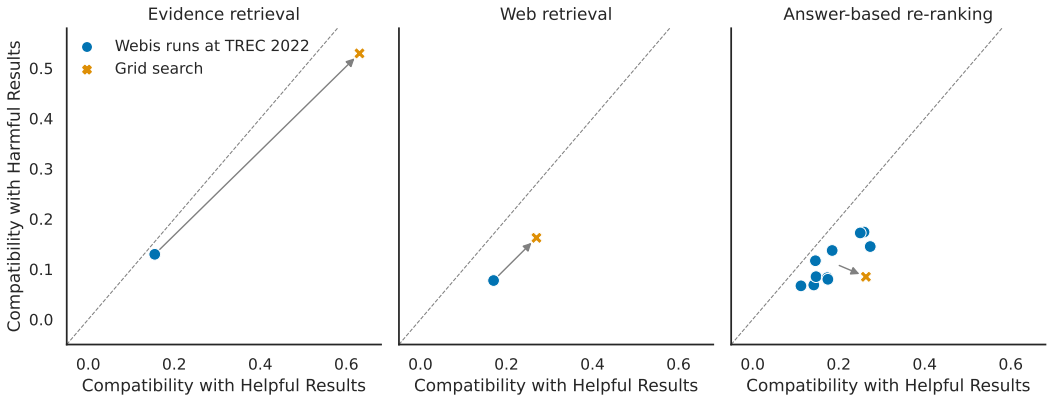


Figure 5.7: Compatibility with helpful and harmful results on TREC 2022 Health Misinformation topics comparing our TREC runs with the grid search optimized runs. Good runs are helpful but not harmful (lower right corner). The reference lines indicate equal helpful and harmful compatibility.

used the same configuration for evidence retrieval (see Section 5.5.2), we only report their results once in Table 5.6. Our grid search optimized approach (run (a) in Table 5.6) has significantly increased compatibility with helpful and harmful results compared to the approach that we submitted to TREC (run (b)). This improvement in helpful compatibility and worsening of harmful compatibility is also indicated in the leftmost plot of Figure 5.7. The compatibility difference, indicated as the distance to the reference line in the plot, only slightly increases from 0.03 to 0.10, though not significantly. Because evidence retrieval was not considered a sub-task of the TREC 2022 Health Misinformation track, we cannot compare our runs to third-party approaches. We notice that the evidence retrieval effectiveness also largely varies between different queries, resulting in a standard deviation of 0.40 for the compatibility difference. For 9 of the 50 evaluated topics, we measure an nDCG of 0 on graded relevance labels. This indicates that for some queries, (almost) no evidence can be found on PubMed. Manual inspection of the affected queries revealed that the worst-performing queries are often about alternative remedies such as for the question "Can you use WD-40 for arthritis?" Integrating other sources of evidence might therefore be beneficial.

Answer Inference Answer inference effectiveness is measured as the area under the receiver operating characteristic curve (AUC) among other classification measures and is shown in Table 5.7. We also include the results reported for the best-performing automatic run (h2oloo-gpt3b) at TREC 2022 [CMS+22] and the best reported effectiveness from Pugachev et al. [PAB+23] (using 113 topics distinct from our test set) for reference. None of our evidence-based answer inference

Table 5.7: Answer inference effectiveness on TREC 2022 Health Misinformation topics, reported as area under the ROC curve (AUC), accuracy (Acc.), false positive rate (FPR), and true positive rate (TPR). Results sorted by the AUC scores or by the next metric in case of ties. The lower half shows runs evaluated on a different set of topics. Third-party runs are greyed out. Best results per group are highlighted in bold.

Run	AUC	Acc.	FPR	TPR
h2oloo-gpt3b (best run at TREC 2022 [CMS+22])	**0.95**	**0.86**	**0.04**	0.76
Grid search	0.83	0.72	0.40	0.84
Webis-verasent-dis	0.81	0.70	0.40	0.80
Webis-longck-dis	0.79	0.64	0.36	0.64
Webis-nlm-boolq-abs	0.69	0.52	0.96	**1.00**
Webis-longck-uniqa-dis	0.66	0.62	0.48	0.72
Webis-uniqa-dis	0.66	0.62	0.48	0.72
Webis-longck-uniqa-ax-dis	0.66	0.60	0.48	0.68
Webis-goo-boolq-abs	0.65	0.52	0.96	**1.00**
Webis-nlm-lbert-abs	0.48	0.50	0.80	0.80
Webis-goo-lbert-title-abs	0.48	0.50	0.92	0.92
Webis-goo-lbert-abs	0.48	0.50	0.88	0.88
YouChat (15 topics)	**1.00**	**1.00**	**0.00**	**1.00**
ChatGPT (15 topics)	0.93	0.93	0.14	**1.00**
Pugachev et al. [PAB+23] (113 topics)	0.82	—	—	—

runs outperforms the best automatic run submitted at TREC 2022. The best run we submitted to TREC reaches an AUC score of 0.81. Our multi-stage grid search optimized run (AUC: 0.83, using Vera [PMN+21] for article answer prediction) slightly surpasses the AUC score of our best TREC run by 0.02 but still falls short of the GPT-3-based [BMR+20] h2oloo-gpt3b run by 0.12.

From our proposed answer aggregation strategies (see Section 5.2), the nDCA_k strategy (used in our best TREC runs and the grid search run) is the most effective in predicting the correct answer with consistently lower false positive rates and higher accuracy than with simple averaging and in most cases higher AUC scores. A similar ordering can be observed for the answer prediction model. Here, the claim verification models (Vera [PMN+21] and MultiVerS [WLW+22]) outperform the question answering models. The receiver operating characteristic curves shown in Figure 5.8 indicate that claim verification models are better able to identify most true positives at lower false positive rates than question answering models. But because incorrectly predicting that a treatment would be effective (false positive) is considered more harmful than incorrectly predicting that a treatment would be ineffective (false negative), we also consider the false positive rate as an important measure. Here, the MultiVerS model [WLW+22] is the most effective.

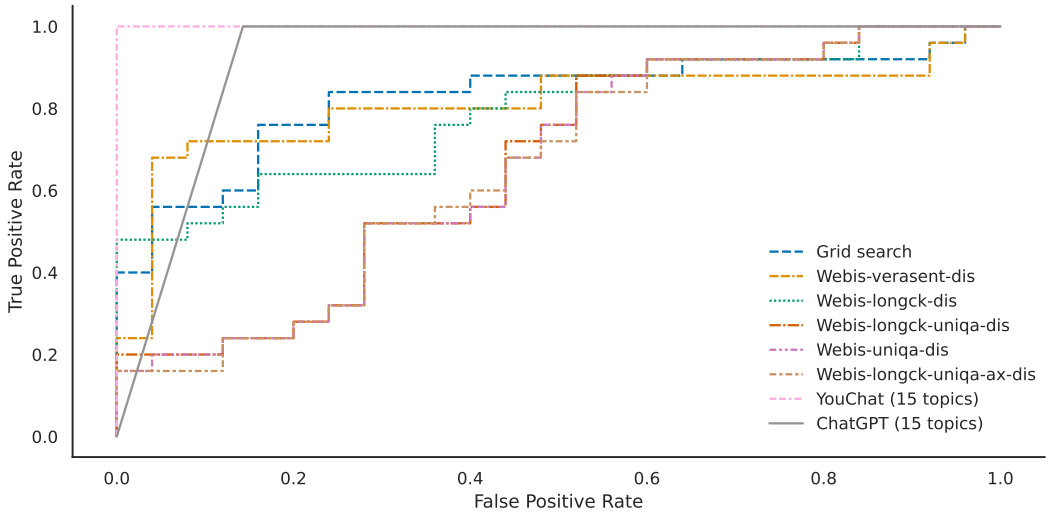


Figure 5.8: Receiver operating characteristic curves of our answer inference approaches for answering health-related yes-no questions. Better models have a larger area under the curve, perfect models touch the top left corner (i.e., FPR: 0, TPR: 1).

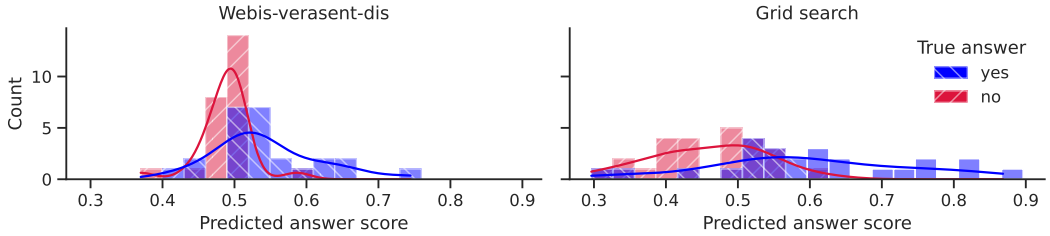


Figure 5.9: Histograms of the predicted answer scores for true answers "yes" or "no" from our best TREC run and grid search optimization. The lines show the kernel-density estimates (Scott's rule [Sco92, pp. 200 sqq.]).

Even though our grid search optimized approach used a more effective evidence retrieval than our TREC runs, answer prediction based on that evidence only slightly improves. As shown in Figure 5.9, the answer scores predicted by the grid search optimized approach span a wider range than the scores predicted by our best TREC run, but the predictions for the conflicting true answers "yes" and "no" largely overlap in both cases. A possible explanation might be that even though the more effective evidence retrieval can capture stronger signal for the correct answer in some cases, it also includes "stronger" evidence for the conflicting answers. Figure 5.9 also highlights that the "no" answer is predicted more accurately.

The two conversational language models perform very well on the 15 sampled questions. You Chat even predicts all 15 questions correctly, and hence, achieves

a perfect AUC score of 1. ChatGPT comes close with an AUC score of 0.93. Even though the results cannot be directly compared due to different test sets (conversational language models were only tested on 15 questions), the results indicate a strong potential for conversational language models to be used for answer inference. The best-performing run at the TREC 2022 Health Misinformation track [CMS+22] also uses a large language model, GPT-3 [BMR+20] (in a few-shot setting) and outperforms the remaining runs. Other evidence-based approaches from related work, e.g., Pugachev et al. [PAB+23] using question answering models, achieve comparable AUC scores like with our evidence-based approach, but direct comparison is not possible because their approach was evaluated on different topics. Even though large language models clearly outperform evidence-based answer inference, their use is questionable due to their reported tendencies to hallucinate and to return factual incorrect or contradicting answers, especially in health-related contexts [BGM+21; KM21; WMR+21]. Recent studies also attribute a lot of the strength of large language models in benchmarks to train-test leakage [FAP+22]. The pre-trained models can only encode knowledge available at the time they were trained, which can quickly become outdated in medical sciences. Because our approach is grounded on credible evidence stored in an easily updatable search index, we therefore still consider evidence-based answer inference to be a promising approach. With our discounting strategy based on the evidence rank, we also make the answer more explainable, because it is understandable that the predicted answer is based on topically relevant articles. In-depth analyses of potential biases in evidence-based approaches are needed to support this claim.

Web Retrieval To assess web retrieval effectiveness, we again measure compatibility [CSV20; CVS20] and nDCG [JK02]. The results are measured using the official relevance judgments from the TREC 2022 Health Misinformation track [CMS+22] and are shown in Table 5.6. All of our TREC runs used the same configuration for web retrieval (see Section 5.5.2) and hence, we only report the results of this configuration once (run (d) in Table 5.6). The effectiveness of the best baseline run at TREC (i.e., without answer-based re-ranking) is reported for reference. The results are similar to our results for evidence retrieval: Our grid search run is more compatible to both helpful and harmful results than our TREC runs and more effective when considering graded relevance judgments. The best difference between helpful and harmful compatibility is measured for the grid search run (0.10) but the compatibility difference is not significantly increased from our TREC runs (0.09). The middle plot in Figure 5.7 also shows that our grid search run improves the compatibility of both helpful and harmful results at a similar rate. Overall, the effectiveness of our web retrieval stage is comparable to the best baseline run at the TREC 2022 Health Misinformation track [CMS+22].

Answer-Based Re-Ranking Our answer-based re-ranking stage is also evaluated using relevance judgments from the TREC 2022 Health Misinformation track [CMS+22]. In Table 5.6, we compare the ten runs we submitted to TREC (runs (h)–(q), see Section 5.5.2), our best run from grid search optimization on TREC 2021 topics (run (g), see Section 5.5.3), and the best participating run at TREC 2022 (run (f)) with respect to the compatibility to helpful or harmful results and the nDCG effectiveness on graded relevance judgments. Even though none of our runs outperform the best participating run from TREC 2022 (run (f)) that achieves a compatibility difference of 0.26 [CMS+22], the variety of re-ranking approaches used in our runs allows us to draw conclusions about the effectiveness of answer score combinations and answer prediction models for reducing misinformation in web search.

Significant effects were only observed for the compatibility to helpful results and nDCG. Our runs featuring a weighted score combination of the original retrieval score and the closeness to the topic answer (runs (g), (h), (l), and (m) in Table 5.6) are often significantly more effective on graded relevance judgments and sometimes also more compatible with helpful results than runs that use linear or polynomial score boosting. Both the highest compatibility difference (0.19, run (g)) and the highest nDCG effectiveness (0.66, run (h)) are measured on runs featuring a weighted score combination, advocating for the use of weighted combinations to combine (estimated) topical relevance and correctness. An explanation for the worse effectiveness of score boosting (linear, polynomial, logarithmic) can be found in the behavior on negative retrieval scores such as returned by monoT5 or duoT5 [PNL21].[16] Figure 5.6 in Section 5.2 shows that the multiplication with the answer closeness only considers the amplitude of the original retrieval score disregarding of its sign. Hence, the retrieval scores of correct documents are actually reduced more than the scores of incorrect documents, exactly the opposite of the intended behavior. One way to mitigate this effect could be to normalize the retrieval scores to positive values prior to answer-based re-ranking.

Apart from the answer score combination, we also observe small differences with different answer prediction models. From our TREC runs using a weighted score combination, the run predicting answer scores with the MultiVerS claim verification model (run (h) in Table 5.6) has a slightly improved compatibility difference and nDCG compared to runs that use the UnifiedQA [KMK+20] question answering model (run (l)) or a combination of both models (run (m)). With score boosting, the opposite effect can be observed, which we attribute to the aforementioned undesired behavior of score boosting on negative retrieval scores. These changes are, however, not significant. Axiomatic re-ranking to resolve answer conflicts (runs (i) and (j) in Table 5.6) also did not change compatibility or effectiveness.

[16] These two T5-based models return log probabilities that are always negative.

Our best runs from the grid search optimization (run (g)) and from TREC (run (h)) are configured very similarly: Both use a weighted score combination of the retrieval score and answer closeness and rely on claim verification models to predict answers. The most important difference between the two runs is that the grid search run uses the Vera model [PMN+21] for answer prediction while the TREC run uses the MultiVerS model [WLW+22]. Both runs have a similarly high compatibility to helpful results (0.26 and 0.27) but our best grid search optimized configuration (run (g)) has a lower compatibility to harmful results of just 0.08 compared to our best TREC run (run (h)) that has a compatibility to harmful results of 0.15. With a compatibility difference of 0.19, the grid search run features the highest compatibility difference of our runs.

Table 5.6 also allows a comparison of retrieval effectiveness prior to and after answer-based re-ranking (compare run (c) vs. (g) or run (d) vs. (h)–(q)). Interestingly, only one of our TREC runs (run (p)) is able to slightly reduce misinformation (i.e., decrease the compatibility to harmful results) by answer-based re-ranking from 0.08 to 0.07 but at the same time also decreases the compatibility to helpful results from 0.17 to 0.11. The only TREC run with an improved compatibility difference after answer-based re-ranking (run (h)) increases both helpful and harmful compatibility. Both cases are equally undesirable because neither should answer-based re-ranking make the search results more harmful nor should it make them less helpful. The grid search optimized run (run (g)) on the other hand is able to decrease the compatibility to harmful results from 0.16 to 0.08 and only slightly decreases the compatibility to helpful results from 0.27 to 0.26. In the leftmost plot of Figure 5.7 the grid search optimized run (yellow cross mark) is therefore shifted towards the lower right corner.

We conclude that answer-based re-ranking is an effective way to reduce misinformation in web search. To first infer the topic answer, our best approach retrieves medical abstracts from PubMed as evidence, applies the pre-trained Vera claim verification model using the abstracts as context for the model [PMN+21], and considers the ranking position while aggregating the individual answer scores. Evidence can be retrieved with a standard two-stage re-ranking architecture using BM25 scoring and pointwise monoT5 re-ranking [PNL21; RWJ+94]. After retrieving web documents with the same re-ranking architecture, the same Vera claim verification model can be used to predict answers using the web documents as context. A weighted combination of each document's predicted correctness to the binarized topic answer and its original retrieval score is then used to reduce misinformation while preserving most of the retrieval effectiveness. This evidence-based approach can be favorable to current better performing approaches based on large language models, especially in the health-related context studied in the TREC Health Misinformation tracks, because in contrast to language models, our approach can be explained and the influence of each document on the final ranking can be analyzed.

Our analyses also show disadvantages in the generalizability of our approach that should be addressed in future work.

5.6 Summary

In this chapter, we have built a retrieval pipeline for answering health-related yes/no questions regarding the effectiveness of medical treatments. Our approach first retrieves relevant scientific evidence, then predicts an answer based on that evidence, and finally re-ranks retrieved web documents based on their alignment with the predicted answer. We have described our submissions to the TREC 2022 Health Misinformation track [CMS+22] and optimized the hyperparameters of our pipeline in a multi-stage grid search. Evaluations of the evidence retrieval stage indicate that monoT5 re-ranking [PNL21] after retrieval from the PubMed with BM25 scoring [RWJ+94] can achieve high compatibility [CSV20; CVS20] to helpful results but at the same time is nearly as compatible to harmful results (compatibility difference: 0.10, higher is better). Evidence retrieval was not evaluated in previous TREC tracks or other related work, so strong baselines are missing. Based on our evaluations, we therefore cannot conclude an answer to Research Question 7. To infer an answer based on the evidence, our best approach uses the pre-trained Vera claim verification model [PMN+21] to predict article-level answers and aggregates the answers discounted by article ranks (i.e., determined by retrieval scores). On TREC topics we achieve an AUC score of 0.83 with a false positive rate of 0.40, but do not outperform the GPT-3-based [BMR+20] state-of-the-art approach (AUC: 0.95, FPR: 0.04). On a subset of 15 of the 50 evaluated TREC topics, we achieve perfect classification effectiveness with zero-shot prompting the large proprietary You Chat language model. Both results indicate that on a small test set, large language models can answer health-related yes/no questions correctly without spreading misinformation (Research Question 6). However, language models are static, opaque and often lack factual correctness [BGM+21; KM21; WMR+21]. If we add the requirement of being adaptable to new evidence or more explainable, our evidence-based approach is more favorable than the language model-based approach and still mostly yields correct answers (with few exceptions on topics that lack evidence in scientific literature). Our best web retrieval approach is similar to our evidence retrieval approach except that it retrieves from the C4 web corpus [DSM+21]. It also retrieves results that are nearly as compatible with harmful results as with helpful results (compatibility difference: 0.10). We apply answer-based re-ranking to reduce misinformation based on the closeness of the answer predicted for the document and the predicted "true" topic answer. For answer-based re-ranking, our best configuration uses a weighted score combination of the original retrieval score and the answer closeness, and it also uses the Vera claim verification

model [PMN+21]. With this additional stage to mitigate misinformation, we achieve a compatibility difference of 0.19. The best TREC participant run outperforms our approach (compatibility difference: 0.26) by using predictions from GPT-3 [BMR+20]. Yet, our evidence-based approach comes close to the state-of-the-art approach. We therefore see our approach as a promising alternative to the state-of-the-art approaches that are based on large language models and hence suffer from the aforementioned disadvantages, but cannot conclude Research Question 8 positively.

Chapter 6

Conclusion

In this thesis, we have addressed the problems of identifying and answering health-related questions from the perspective of a search engine. The following Section 6.1 concludes the main findings and the main contributions of this thesis: (1) A large-scale, health-related question dataset, (2) effective classifiers to identify health-related and medical questions, and (3) an evidence-based information retrieval system for answering health-related yes-no questions. We also discuss the limitations of our work and propose future work.

6.1 Contributions

The contributions of this thesis are three-fold. In Chapter 3 we have collected a large-scale dataset of health-related and medical questions and applied weak supervision to automatically label the questions. Chapter 4 presents a set of feature-based and neural classifiers to identify health-related and medical questions. Finally, in Chapter 5, we have built a modular information retrieval pipeline to first answer health-related yes-no questions and subsequently retrieve web documents that support the predicted answer. In the following, we summarize the main findings of each chapter and discuss the research questions posed in Chapter 1.

Data Collection and Labeling Our new health-related question answering dataset described in Chapter 3 unites many existing task-specific but smaller medical datasets as well as larger general purpose question datasets from query logs and community platforms. The dataset consists of 8.5 million automatically labeled questions among which 2.0 million are health-related and 1.3 million are medical, making it the largest dataset of health-related questions to date.

The automatic labels are based on a set of heuristic labeling functions whose output was then used to train a label model [RHD+19]. We have found that a question's source dataset or category and occurrences of health-related terms such as drug names and medical conditions are good indicators for health-related questions. But additional heuristics were required to effectively label medical questions. Our evaluation based on a set of 7,500 manually annotated questions

sampled from the dataset shows that our automatic labeling approach achieves a high recall, but a low precision for health-related and medical questions. Yet, the dataset's label distribution (manual: 17 % health-related, 7 % medical; automatic: 23 % health-related, 15 % medical) is similar to the distribution of health-related questions reported in literature (5-24 %). We can therefore give a positive answer to Research Question 1: Our dataset does realistically represent the real world distribution of health-related questions. With regard to Research Question 2, we are unable to give a clear answer. Automatic labeling tends to be overly optimistic in predicting questions as health-related and medical. We therefore recommend to not solely rely on automatic labels to train classifiers if they require high precision. Research Question 3 is inconclusive as well. Our exploratory analysis shows slight differences between the length of medical and non-medical questions but other characteristics such as topical similarity were not yet considered in this thesis.

Identifying Health-Related Questions The second contribution of this thesis is a set of classifiers to identify health-related and medical questions. In Chapter 4 we have trained feature-based classifiers based on sentence embeddings and fine-tuned pre-trained encoder models as well as causal and text-to-text language models to cover a wide spectrum of neural text classification approaches. Our best classifier for health-related questions is a fine-tuned text-to-text language model [GAU+22] that achieves an F_1-score of 0.80 and an AUC score (area under the receiver operating characteristic curve) of 0.96 on the manually labeled test set of our question dataset. The best medical question classifier, a cascading support vector machine [CL11] trained to classify sentence embeddings of questions, achieves an F_1-score of 0.69 and an AUC score of 0.98. Both were trained on the training and validation splits of our manually labeled subset. We can therefore conclude for Research Question 4 that language model fine-tuning is an effective classification approach for health-related questions. Feature-based machine learning on embedded sentences is relatively effective to identify medical questions.

If a higher recall is desired, our results indicate that training encoder models on the larger automatically labeled dataset is more effective. This approach achieves a recall of 0.81 for both classifying either health-related or medical questions and can therefore best be used in web search engines to identify health-related questions. The downside is a lower precision of only 0.67 for health-related questions and 0.42 for medical questions, respectively, that is likely caused by the limited precision of our automatic labeling approach. It therefore largely depends on the notion of effectiveness, i.e., whether identifying health-related questions is considered to be a recall-oriented or precision-oriented task, to answer Research Question 5. Training on the manually annotated gold label dataset yields a higher precision but training on automatic silver labels yields a higher recall.

Answering Health-Related Questions Lastly, in Chapter 5 we have demonstrated a modular retrieval pipeline for answering a subset of the questions that can be identified as health-related: Yes-no questions about the effectiveness of medical treatments. Besides correctly answering the question in most cases, our pipeline also retrieves web documents that support the predicted answer. We have evaluated this question answering and retrieval system on the TREC 2022 Health Misinformation track [CMS+22] and have measured the effectiveness of each of the four stages in the pipeline: (1) Evidence retrieval (by the difference in compatibility to the most "helpful" and "harmful" ranking, see Section 5.5), (2) answer inference (AUC), (3) web retrieval, and (4) answer-based re-ranking (both compatibility difference).

The proposed evidence retrieval stage achieves a high nDCG effectiveness with our best configuration (BM25 scoring [RWJ+94] on a PubMed index and subsequent monoT5 re-ranking [PNL21]) but is nearly as compatible with harmful results as with helpful results (compatibility difference: 0.10). Due to the lack of baselines for evidence retrieval (evidence retrieval effectiveness was not considered at TREC), we cannot conclude an answer to Research Question 7.

Answer inference uses the retrieved evidence to predict the answer with pretrained claim verification models. The best configuration from our experiments uses the Vera claim verification model [PMN+21] on PubMed abstracts and aggregates the answers based on the ranks of the retrieved abstracts. Our evidence-based answer inference approach achieves an AUC score of 0.83 but is still outperformed by the best participating approach at TREC 2022. The state-of-the-art approach uses the GPT-3 language model [BMR+20] yields a higher AUC score (0.95) and lower false positive rates (0.04, our best: 0.40) but has the disadvantage of being less explainable due to using outputs of a black-box commercial language model. We confirm the higher effectiveness of large language models in a pilot study using two conversational language models and hence draw a positive conclusion to Research Question 6. Because of the recently emerging questions regarding the fairness, accountability and transparency of such models [BGM+21; KM21; WMR+21], a reasonable reformulation of the research question might also be: *Can we answer* current *health-related yes-no questions correctly,* explainable*, and without spreading misinformation?* Because our answer inference approach outperforms other approaches based on evidence retrieval (which is both adaptable to current topics and more explainable than large language models), we can still answer the adjusted research question positively.

Web retrieval suffers from the same problem as evidence retrieval. It is also nearly as compatible with harmful results as with helpful results (compatibility difference: 0.10) because it does not (yet) consider the true answer to the health-related question. But even without answer-based re-ranking, our best configuration (monoT5 re-ranking [PNL21] after BM25) improves the compatibility to helpful results compared to the best baseline run from TREC (just BM25, compatibility dif-

ference: 0.05). The best answer-based re-ranking configuration in our experiments uses a weighted score combination of retrieval score and answer closeness and relies on the same claim verification model used for answer inference to predict answers (Vera [PMN+21]). This answer-based re-ranking approach achieves a compatibility difference of 0.19 by reducing the compatibility to harmful results. Again, the best participating run at TREC outperforms our approach (compatibility difference: 0.26) but because it uses answer predictions from GPT-3, it suffers from the same biases as mentioned earlier for answer inference. The evidence-based approach proposed in this thesis comes close to the state-of-the-art approach. To answer Research Question 8, answer-based re-ranking is a promising way to retrieve web documents that support the true answer to a health-related question, but large language models are yet more effective. Our results also indicate that re-ranking depends on the answer prediction effectiveness. In the future, we therefore plan to decouple the evaluation of answer-based re-ranking by experimenting on ground-truth answers.

6.2 Future Work

Our work has demonstrated that health-related questions can effectively be identified and answered by exploiting advances in machine learning, natural language processing, and information retrieval. However, there still is a plethora of challenges that need to be addressed in the future. In the following, we discuss some challenges and advances that we plan to address in future work to build an end-to-end search engine for health-related information needs.

Considering that not all queries in search engines are formulated as natural language questions, the most obvious next step is to not only consider questions to identify health-related information needs but also keyword queries. A similar approach as described in Chapter 3 could be applied to automatically label query logs with health-related queries. Initial experiments have shown that the most difficult part of this extension would be to find suitable labeling functions to automatically label health-related queries. In contrast to question answering datasets, query logs are rarely annotated with categories and to our knowledge, there is only one query log that only contains health-related queries [RLS+21]. Building a large-scale query log with health-related queries could also benefit reproducing studies on the prevalence of health-related queries and the biases that have been observed in web search [AS19; BZE22; WA14; WH15]. These studies are often based on smaller or proprietary query logs, so a large scale analysis of biases in health-related search across multiple search engines is a valuable contribution to the field.

To extend our dataset of health-related questions, multilingual question answering datasets could also be a valuable addition, for example in Chinese [YZL+14] or Russian [BKB13]. After finalization of our dataset, there have also been new

question answering datasets released that could be added to our dataset, for example HealthSearchQA [SAT+22], CausalQA [BWH+22], and RedHOT [WKA+22]. By analyzing large datasets of questions and queries, we could also build a taxonomy of health-related information needs similar to already existing taxonomies for web search [AKV22; Bro02; CTS+21]. To further improve identifying health-related or medical questions or queries, interesting directions include the application of zero-shot language models like T0 [SWR+22] or improving efficiency by distilling well-performing larger text-to-text language models into smaller models.

Our retrieval and question answering pipeline is currently only capable of answering health-related yes-no questions. Generalizing our approaches to open-ended questions is thus a natural next step. Due to the criticism of large language models for hallucination, factual incorrectness, contradictions, and possible train-test-leakage effects [BGM+21; FAP+22; KM21; LHE22], we still see our evidence-based approach as a promising approach to answer open-ended questions. Though, we have already noticed that retrieving evidence solely from scientific literature is limited in the sense that it often lacks information about alternative remedies. Here, we see potential in including high-quality newspapers, websites from health organizations, and health-related knowledge bases to provide evidence for open-ended questions. Another direction is to approach health-related question answering in a multi-hop fashion, thereby reducing the complexity of individual tasks or to employ a conversational system that allows users to ask follow-up questions to better understand the answer.

We also believe that evidence retrieval is conceptually similar to finding studies for a systematic review. Future work should therefore also exploit automated procedures from systematic reviews to improve evidence retrieval for health-related question answering systems. This includes to not only look at the abstract and title of each article but rather their full text which is often publicly available.[1] Also, the type of evidence could probably be used to improve the ranking of evidence retrieval results. Clearly, a systematic review can be trusted more than a case study. A promising way to use this metadata could be retrieval axioms. Because some neural models cannot effectively work on longer texts such as full scientific articles, we could apply summarization to improve health-related claim verification, an approach that has been used previously to identify facts in news articles [BKC22]. Summarization techniques are also necessary to make the evidence used from scientific articles understandable to the public.

Different query variants of the same underlying information need were found to have adverse effects on the effectiveness of information retrieval systems [ZPH16]. In this regard, the single-query evaluation setting of the TREC Health Misinformation tracks [CMS+22] can be problematic. In the future, we therefore suggest

[1] E.g., on PubMed Central: `https://ncbi.nlm.nih.gov/pmc`

that health-related question answering systems be evaluated with respect to user query variants, where the same question is asked in different ways. Our dataset of health-related questions can serve as a first building block to match similar questions to existing benchmarks' topics. We could also use our classifiers to identify health-related questions in existing datasets of user query variations [BMS+16]. Larger query logs can similarly augment the topics from benchmark collections to make shared task evaluations more realistic.

Finally, we also plan to conduct end-to-end user studies. An interesting question is how users perceive search results from existing search engines like Google or Bing that do not warn users about health misinformation. Comparing user satisfaction to a new system that identifies health-related questions and answers them correctly, maybe even in a specialized search interface, could serve as empirical arguments to bring recent advances in the detection of health misinformation (e.g., at the TREC shared tasks [CMS21]) to real-world search engines. A joint end-to-end evaluation of systems for health-related and non-health-related search would also be a first step to tackle the criticized modular evaluation of current health-related question answering and information retrieval systems [JYX+23].

Appendix A

Annotator Instructions

The following sections contain the instructions that were given to the annotators for the three annotation tasks conducted in this thesis: Annotating yes-no questions (Section A.1), annotating health-related and medical questions (Section A.2), and annotating relevance and answers for evidence retrieval (Section A.3).

A.1 Yes-No Question Annotation

The task is to identify whether a text is a question and whether it is a yes-no question.

Question or not?

A piece of text can contain several sentences.

Label a text as a question if:

1. it is one single interrogative sentence, that is, a question
   Examples:
   - Does aspirin help with headaches?
   - How much aspirin should I take?
2. it contains several sentences, one of which is a question (see above)
   Examples:
   - I have a headache. Does aspirin help?
   - Where can I buy aspirin? My head hurts!
3. typos, smaller grammatical errors, or incorrect word ordering should be ignored (label as a question)
   Examples:
   - Do aspirin help with headache? (typos)
   - Does aspirin help headache? (missing word)
   - Does aspirin help with headaches (missing question mark)

Label a text as **not** a question if all sentences within the text are:

1. declarative sentences (assertions or statements)
   Examples:
   - Aspirin is good for headaches.
   - I use aspirin when I have a headache.
2. imperative sentences, i.e., commands
   Examples:
   - Use aspirin if you have headaches.
   - Tell me if aspirin is good for headaches.
3. exclamatory sentences, i.e., exclamations
   Examples:
   - How good aspirin is!
   - What a bad drug aspirin is!
4. ill-formed questions, i.e., questions that do not start with a question word or auxiliary verb
   Examples:
   - Aspirin helps for headache?
   - Take aspirin at night?

Yes-no question or not?

Label a text as a yes-no question if:

1. the text is a question (see above) *and*
2. the question can be answered by either "yes" or "no" *Examples:*
   - Should I use aspirin?
   - Does aspirin help with headaches?

Label a text as **not** a yes-no question if:

- the text is *not* a question (see above) *or*
- the question is open-ended, i.e., it starts with a question word (where? who? and so on)
  Examples:
  - Who can use aspirin?
  - How effective is aspirin for headaches?
- the question has multiple explicit choices, e.g., comparisons
  Examples:
  - I have a headache. Should I use aspirin or paracetamol?
  - Does aspirin work better for children or adults?

Summary

Please read each text and label whether the text is a yes-no question (shortcut 1), a question (shortcut 2), or not interrogative at all (shortcut 3).

A.2 Health-Related and Medical Question Annotation

The task is to identify whether a natural question is health-related and/or medical.

Health-related or not?

Health-related question can be any question about human or animal health, both physical and mental. Some of the topics such questions can address are:

- physical, mental, social well-being
  Examples: I'm not feeling well, what should I do? How to cope with being bullied in class?
- diseases, illnesses, disorders, medical conditions
  Examples: I got the flu, should I call a doctor? How do I know I have measles?
- physical or mental states (like pregnancy, aging, etc.)
  Examples: Can I smoke during pregnancy? What is a normal age to get hair loss?
- diagnosis, prevention, risk factors
  Examples: How to test for covid? Are men more likely to get a heart attack?
- treatments, medication, drugs, exercises
  Examples: Can aspirin reduce fever? I had a stroke, which exercises can I do?
- healthcare service, social measures
  Examples: Where's the next physician? Does the american healthcare system work?
- anatomy, biochemical processes
  Examples: How much oxygen does a fish need to survive? How tall is the average woman?
- fitness, sports, lifestyle, sex, and nutrition
  Examples: How often should I go to the gym? Should I use a condom? Is green tea healthy?

Label a question as **not** health-related if:

- it is only asking for references or navigation
  Examples: What papers are worth reading about cancer? Where can I find websites about pilates?
- it is a purely factual biological question (i.e., there is no direct influence on

health)
Examples: Can fish breathe? Are bacteria animals?

Medical or not?

Medical questions are health-related questions that require additional, professional expertise, that often motivate clinical studies and research.

Label a question as medical if:

- the question is health-related (see above) *and*
- one would usually seek professional advice (e.g., from a doctor, nurse, pharmacist, or therapist)
  Examples:
  - Why do I have diarrhea for 3 days?
  - I haven't had my period. Am I pregnant?
- it cannot be answered by laymen
  Examples:
  - Which variant of Covid is the most dangerous?
  - What are the best therapy options for depression?
- wrong answers could cause severe harm
  Examples:
  - What is the best dosage of ibuprofen for a child?
  - How to cure a snakebite?

Label a question as **not** medical if:

- the question is *not* health-related (see above) *or*
- the answer is mostly common sense, even for laymen
  Examples:
  - Are soft or hard mattresses healthier?
  - Do bacteria cause harm to my health?
  - Can I drink too much beer?
- the answer mostly depends on personal preference
  Examples:
  - How many steps should I do each day?
  - What's your preferred way of losing weight?

Summary

Please read each question and check whether it is medical (shortcut 1), otherwise health-related (shortcut 2), or not health-related at all (shortcut 3).

A.3 Evidence Relevance and Answer Annotation

The task is to identify whether a scientific abstract is relevant to answer the given health-related question.
For each question, assess abstracts on:

- How relevant is this abstract for answering the question?
- For abstracts that are relevant, what answer do they support? "Yes" or "no"?

Relevance

- Relevant (*see answer labels below*)
  The abstract either directly answers the question or provides enough information to determine an answer. A relevant abstract must address *all* parts of the question and help to make a yes/no decision.
- Not relevant (shortcut 3)
  The abstract either does not address the question, or fails to address all parts of a question.
  An abstract is **not** relevant if it:
  - only asks about the effectiveness of a specific treatment but only merely mentions the health issue or treatment of the question
  - describes an animal study and does not explicitly mention applicability to humans
  - is not English
  - contains adult material, or
  - is garbled, empty, unreadable or otherwise broken.

  Example: If the question is "Does yoga improve the management of asthma?", and the abstract only talks about yoga without talking about asthma or talks about asthma but not yoga, then the abstract is not-relevant.

Important: For a relevant abstract, it does not matter whether you believe the information provided in that abstract is correct or incorrect. Only judge whether a user would likely find the information relevant regardless of its correctness.

Example: Two relevant abstracts could have different answers (i.e., yes and no) to the same question, but both would be viewed to be high quality results from credible sources suitable as top 10 web search results.

Answer

For relevant abstracts, judge whether the answer to the question is "yes" or "no" according to the abstract.

- Yes (shortcut `0`)
  The abstract says the answer to the question is "yes" or provides strong support that would lead to the conclusion that the answer is "yes".
- No (shortcut `1`)
  The abstract says the answer to the question is "no" or provides strong support that would lead to the conclusion that the answer is "no".
- Unclear (shortcut `2`)
  The abstract addresses the question, but a user would not be able to conclude either "yes" or "no" given the abstract.
  An abstract has an unclear answer if it:
  - is a meta-analysis or systematic review without a final conclusion
  - concludes an answer only for animals but not for humans
  - the answer could be found in the full study but is not provided in the abstract

Summary

Please read each topic's question and description. Then rate each abstract's relevance and answer to the question (only if abstract is relevant).

Appendix B

Used Models

In Table B.1, we provide direct links to all models used in this thesis.

Table B.1: Links to the model checkpoints used for the classification of health-related and medical questions, claim verification, and question answering, sorted alphabetically.

Model	Link
BART	https://huggingface.co/facebook/bart-base
BERT	https://huggingface.co/bert-base-uncased
BioGPT	https://huggingface.co/microsoft/biogpt
BioLinkBERT	https://huggingface.co/michiyasunaga/BioLinkBERT-base
BioMedLM	https://huggingface.co/stanford-crfm/BioMedLM
duoT5	https://huggingface.co/castorini/duot5-3b-med-msmarco
	https://huggingface.co/castorini/duot5-3b-msmarco
	https://huggingface.co/castorini/duot5-base-msmarco
Flan-T5	https://huggingface.co/google/flan-t5-base
Galactica	https://huggingface.co/facebook/galactica-125m
GPT-2	https://huggingface.co/gpt2
GPT-Neo	https://huggingface.co/EleutherAI/gpt-neo-125M
Instructor	https://huggingface.co/hkunlp/instructor-base
LongT5	https://huggingface.co/google/long-t5-tglobal-base
MiniLM	https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
	https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2
monoT5	https://huggingface.co/castorini/monot5-3b-med-msmarco
	https://huggingface.co/castorini/monot5-3b-msmarco
	https://huggingface.co/castorini/monot5-base-msmarco
MPNet	https://huggingface.co/sentence-transformers/all-mpnet-base-v2
MultiVerS	https://scifact.s3.us-west-2.amazonaws.com/longchecker/latest/checkpoints/fever.ckpt
	https://scifact.s3.us-west-2.amazonaws.com/longchecker/latest/checkpoints/fever_sci.ckpt
	https://scifact.s3.us-west-2.amazonaws.com/longchecker/latest/checkpoints/healthver.ckpt
	https://scifact.s3.us-west-2.amazonaws.com/longchecker/latest/checkpoints/scifact.ckpt
OPT	https://huggingface.co/facebook/opt-125m
PubMedBERT	https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract
RoBERTa	https://huggingface.co/roberta-base
	https://huggingface.co/apugachev/roberta-large-boolq-finetuned
UnifiedQA	https://huggingface.co/allenai/unifiedqa-t5-large
SciBERT	https://huggingface.co/allenai/scibert_scivocab_uncased
SciFive	https://huggingface.co/razent/SciFive-base-Pubmed
T5	https://huggingface.co/t5-base
Vera	gs://castorini/vera/experiments/3B (Google file system)

Bibliography

[46] "Constitution of the World Health Organization." In: *American Journal of Public Health and the Nation's Health* 36.11 (Nov. 1, 1946), pp. 1315–1323. ISSN: 0002-9572. DOI: 10.2105/ajph.36.11.1315 (cit. on p. 32).

[ACG+21] Mustafa Abualsaud, Irene Xiangyi Chen, Kamyar Ghajar, Linh Nhi Phan Minh, Mark D. Smucker, Amir Vakili Tahami, and Dake Zhang. "UWaterlooMDS at the TREC 2021 Health Misinformation Track." In: 30th Text REtrieval Conference. TREC 2021 (Gaithersburg, Maryland, United States, Nov. 15–19, 2021). Vol. 500-335. NIST Special Publication. National Institute of Standards and Technology, 2021. URL: https://trec.nist.gov/pubs/trec30/papers/UwaterlooMDS-HM.pdf (visited on 03/28/2023) (cit. on p. 16).

[ACN08] Nir Ailon, Moses Charikar, and Alantha Newman. "Aggregating Inconsistent Information: Ranking and Clustering." In: *Journal of the ACM* 55.5, 23 (2008), 23:1–23:27. ISSN: 1557-735X. DOI: 10.1145/1411509.1411513 (cit. on p. 71).

[AKV22] Daria Alexander, Wojciech Kusa, and Arjen P. de Vries. "ORCAS-I: Queries Annotated with Intent using Weak Supervision." In: 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2022 (Madrid, Spain, July 11–15, 2022). Association for Computing Machinery, 2022, pp. 3057–3066. DOI: 10.1145/3477495.3531737 (cit. on pp. 12, 99).

[AR22] Gavin Abercrombie and Verena Rieser. "Risk-graded Safety for Handling Medical Queries in Conversational AI." In: 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and 12th International Joint Conference on Natural Language Processing. AACL/IJCNLP 2022 (Virtual Event, Nov. 20–23, 2022). Vol. 2, Short Papers. Association for Computational Linguistics, 2022, pp. 234–243. URL: https://aclanthology.org/2022.aacl-short.30 (visited on 03/28/2023) (cit. on pp. 10, 25).

[AS19] Mustafa Abualsaud and Mark D. Smucker. "Exposure and Order Effects of Misinformation on Health Search Decisions." In: Workshop on Reducing Online Misinformation Exposure. ROME 2019 (Paris,

France, July 25–25, 2019). 2019. URL: https://rome2019.github.io/papers/Abualsaud_Smucker_ROME2019.pdf (visited on 03/28/2023) (cit. on pp. 5, 72, 98).

[ASL+19] C. Lioma Abualsaud, Mark D. Smucker, Christina Lioma, Maria Maistro, and Guido Zuccon. "Overview of the TREC 2019 Decision Track." In: 28th Text REtrieval Conference. TREC 2019 (Gaithersburg, Maryland, United States, Nov. 13–15, 2019). Vol. 500-331. NIST Special Publication. National Institute of Standards and Technology, 2019. URL: https://trec.nist.gov/pubs/trec28/papers/OVERVIEW.D.pdf (visited on 09/30/2022) (cit. on pp. 2, 10, 16, 17, 23, 67, 76, 78, 79).

[Azz21] Leif Azzopardi. "Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval." In: ACM SIGIR Conference on Human Information Interaction and Retrieval. CHIIR 2021 (Canberra, Australia, Mar. 14–19, 2021). Association for Computing Machinery, 2021, pp. 27–37. DOI: 10.1145/3406522.3446023 (cit. on p. 5).

[BC00] David M. Baorto and James J. Cimino. "An 'Infobutton' for Enabling Patients to Interpret On-line Pap Smear Reports." In: American Medical Informatics Association Annual Symposium. AMIA 2000 (Los Angeles, California, United States, Nov. 4–8, 2000). American Medical Informatics Association, 2000, pp. 47–50. PMCID: PMC2243949 (cit. on p. 13).

[BCI07] Francesca Borrelli, Raffaele Capasso, and Angelo A. Izzo. "Garlic (Allium sativum L.): Adverse Effects and Drug Interactions in Humans." In: *Molecular Nutrition & Food Research* 51.11 (Oct. 26, 2007), pp. 1386–1397. ISSN: 1613-4133. DOI: 10.1002/mnfr.200700072 (cit. on pp. 1, 2, 5).

[BD16] Asma Ben Abacha and Dina Demner-Fushman. "Recognizing Question Entailment for Medical Question Answering." In: American Medical Informatics Association Annual Symposium. AMIA 2016 (Chicago, Illinois, United States, Nov. 12–16, 2016). American Medical Informatics Association, 2016, pp. 310–318. PMCID: PMC5333286 (cit. on p. 22).

[BD19a] Asma Ben Abacha and Dina Demner-Fushman. "A Question-Entailment Approach to Question Answering." In: *BMC Bioinformatics* 20.1, 511 (2019), 511:1–511:23. ISSN: 1471-2105. DOI: 10.1186/s12859-019-3119-4 (cit. on pp. 10, 20, 22).

[BD19b] Asma Ben Abacha and Dina Demner-Fushman. "On the Summarization of Consumer Health Questions." In: 57th Conference of the Association for Computational Linguistics. ACL 2019 (Florence, Italy, July 28–Aug. 2, 2019). Vol. 1, Long Papers. Association for Computational Linguistics, 2019, pp. 2228–2234. DOI: 10.18653/v1/p19-1215 (cit. on pp. 10, 21).

[BFC22] Patrick V. Barnwell, Erick J. Fedorenko, and Richard J. Contrada. "Healthy or Not? The Impact of Conflicting Health-Related Information on Attentional Resources." In: *Journal of Behavioral Medicine* 45.2 (Apr. 2022), pp. 306–317. ISSN: 1573-3521. DOI: 10.1007/s10865-021-00256-4 (cit. on p. 5).

[BFG+22] Alexander Bondarenko, Maik Fröbe, Lukas Gienapp, Alexander Pugachev, Jan Heinrich Reimer, Ferdinand Schlatt, Ekaterina Artemova, Martin Potthast, Benno Stein, Pavel Braslavski, and Matthias Hagen. "Webis at TREC 2022: Deep Learning and Health Misinformation." In: 31st Text REtrieval Conference. TREC 2022 (Virtual Event, Nov. 14–18, 2022). Vol. 500-338. NIST Special Publication. National Institute of Standards and Technology, 2022. URL: https://trec.nist.gov/pubs/trec31/papers/Webis.DH.pdf (visited on 03/28/2023) (cit. on pp. 17, 78).

[BFK+19] Alexander Bondarenko, Maik Fröbe, Vaibhav Kasturia, Matthias Hagen, Michael Völske, and Benno Stein. "Webis at TREC 2019: Decision Track." In: 28th Text REtrieval Conference. TREC 2019 (Gaithersburg, Maryland, United States, Nov. 13–15, 2019). Vol. 500-331. NIST Special Publication. National Institute of Standards and Technology, 2019. URL: https://trec.nist.gov/pubs/trec28/papers/Webis.D.pdf (visited on 03/28/2023) (cit. on p. 17).

[BFR+22] Alexander Bondarenko, Maik Fröbe, Jan Heinrich Reimer, Benno Stein, Michael Völske, and Matthias Hagen. "Axiomatic Retrieval Experimentation with ir_axioms." In: 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2022 (Madrid, Spain, July 11–15, 2022). Association for Computing Machinery, 2022, pp. 3131–3140. DOI: 10.1145/3477495.3531743 (cit. on pp. 17, 71).

[BGM+21] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: 4th ACM Conference on Fairness, Accountability, and Transparency. FAccT 2021 (Virtual Event, Mar. 3–10, 2021). Association for Computing Machinery, 2021, pp. 610–623. DOI: 10.1145/3442188.3445922 (cit. on pp. 90, 93, 97, 99).

[BHY+22] Elliot Bolton, David Hall, Michihiro Yasunaga, Tony Lee, Chris Manning, and Percy Liang. *PubMedGPT 2.7B*. Dec. 15, 2022. URL: https://crfm.stanford.edu/2022/12/15/pubmedgpt.html (visited on 03/28/2023) (cit. on pp. 14, 48).

[Bie20] Lukas Biewald. *Experiment Tracking with Weights and Biases*. 2020. URL: https://wandb.com (visited on 04/06/2021) (cit. on pp. 47, 50, 53, 82).

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. 1st ed. Information Science and Statistics. Springer, 2006. 738 pp. ISBN: 978-0-387-31073-2 (cit. on p. 59).

[BKB13] Alexander Beloborodov, Artem Kuznetsov, and Pavel Braslavski. "Characterizing Health-Related Community Question Answering." In: *Advances in Information Retrieval*. 35th European Conference on IR Research. ECIR 2013 (Moscow, Russia, Mar. 24–27, 2013). Vol. 7814. Lecture Notes in Computer Science. Springer, 2013, pp. 680–683. DOI: 10.1007/978-3-642-36973-5_59 (cit. on p. 98).

[BKC22] Varad Bhatnagar, Diptesh Kanojia, and Kameswari Chebrolu. "Harnessing Abstractive Summarization for Fact-Checked Claim Detection." In: 29th International Conference on Computational Linguistics. COLING 2022 (Gyeongju, Republic of Korea, Oct. 12–17, 2022). International Committee on Computational Linguistics, 2022, pp. 2934–2945. URL: https://aclanthology.org/2022.coling-1.259 (visited on 03/28/2023) (cit. on p. 99).

[BKK+03] Nicholas J. Belkin, Diane Kelly, Giyeong Kim, Ja-Young Kim, Hyuk-Jin Lee, Gheorghe Muresan, Muh-Chyun (Morris) Tang, Xiaojun Yuan, and Colleen Cool. "Query Length in Interactive Information Retrieval." In: 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2003 (Toronto, Ontario, Canada, July 28–Aug. 1, 2003). Association for Computing Machinery, 2003, pp. 205–212. DOI: 10.1145/860435.860474 (cit. on pp. 28, 43).

[BKK+15] Georgios Balikas, Aris Kosmopoulos, Anastasia Krithara, Georgios Paliouras, and Ioannis A. Kakadiaris. "Results of the BioASQ Tasks of the Question Answering Lab at CLEF 2015." In: *Working Notes of CLEF 2015*. 6th International Conference of the CLEF Association. CLEF 2015 (Toulouse, France, Sept. 8–11, 2015). Vol. 1391. CEUR Workshop Proceedings. CEUR-WS.org, 2015. URL: https://ceur-ws.org/Vol-1391/inv-pap7-CR.pdf (visited on 03/28/2023) (cit. on pp. 10, 21).

[BL04] Steven Bird and Edward Loper. "NLTK: The Natural Language Toolkit." In: 42nd Annual Meeting of the Association for Computational Linguistics. ACL 2004 (Barcelona, Spain, July 21–26, 2004). Association for Computational Linguistics, 2004, pp. 214–217. URL: https://aclanthology.org/P04-3031 (visited on 09/30/2022) (cit. on pp. 27, 31, 36, 43).

[BLC19] Iz Beltagy, Kyle Lo, and Arman Cohan. "SciBERT: A Pretrained Language Model for Scientific Text." In: Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing. EMNLP-IJCNLP 2019 (Hong Kong, China, Nov. 3–7, 2019). Association for Computational Linguistics, 2019, pp. 3613–3618. DOI: 10.18653/v1/D19-1371 (cit. on pp. 11, 48).

[BLW+21] Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. *GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow*. Version 1.0. Mar. 21, 2021. DOI: 10.5281/zenodo.5297715 (cit. on p. 48).

[BMR+20] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. "Language Models are Few-Shot Learners." In: *Advances in Neural Information Processing Systems*. 34th Annual Conference on Neural Information Processing Systems. NeurIPS 2020 (Virtual Event, Dec. 6–12, 2020). NeurIPS.cc, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (visited on 03/28/2023) (cit. on pp. 88, 90, 93, 94, 97).

[BMS+16] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. "UQV100: A Test Collection with Query Variability." In: 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2016 (Pisa, Italy, July 17–21, 2016). Association for Computing Machinery, 2016, pp. 725–728. DOI: 10.1145/2911451.2914671 (cit. on p. 100).

[BMS+19] Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R. Goodwin, Sonya E. Shooshan, and Dina Demner-Fushman. "Bridging the Gap

Between Consumers' Medication Questions and Trusted Answers." In: *Health and Wellbeing e-Networks for All*. 17th World Congress on Medical and Health Informatics. MEDINFO 2019 (Lyon, France, Aug. 25–30, 2019). Vol. 264. Studies in Health Technology and Informatics. IOS Press, 2019, pp. 25–29. DOI: 10.3233/SHTI190176 (cit. on pp. 3, 10, 22).

[BPC20] Iz Beltagy, Matthew E. Peters, and Arman Cohan. "Longformer: The Long-Document Transformer." In: *CoRR* abs/2004.05150 (2020). ISSN: 2331-8422. DOI: 10.48550/arXiv.2004.05150 (cit. on p. 15).

[BPN+14] Georgios Balikas, Ioannis Partalas, Axel-Cyrille Ngonga Ngomo, Anastasia Krithara, and Georgios Paliouras. "Results of the BioASQ Track of the Question Answering Lab at CLEF 2014." In: *Working Notes of CLEF 2014*. 5th International Conference of the CLEF Initiative. CLEF 2014 (Sheffield, United Kingdom, Aug. 15–18, 2014). Vol. 1180. CEUR Workshop Proceedings. CEUR-WS.org, 2014, pp. 1181–1193. URL: https://ceur-ws.org/Vol-1180/CLEF2014wn-QA-BalikasEt2014.pdf (visited on 03/28/2023) (cit. on pp. 10, 21).

[Bre01] Leo Breiman. "Random Forests." In: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324 (cit. on pp. 4, 48).

[Bro02] Andrei Z. Broder. "A Taxonomy of Web Search." In: *SIGIR Forum* 36.2 (2002), pp. 3–10. ISSN: 0163-5840. DOI: 10.1145/792550.792552 (cit. on pp. 1, 99).

[BSD+20] Rakesh Bal, Sayan Sinha, Swastika Dutta, Rishabh Joshi, Sayan Ghosh, and Ritam Dutt. "Analysing the Extent of Misinformation in Cancer Related Tweets." In: 14th International AAAI Conference on Web and Social Media. ICWSM 2020 (Virtual Event, June 8–11, 2020). Association for the Advancement of Artificial Intelligence, 2020, pp. 924–928. DOI: 10.1609/icwsm.v14i1.7359 (cit. on p. 3).

[BSD+21] Alexander Bondarenko, Ekaterina Shirshakova, Marina Driker, Matthias Hagen, and Pavel Braslavski. "Misbeliefs and Biases in Health-Related Searches." In: 30th ACM International Conference on Information and Knowledge Management. CIKM 2021 (Virtual Event, Nov. 1–5, 2021). Association for Computing Machinery, 2021, pp. 2894–2899. DOI: 10.1145/3459637.3482141 (cit. on pp. 1, 2, 4, 5, 10, 17, 24, 76, 78).

[BSD19] Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. "Overview of the MEDIQA 2019 Shared Task on Textual Inference, Question Entailment and Question Answering." In: 18th BioNLP Workshop and Shared Task co-located with the 57th Conference of the Association for Computational Linguistics. BioNLP@ACL 2019 (Florence, Italy, Aug. 1–1, 2019). Association for Computational Linguistics, 2019, pp. 370–379. DOI: 10.18653/v1/w19-5039 (cit. on pp. 10, 20, 21).

[BTD+12] Michael S. Bernstein, Jaime Teevan, Susan T. Dumais, Daniel J. Liebling, and Eric Horvitz. "Direct Answers for Search Queries in the Long Tail." In: ACM CHI Conference on Human Factors in Computing Systems. CHI 2012 (Austin, Texas, United States, May 5–10, 2012). Association for Computing Machinery, 2012, pp. 237–246. DOI: 10.1145/2207676.2207710 (cit. on p. 11).

[BTS12] Sanmitra Bhattacharya, Hung Tran, and Padmini Srinivasan. "Discovering Health Beliefs in Twitter." In: *Information Retrieval and Knowledge Discovery in Biomedical Text.* AAAI 2012 Fall Symposium (Arlington, Virginia, United States, Nov. 2–4, 2012). Vol. FS-12-05. AAAI Technical Report. Association for the Advancement of Artificial Intelligence, 2012. URL: https://homepage.cs.uiowa.edu/~psriniva/Papers/aaai2012-camera-ready.pdf (visited on 03/18/2023) (cit. on p. 17).

[BWH+22] Alexander Bondarenko, Magdalena Wolska, Stefan Heindorf, Lukas Blübaum, Axel-Cyrille Ngonga Ngomo, Benno Stein, Pavel Braslavski, Matthias Hagen, and Martin Potthast. "CausalQA: A Benchmark for Causal Question Answering." In: 29th International Conference on Computational Linguistics. COLING 2022 (Gyeongju, Republic of Korea, Oct. 12–17, 2022). International Committee on Computational Linguistics, 2022, pp. 3296–3308. URL: https://aclanthology.org/2022.coling-1.291 (visited on 03/28/2023) (cit. on p. 99).

[BWS+03] Laurence Baker, Todd H. Wagner, Sara Singer, and M. Kate Bundorf. "Use of the Internet and E-mail for Health Care Information: Results From a National Survey." In: *Journal of the American Medical Association* 289.18 (May 14, 2003), pp. 2400–2406. ISSN: 1538-3598. DOI: 10.1001/jama.289.18.2400 (cit. on pp. 1, 3).

[BZE22] Markus Bink, Steven Zimmerman, and David Elsweiler. "Featured Snippets and their Influence on Users' Credibility Judgements." In: ACM SIGIR Conference on Human Information Interaction and Retrieval. CHIIR 2022 (Regensburg, Germany, Mar. 14–18, 2022). Associ-

ation for Computing Machinery, 2022, pp. 113–122. DOI: 10.1145/3498366.3505766 (cit. on pp. 5, 98).

[CCM+20] Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. "ORCAS: 20 Million Clicked Query-Document Pairs for Analyzing Search." In: 29th ACM International Conference on Information and Knowledge Management. CIKM 2020 (Virtual Event, Oct. 19–23, 2020). Association for Computing Machinery, 2020, pp. 2983–2989. DOI: 10.1145/3340531.3412779 (cit. on p. 12).

[CG16] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System." In: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD 2016 (San Francisco, California, United States, Aug. 13–17, 2016). Association for Computing Machinery, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785 (cit. on pp. 4, 48, 50).

[CHL+22] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. "Scaling Instruction-Finetuned Language Models." In: *CoRR* abs/2210.11416 (2022). ISSN: 2331-8422. DOI: 10.48550/arXiv.2210.11416 (cit. on p. 48).

[CL11] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: A Library for Support Vector Machines." In: *ACM Transactions on Intelligent Systems and Technology* 2.3, 27 (2011), 27:1–27:27. ISSN: 2157-6912. DOI: 10.1145/1961189.1961199 (cit. on pp. 4, 48, 96).

[CLC+19] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. "BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions." In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. NAACL-HLT 2019 (Minneapolis, Minnesota, United States, June 2–7, 2019). Vol. 1, Long and Short Papers. Association for Computational Linguistics, 2019, pp. 2924–2936. DOI: 10.18653/v1/n19-1300 (cit. on p. 17).

[CMS+22] Charles L. A. Clarke, Maria Maistro, Mahsa Seifikar, and Mark D. Smucker. "Overview of the TREC 2022 Health Misinformation Track." In: 30th REtrieval Conference. TREC 2021 (Gaithersburg, Maryland, United States, Nov. 15–19, 2021). Vol. 500-338. NIST Special Publica-

tion. To appear. National Institute of Standards and Technology, 2022 (cit. on pp. 2, 10, 16, 17, 23, 67, 75, 76, 78, 85–88, 90, 91, 93, 97, 99).

[CMS21] Charles L. A. Clarke, Maria Maistro, and Mark D. Smucker. "Overview of the TREC 2021 Health Misinformation Track." In: 30th REtrieval Conference. TREC 2021 (Gaithersburg, Maryland, United States, Nov. 15–19, 2021). Vol. 500-335. NIST Special Publication. National Institute of Standards and Technology, 2021. URL: https://trec.nist.gov/pubs/trec30/papers/Overview-HM.pdf (visited on 09/30/2022) (cit. on pp. 2, 10, 16, 17, 23, 67, 75, 76, 78, 100).

[CMW14] Munmun De Choudhury, Meredith Ringel Morris, and Ryen W. White. "Seeking and Sharing Health Information Online: Comparing Search Engines and Social Media." In: ACM CHI Conference on Human Factors in Computing Systems. CHI 2014 (Toronto, Ontario, Canada, Apr. 26–May 1, 2014). Association for Computing Machinery, 2014, pp. 1365–1376. DOI: 10.1145/2556288.2557214 (cit. on pp. 1, 3).

[CMY+21] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Ellen M. Voorhees, and Ian Soboroff. "TREC Deep Learning Track: Reusable Test Collections in the Large Data Regime." In: 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2021 (Virtual Event, July 11–15, 2021). Association for Computing Machinery, 2021, pp. 2369–2375. DOI: 10.1145/3404835.3463249 (cit. on p. 25).

[Coh60] Jacob Cohen. "A Coefficient of Agreement for Nominal Scales." In: *Educational and Psychological Measurement* 20.1 (Apr. 1960), pp. 37–46. ISSN: 1552-3888. DOI: 10.1177/001316446002000104 (cit. on p. 40).

[CRS+20] Charles L. A. Clarke, Saira Rizvi, Mark D. Smucker, Maria Maistro, and Guido Zuccon. "Overview of the TREC 2020 Health Misinformation Track." In: 29th Text REtrieval Conference. TREC 2020 (Virtual Event, Nov. 16–20, 2020). Vol. 1266. NIST Special Publication. National Institute of Standards and Technology, 2020. URL: https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.HM.pdf (visited on 09/30/2022) (cit. on pp. 1, 2, 6, 10, 16, 17, 23, 67, 75–78, 82).

[CSV20] Charles L. A. Clarke, Mark D. Smucker, and Alexandra Vtyurina. "Offline Evaluation by Maximum Similarity to an Ideal Ranking." In: 29th ACM International Conference on Information and Knowledge Management. CIKM 2020 (Virtual Event, Oct. 19–23, 2020). Association for Computing Machinery, 2020, pp. 225–234. DOI: 10.1145/3340531.3411915 (cit. on pp. 7, 16, 75, 82, 85, 90, 93).

[CTS+21] Berkant Barla Cambazoglu, Leila Tavakoli, Falk Scholer, Mark Sanderson, and W. Bruce Croft. "An Intent Taxonomy for Questions Asked in Web Search." In: ACM SIGIR Conference on Human Information Interaction and Retrieval. CHIIR 2021 (Canberra, Australia, Mar. 14–19, 2021). Association for Computing Machinery, 2021, pp. 85–94. DOI: 10.1145/3406522.3446027 (cit. on pp. 1, 4, 99).

[CVS20] Charles L. A. Clarke, Alexandra Vtyurina, and Mark D. Smucker. "Offline Evaluation without Gain." In: 10th ACM SIGIR International Conference on the Theory of Information Retrieval. ICTIR 2020 (Virtual Event, Aug. 14–17, 2020). Association for Computing Machinery, 2020, pp. 185–192. DOI: 10.1145/3409256.3409816 (cit. on pp. 7, 16, 75, 82, 85, 90, 93).

[CWH11] Marc-Allen Cartright, Ryen W. White, and Eric Horvitz. "Intentions and Attention in Exploratory Health Search." In: 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2011 (Beijing, China, July 25–29, 2011). Association for Computing Machinery, 2011, pp. 65–74. DOI: 10.1145/2009916.2009929 (cit. on p. 4).

[DCL+19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. NAACL-HLT 2019 (Minneapolis, Minnesota, United States, June 2–7, 2019). Vol. 1, Long and Short Papers. Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/n19-1423 (cit. on pp. 4, 11, 48, 52, 54).

[DDH07] Doug Downey, Susan T. Dumais, and Eric Horvitz. "Heads and Tails: Studies of Web Search With Common and Rare Queries." In: 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2007 (Amsterdam, The Netherlands, July 23–27, 2007). Association for Computing Machinery, 2007, pp. 847–848. DOI: 10.1145/1277741.1277939 (cit. on p. 11).

[DOM+21] Kevin Dadaczynski, Orkan Okan, Melanie Messer, Angela Yee Man Leung, Rafaela Rosário, Emily Joan Darlington, and Katharina Rathmann. "Digital Health Literacy and Web Information-Seeking Behaviors of University Students in Germany During the COVID-19 Pandemic: Cross-sectional Survey Study." In: *Journal of Medical Internet Research* 23.1 (Jan. 15, 2021). ISSN: 1438-8871. DOI: 10.2196/24097 (cit. on p. 5).

[DSM+21] Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, and Matt Gardner. "Documenting the English Colossal Clean Crawled Corpus." In: *CoRR* abs/2104.08758 (2021). ISSN: 2331-8422. DOI: 10.48550/arXiv.2104.08758 (cit. on pp. 72, 83, 93).

[DSW20] Enyan Dai, Yiwei Sun, and Suhang Wang. "Ginger Cannot Cure Cancer: Battling Fake Health News with a Comprehensive Data Repository." In: 14th International AAAI Conference on Web and Social Media. ICWSM 2020 (Virtual Event, June 8–11, 2020). Association for the Advancement of Artificial Intelligence, 2020, pp. 853–862. DOI: 10.1609/icwsm.v14i1.7350 (cit. on p. 15).

[DXC+18] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. "Convolutional Neural Networks for Soft-Matching N-Grams in Ad-hoc Search." In: 11th ACM International Conference on Web Search and Data Mining. WSDM 2018 (Marina del Rey, California, United States, Feb. 5–9, 2018). Association for Computing Machinery, 2018, pp. 126–134. DOI: 10.1145/3159652.3159659 (cit. on p. 13).

[EK03] Gunther Eysenbach and Christian Köhler. "What is the Prevalence of Health-Related Searches on the World Wide Web? Qualitative and Quantitative Analysis of Search Engine Queries on the Internet." In: American Medical Informatics Association Annual Symposium. AMIA 2003 (Washington, D.C., United States, Nov. 8–12, 2003). American Medical Informatics Association, 2003, pp. 225–229. PMCID: PMC1480194 (cit. on pp. 1, 3, 4, 9, 10, 66).

[Eys04] Gunther Eysenbach. "Health-Related Searches on the Internet." In: *Journal of the American Medical Association* 291.24 (June 23, 2004), pp. 2946–2946. ISSN: 1538-3598. DOI: 10.1001/jama.291.24.2946 (cit. on pp. 1, 3, 9, 11, 19).

[FAP+22] Maik Fröbe, Christopher Akiki, Martin Potthast, and Matthias Hagen. "How Train-Test Leakage Affects Zero-Shot Retrieval." In: 29th International Symposium on String Processing and Information Retrieval. SPIRE 2022 (Concepción, Chile, Nov. 8–10, 2022). Vol. 13617. Lecture Notes in Computer Science. Springer, 2022, pp. 147–161. DOI: 10.1007/978-3-031-20643-6_11 (cit. on pp. 90, 99).

[FBG+19] Lila J. Finney Rutten, Kelly D. Blake, Alexandra J. Greenberg-Worisek, Summer V. Allen, Richard P. Moser, and Bradford W. Hesse. "Online Health Information Seeking Among US Adults: Measuring Progress Toward a Healthy People 2020 Objective." In: *Public Health Reports*

134.6 (Nov. 12, 2019), pp. 617–625. ISSN: 1468-2877. DOI: 10.1177/0033354919874074 (cit. on p. 3).

[FBP+20] Maik Fröbe, Jan Philipp Bittner, Martin Potthast, and Matthias Hagen. "The Effect of Content-Equivalent Near-Duplicates on the Evaluation of Search Engines." In: *Advances in Information Retrieval*. 42nd European Conference on IR Research. ECIR 2020 (Lisbon, Portugal, Apr. 14–17, 2020). Vol. 12036.2. Lecture Notes in Computer Science. Springer, 2020, pp. 12–19. DOI: 10.1007/978-3-030-45442-5_2 (cit. on p. 77).

[FBR+20] Maik Fröbe, Janek Bevendorff, Jan Heinrich Reimer, Martin Potthast, and Matthias Hagen. "Sampling Bias Due to Near-Duplicates in Learning to Rank." In: 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2020 (Virtual Event, July 25–30, 2020). Association for Computing Machinery, 2020, pp. 1997–2000. DOI: 10.1145/3397271.3401212 (cit. on p. 77).

[FGB+22] Maik Fröbe, Sebastian Günther, Alexander Bondarenko, Johannes Huck, and Matthias Hagen. "Using Keyqueries to Reduce Misinformation in Health-Related Search Results." In: 2nd Workshop Reducing Online Misinformation through Credible Information Retrieval co-located with the 44th European Conference on Information Retrieval. ROMCIR 2022 (Stavanger, Norway, Apr. 10–10, 2022). Vol. 3138. CEUR Workshop Proceedings. CEUR-WS.org, 2022, pp. 1–10. URL: https://ceur-ws.org/Vol-3138/paper1_jot.pdf (visited on 09/30/2022) (cit. on p. 16).

[FKH18] Stefan Falkner, Aaron Klein, and Frank Hutter. "BOHB: Robust and Efficient Hyperparameter Optimization at Scale." In: 35th International Conference on Machine Learning. ICML 2018 (Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018). Vol. 80. Proceedings of Machine Learning Research, 2018, pp. 1436–1445. URL: https://proceedings.mlr.press/v80/falkner18a.html (visited on 03/28/2023) (cit. on pp. 12, 47, 49, 52, 56).

[Fle71] Joseph L. Fleiss. "Measuring Nominal Scale Agreement Among Many Raters." In: *Psychological Bulletin* 76.5 (Nov. 1971), pp. 378–382. ISSN: 1939-1455. DOI: 10.1037/h0031619 (cit. on pp. 40, 41, 76).

[FLP22] Marcos Fernández-Pichel, David E. Losada, and Juan Carlos Pichel. "A Multistage Retrieval System for Health-Related Misinformation Detection." In: *Engineering Applications of Artificial Intelligence* 115, 105211 (2022), 105211:1–105211:17. ISSN: 1873-6769. DOI: 10.1016/j.engappai.2022.105211 (cit. on pp. 15, 17).

[GAU+22] Mandy Guo, Joshua Ainslie, David C. Uthus, Santiago Ontañón, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. "LongT5: Efficient Text-to-Text Transformer for Long Sequences." In: *Findings of the Association for Computational Linguistics*. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. NAACL-HLT 2022 (Seattle, Washington, United States, July 10–15, 2022). Association for Computational Linguistics, 2022, pp. 724–736. DOI: 10.18653/v1/2022.findings-naacl.55 (cit. on pp. 48, 57, 96).

[GBB+21] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. "The Pile: An 800GB Dataset of Diverse Text for Language Modeling." In: *CoRR* abs/2101.00027 (2021). ISSN: 2331-8422. DOI: 10.48550/arXiv.2101.00027 (cit. on p. 14).

[GBC16] Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. *Deep Learning*. Adaptive Computation and Machine Learning. MIT Press, Nov. 18, 2016. 800 pp. ISBN: 978-0-262-03561-3 (cit. on pp. 12, 55).

[GDL+22] Travis R. Goodwin, Dina Demner-Fushman, Kyle Lo, Lucy Lu Wang, Hoa T. Dang, and Ian M. Soboroff. "Automatic Question Answering for Multiple Stakeholders, the Epidemic Question Answering Dataset." In: *Scientific Data* 9.1, 432 (July 21, 2022), 432:1–432:11. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01533-w (cit. on pp. 10, 24).

[GO09] Jeffrey S. Gerber and Paul A. Offit. "Vaccines and Autism: A Tale of Shifting Hypotheses." In: *Clinical Infectious Diseases* 48.4 (Feb. 15, 2009), pp. 456–461. ISSN: 1537-6591. DOI: 10.1086/596476 (cit. on p. 71).

[GTC+22] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing." In: *ACM Transactions on Computing for Healthcare* 3.1, 2 (2022), 2:1–2:23. ISSN: 2637-8051. DOI: 10.1145/3458754 (cit. on pp. 11, 48, 53).

[HAS+22] Sebastian Hofstätter, Sophia Althammer, Mete Sertkan, and Allan Hanbury. "Establishing Strong Baselines For TripClick Health Retrieval." In: *Advances in Information Retrieval*. 44th European Conference on IR Research. ECIR 2022 (Stavanger, Norway, Apr. 10–14, 2022). Vol. 13186.2. Lecture Notes in Computer Science. Springer, 2022, pp. 144–152. DOI: 10.1007/978-3-030-99739-7_17 (cit. on p. 13).

[HSW+20] Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. "CauseNet: Towards a Causality Graph Extracted from the Web." In: 29th ACM International Conference on Information and Knowledge Management. CIKM 2020 (Virtual Event, Oct. 19–23, 2020). Association for Computing Machinery, 2020, pp. 3023–3030. DOI: 10.1145/3340531.3412763 (cit. on p. 11).

[HVG+16] Matthias Hagen, Michael Völske, Steve Göring, and Benno Stein. "Axiomatic Result Re-Ranking." In: 25th ACM International Conference on Information and Knowledge Management. CIKM 2016 (Indianapolis, Indiana, United States, Oct. 24–28, 2016). Association for Computing Machinery, 2016, pp. 721–730. DOI: 10.1145/2983323.2983704 (cit. on pp. 70, 71, 80).

[HWJ+21] James E. Harrison, Stefanie Weber, Robert Jakob, and Christopher G. Chute. "ICD-11: An International Classification of Diseases for the Twenty-first Century." In: *BMC Medical Informatics and Decision Making* 21, 206 (Supplement 6 Nov. 2021), 206:1–206:10. ISSN: 1472-6947. DOI: 10.1186/s12911-021-01534-6 (cit. on p. 36).

[JDL+19] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. "PubMedQA: A Dataset for Biomedical Research Question Answering." In: Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing. EMNLP-IJCNLP 2019 (Hong Kong, China, Nov. 3–7, 2019). Association for Computational Linguistics, 2019, pp. 2567–2577. DOI: 10.18653/v1/D19-1259 (cit. on pp. 10, 11, 13–15, 17, 20, 22, 23, 80).

[JGB+16] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomás Mikolov. "FastText.zip: Compressing Text Classification Models." In: *CoRR* abs/1612.03651 (2016). ISSN: 2331-8422. DOI: 10.48550/arXiv.1612.03651 (cit. on p. 28).

[JGB+17] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomás Mikolov. "Bag of Tricks for Efficient Text Classification." In: 15th Conference of the European Chapter of the Association for Computational Linguistics. EACL 2017 (Valencia, Spain, Apr. 3–7, 2017). Vol. 2, Short Papers. Association for Computational Linguistics, 2017, pp. 427–431. DOI: 10.18653/v1/e17-2068 (cit. on p. 28).

[JK02] Kalervo Järvelin and Jaana Kekäläinen. "Cumulated Gain-Based Evaluation of IR Techniques." In: *ACM Transactions on Information Systems* 20.4 (2002), pp. 422–446. ISSN: 1558-2868. DOI: 10.1145/582415.582418 (cit. on pp. 71, 75, 85, 90).

[Jon72] Karen Spärck Jones. "A Statistical Interpretation of Term Specificity and its Application in Retrieval." In: *Journal of Documentation* 28.1 (1972), pp. 11–21. ISSN: 0022-0418. DOI: 10.1108/00220410410560573 (cit. on p. 15).

[JPG+18] Skyler B. Johnson, Henry S. Park, Cary P. Gross, and James B. Yu. "Use of Alternative Medicine for Cancer and Its Impact on Survival." In: *Journal of the National Cancer Institute* 110 (1 Jan. 2018), pp. 121–124. ISSN: 1460-2105. DOI: 10.1093/jnci/djx145 (cit. on p. 1).

[JS06] Bernard J. Jansen and Amanda Spink. "How Are We Searching the World Wide Web? A Comparison of Nine Search Engine Transaction Logs." In: *Information Processing and Management* 42.1 (2006), pp. 248–263. ISSN: 1873-5371. DOI: 10.1016/j.ipm.2004.10.007 (cit. on pp. 1, 3).

[JYX+23] Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. "Biomedical Question Answering: A Survey of Approaches and Challenges." In: *ACM Computing Surveys* 55.2, 1 (2023), 35:1–35:36. ISSN: 1557-7341. DOI: 10.1145/3490238 (cit. on pp. 3, 4, 14, 100).

[KB15] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization." In: 3rd International Conference on Learning Representations. ICLR 2015 (San Diego, California, United States, May 7–9, 2015). 2015. DOI: 10.48550/arXiv.1412.6980 (cit. on pp. 52, 55, 56).

[KBM+18] Halil Kilicoglu, Asma Ben Abacha, Yassine Mrabet, Sonya E. Shooshan, Laritza Rodriguez, Kate Masterton, and Dina Demner-Fushman. "Semantic Annotation of Consumer Health Questions." In: *BMC Bioinformatics* 19.1, 34 (Dec. 2018), 34:1–34:28. ISSN: 1471-2105. DOI: 10.1186/s12859-018-2045-1 (cit. on pp. 11, 21).

[KCL+10] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. "A Side Effect Resource to Capture Phenotypic Effects of Drugs." In: *Molecular Systems Biology* 6, 343 (2010), 343:1–343:6. ISSN: 1744-4292. DOI: 10.1038/msb.2009.98 (cit. on pp. 35, 36).

[KM21] Diane M. Korngiebel and Sean D. Mooney. "Considering the Possibilities and Pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in Healthcare Delivery." In: *npj Digital Medicine* 4.1, 93 (June 3, 2021), 93:1–93:3. ISSN: 2398-6352. DOI: 10.1038/s41746-021-00464-x (cit. on pp. 90, 93, 97, 99).

[KMK+20] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. "UnifiedQA: Crossing Format Boundaries With a Single QA System." In: *Findings of the Association for Computational Linguistics*. Conference on Empirical Methods in Natural Language Processing. EMNLP 2020 (Virtual Event, Nov. 16–20, 2020). Association for Computational Linguistics, 2020, pp. 1896–1907. DOI: 10.18653/v1/2020.findings-emnlp.171 (cit. on pp. 14, 69, 70, 79–81, 91).

[KNK+21] Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. "GooAQ: Open Question Answering with Diverse Answer Types." In: *Findings of the Association for Computational Linguistics*. Conference on Empirical Methods in Natural Language Processing. EMNLP 2021 (Punta Cana, Dominican Republic, Nov. 16–20, 2021). Association for Computational Linguistics, 2021, pp. 421–433. DOI: 10.18653/v1/2021.findings-emnlp.38 (cit. on pp. 11, 26).

[KNP+16] Anastasia Krithara, Anastasios Nentidis, Georgios Paliouras, and Ioannis Kakadiaris. "Results of the 4th Edition of BioASQ Challenge." In: 4th BioASQ Workshop. BioASQ 2016 (Berlin, Germany, Aug. 12–12, 2016). Association for Computational Linguistics, 2016, pp. 1–7. DOI: 10.18653/v1/W16-3101 (cit. on pp. 10, 21).

[KPR+19] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. "Natural Questions: a Benchmark for Question Answering Research." In: *Transactions of the Association for Computational Linguistics* 7 (Aug. 1, 2019), pp. 452–466. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00276 (cit. on pp. 1, 26).

[KSF+22] Damian P. Kotevski, Robert I. Smee, Matthew Field, Yvonne N. Nemes, Kathryn Broadley, and Claire M. Vajdic. "Evaluation of an Automated Presidio Anonymisation Model for Unstructured Radiation Oncology Electronic Medical Records in an Australian Setting." In: *International Journal of Medical Informatics* 168, 104880 (2022), 104880:1–104880:8. ISSN: 1872-8243. DOI: 10.1016/j.ijmedinf.2022.104880 (cit. on p. 29).

[KZ20] Omar Khattab and Matei Zaharia. "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT." In: 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2020 (Virtual Event, July 25–30,

2020). Association for Computing Machinery, 2020, pp. 39–48. DOI: 10.1145/3397271.3401075 (cit. on p. 13).

[LAY11] Feifan Liu, Lamont D. Antieau, and Hong Yu. "Toward Automated Consumer Question Answering: Automatically Separating Consumer Questions from Professional Questions in the Healthcare Domain." In: *Journal of Biomedical Informatics* 44.6 (2011), pp. 1032–1038. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2011.08.008 (cit. on pp. 10, 66).

[LCZ+06] Minsuk Lee, James Cimino, Hai R. Zhu, Carl Sable, Vijay Shanker, John Ely, and Hong Yu. "Beyond Information Retrieval: Medical Question Answering." In: American Medical Informatics Association Annual Symposium. AMIA 2006 (Washington, D.C., United States, Nov. 11–15, 2006). American Medical Informatics Association, 2006, pp. 469–473. PMCID: PMC1839371 (cit. on p. 13).

[LDB+19] Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. "Inferring Which Medical Treatments Work from Reports of Clinical Trials." In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. NAACL-HLT 2019 (Minneapolis, Minnesota, United States, June 2–7, 2019). Vol. 1, Long and Short Papers. Association for Computational Linguistics, 2019, pp. 3705–3717. DOI: 10.18653/v1/n19-1371 (cit. on pp. 15, 80).

[LHE22] Stephanie Lin, Jacob Hilton, and Owain Evans. "TruthfulQA: Measuring How Models Mimic Human Falsehoods." In: 60th Annual Meeting of the Association for Computational Linguistics. ACL 2022 (Dublin, Ireland, May 22–27, 2022). Vol. 1, Long Papers. Association for Computational Linguistics, 2022, pp. 3214–3252. DOI: 10.18653/v1/2022.acl-long.229 (cit. on p. 99).

[LHW22] Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. "Can Large Language Models Reason About Medical Questions?" In: *CoRR* abs/2207.08143 (2022). ISSN: 2331-8422. DOI: 10.48550/arXiv.2207.08143 (cit. on p. 14).

[Lin22] Jimmy Lin. "Building a Culture of Reproducibility in Academic Research." In: *CoRR* abs/2212.13534 (2022). ISSN: 2331-8422. DOI: 10.48550/arXiv.2212.13534 (cit. on p. 67).

[LJD+17] Lisha Li, Kevin G. Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. "Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization." In: *Journal of Machine Learning Research* 18, 185 (2017), 185:1–185:52. ISSN: 1533-7928. URL: https:

//jmlr.org/papers/v18/16-558.html (visited on 03/28/2023) (cit. on pp. 12, 47).

[LK77] J. Richard Landis and Gary G. Koch. "The Measurement of Observer Agreement for Categorical Data." In: *Biometrics* 33.1 (Mar. 1977), pp. 159–174. ISSN: 0006-341X. DOI: 10.2307/2529310 (cit. on pp. 40, 41, 77).

[LLG+20] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." In: 58th Annual Meeting of the Association for Computational Linguistics. ACL 2020 (Virtual Event, July 5–10, 2020). Association for Computational Linguistics, 2020, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703 (cit. on pp. 48, 54).

[LOG+19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." In: *CoRR* abs/1907.11692 (2019). ISSN: 2331-8422. DOI: 10.48550/arXiv.1907.11692 (cit. on pp. 17, 48).

[LSC20] Andre Lamurias, Diana Sousa, and Francisco M. Couto. "Generating Biomedical Question Answering Corpora From Q&A Forums." In: *IEEE Access* 8 (2020), pp. 161042–161051. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.3020868 (cit. on pp. 10, 23).

[LSX+22] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. "BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining." In: *CoRR* abs/2210.10341 (2022). ISSN: 2331-8422. DOI: 10.48550/arXiv.2210.10341 (cit. on pp. 48, 57, 59).

[MCG20] Sean MacAvaney, Arman Cohan, and Nazli Goharian. "SLEDGE-Z: A Zero-Shot Baseline for COVID-19 Literature Search." In: Conference on Empirical Methods in Natural Language Processing. EMNLP 2020 (Virtual Event, Nov. 16–20, 2020). Association for Computational Linguistics, 2020, pp. 4171–4179. DOI: 10.18653/v1/2020.emnlp-main.341 (cit. on p. 13).

[MKC+22] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. "Deep Learning–Based Text Classification: A Comprehensive Review." In: *ACM Computing Surveys* 54.3, 62 (2022), 62:1–62:40. ISSN: 1557-7341. DOI: 10.1145/3439726 (cit. on p. 1).

[MMO22] Sean MacAvaney, Craig Macdonald, and Iadh Ounis. "Streamlining Evaluation with ir-measures." In: *Advances in Information Retrieval*. 44th European Conference on IR Research. ECIR 2022 (Stavanger, Norway, Apr. 10–14, 2022). Vol. 13186.2. Lecture Notes in Computer Science. Springer, 2022, pp. 305–310. DOI: 10.1007/978-3-030-99739-7_38 (cit. on pp. 14, 68).

[MRJ+20] Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. "COVID-QA: A Question Answering Dataset for COVID-19." In: 1st Workshop on NLP for COVID-19 co-located with the 58th Annual Meeting of the Association for Computational Linguistics. NLP-COVID19 (Virtual Event, July 9–10, 2020). Association for Computational Linguistics, 2020. URL: https://aclanthology.org/2020.nlpcovid19-acl.18 (visited on 03/28/2023) (cit. on pp. 10, 24).

[MSM93] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. "Building a Large Annotated Corpus of English: The Penn Treebank." In: *Computational Linguistics* 19.2 (1993), pp. 313–330. ISSN: 0891-2017. URL: https://dl.acm.org/doi/10.5555/972470.972475 (visited on 03/28/2023) (cit. on p. 23).

[MTM+21] Craig Macdonald, Nicola Tonellotto, Sean MacAvaney, and Iadh Ounis. "PyTerrier: Declarative Experimentation in Python from BM25 to Dense Retrieval." In: 30th ACM International Conference on Information and Knowledge Management. CIKM 2021 (Virtual Event, Nov. 1–5, 2021). Association for Computing Machinery, 2021, pp. 4526–4533. DOI: 10.1145/3459637.3482013 (cit. on pp. 14, 67, 68).

[MYF+21] Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. "Simplified Data Wrangling with ir_datasets." In: 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2021 (Virtual Event, July 11–15, 2021). Association for Computing Machinery, 2021, pp. 2429–2436. DOI: 10.1145/3404835.3463254 (cit. on pp. 14, 26).

[NBK+17] Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, Georgios Paliouras, and Ioannis A. Kakadiaris. "Results of the Fifth Edition of the BioASQ Challenge." In: 13th Workshop on Biomedical Natural Language Processing. BioNLP 2017 (Vancouver, British Columbia, Canada, Aug. 4–4, 2017). Association for Computational Linguistics, 2017, pp. 48–57. DOI: 10.18653/v1/W17-2306 (cit. on pp. 10, 21).

[NBK+19] Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. "Results of the Seventh Edition of the BioASQ Challenge." In: *Machine Learning and Knowledge Discovery in Databases*. International Workshops of the 19th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. ECML PKDD 2019 (Würzburg, Germany, Aug. 16–20, 2019). Vol. 1168.2. Communications in Computer and Information Science. Springer, 2019, pp. 553–568. DOI: 10.1007/978-3-030-43887-6_51 (cit. on pp. 10, 21).

[ND22] Devesh Narayanan and David De Cremer. "'Google Told Me So!' On the Bent Testimony of Search Engine Algorithms." In: *Philosophy & Technology* 35.2, 22 (Mar. 26, 2022), 22:1–22:19. ISSN: 2210-5441. DOI: 10.1007/s13347-022-00521-7 (cit. on pp. 1, 5).

[NKB+18] Anastasios Nentidis, Anastasia Krithara, Konstantinos Bougiatiotis, Georgios Paliouras, and Ioannis Kakadiaris. "Results of the Sixth Edition of the BioASQ Challenge." In: *A Challenge on Large-scale Biomedical Semantic Indexing and Question Answering*. 6th BioASQ Workshop. BioASQ 2018 (Brussels, Belgium, Nov. 1–1, 2018). Association for Computational Linguistics, 2018, pp. 1–10. DOI: 10.18653/v1/W18-5301 (cit. on pp. 10, 21).

[NKB+20] Anastasios Nentidis, Anastasia Krithara, Konstantinos Bougiatiotis, and Georgios Paliouras. "Overview of BioASQ 8a and 8b: Results of the Eighth Edition of the BioASQ Tasks a and b." In: *Working Notes of CLEF 2020*. 11th International Conference of the CLEF Association. CLEF 2020 (Thessaloniki, Greece, Aug. 22–25, 2020). Vol. 2696. CEUR Workshop Proceedings. CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2696/paper_164.pdf (visited on 03/28/2023) (cit. on pp. 10, 21).

[NKV+21] Anastasios Nentidis, Georgios Katsimpras, Eirini Vandorou, Anastasia Krithara, and Georgios Paliouras. "Overview of BioASQ Tasks 9a, 9b and Synergy in CLEF 2021." In: *Working Notes of CLEF 2021*. 12th International Conference of the CLEF Association. CLEF 2021 (Virtual Event, Sept. 21–24, 2021). Vol. 2936. CEUR Workshop Proceedings. CEUR-WS.org, 2021, pp. 157–164. URL: https://ceur-ws.org/Vol-2936/paper-10.pdf (visited on 03/28/2023) (cit. on pp. 10, 21).

[NRS+16] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset." In: Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches co-

located with the 30th Annual Conference on Neural Information Processing Systems. CoCo@NIPS 2016 (Barcelona, Spain, Dec. 9–9, 2016). Vol. 1773. CEUR Workshop Proceedings. CEUR-WS.org, 2016. URL: https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf (visited on 03/28/2023) (cit. on pp. 11, 13, 25, 42, 69, 72, 76).

[OYW12] Sanghee Oh, Yong Jeong Yi, and Adam Worrall. "Quality of Health Answers in Social Q&A." In: *Information, Interaction, Innovation: Celebrating the Past, Constructing the Present and Creating the Future*. 75th ASIS&T Annual Meeting. ASIST 2012 (Baltimore, Maryland, United States, Oct. 26–30, 2012). Vol. 49. 1. Wiley, 2012, pp. 1–6. DOI: 10.1002/meet.14504901075 (cit. on p. 5).

[PAB+23] Alexander Pugachev, Ekaterina Artemova, Alexander Bondarenko, and Pavel Braslavski. "Consumer Health Question Answering Using Off-the-shelf Components." In: *Advances in Information Retrieval*. 45th European Conference on IR Research. ECIR 2023 (Dublin, Ireland, Apr. 2–6, 2023). Lecture Notes in Computer Science. Springer, Mar. 17, 2023, pp. 571–579. DOI: 10.1007/978-3-031-28238-6_48 (cit. on pp. 17, 87, 88, 90).

[PAT+21] Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. "SciFive: A Text-to-Text Transformer Model for Biomedical Literature." In: *CoRR* abs/2106.03598 (2021). ISSN: 2331-8422. DOI: 10.48550/arXiv.2106.03598 (cit. on pp. 48, 57).

[PBK22] Piotr Przybyla, Piotr Borkowski, and Konrad Kaczynski. "Countering Disinformation by Finding Reliable Sources: A Citation-Based Approach." In: International Joint Conference on Neural Networks. IJCNN 2022 (Padua, Italy, July 18–23, 2022). Institute of Electrical and Electronics Engineers, 2022, pp. 1–8. DOI: 10.1109/IJCNN55064.2022.9891941 (cit. on p. 16).

[PBW07] Nina Phan, Peter Bailey, and Ross Wilkinson. "Understanding the Relationship of Information Need Specificity to Search Query Length." In: 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2007 (Amsterdam, The Netherlands, July 23–27, 2007). Association for Computing Machinery, 2007, pp. 709–710. DOI: 10.1145/1277741.1277870 (cit. on pp. 28, 43).

[PGS+17] Frances A. Pogacar, Amira Ghenai, Mark D. Smucker, and Charles L. A. Clarke. "The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments."

In: ACM SIGIR International Conference on Theory of Information Retrieval. ICTIR 2017 (Amsterdam, The Netherlands, Oct. 1–4, 2017). Association for Computing Machinery, 2017, pp. 209–216. DOI: 10.1145/3121050.3121074 (cit. on pp. 5, 6).

[PHF+13] Anselmo Peñas, Eduard H. Hovy, Pamela Forner, Álvaro Rodrigo, Richard F. E. Sutcliffe, and Roser Morante. "QA4MRE 2011-2013: Overview of Question Answering for Machine Reading Evaluation." In: *Information Access Evaluation. Multilinguality, Multimodality, and Visualization.* 4th International Conference of the CLEF Initiative. CLEF 2013 (Valencia, Spain, Aug. 23–26, 2013). Vol. 8138. Lecture Notes in Computer Science. Springer, 2013, pp. 303–320. DOI: 10.1007/978-3-642-40802-1_29 (cit. on pp. 10, 20).

[PHM+16] João R. M. Palotti, Allan Hanbury, Henning Müller, and Charles E. Kahn Jr. "How Users Search and What They Search for in the Medical Domain." In: *Information Retrieval Journal* 19.1-2 (2016), pp. 189–224. ISSN: 1573-7659. DOI: 10.1007/s10791-015-9269-8 (cit. on p. 3).

[PHS20] Martin Potthast, Matthias Hagen, and Benno Stein. "The Dilemma of the Direct Answer." In: *SIGIR Forum* 54.1, 14 (2020), 14:1–14:12. ISSN: 0163-5840. DOI: 10.1145/3451964.3451978 (cit. on p. 5).

[PMM+17] Flaminia Pantano, Giulio Mannocchi, Enrico Marinelli, Sara Gentili, Silvia Graziano, Francesco Paolo Busardò, and Natale Mario di Luca. "Hepatotoxicity Induced by Greater Celandine (Chelidonium majus L.): A Review of the Literature." In: *European Review for Medical and Pharmacological Sciences* 21.1 Supplementary (Mar. 2017), pp. 46–52. ISSN: 2284-0729. URL: https://europeanreview.org/article/12431 (visited on 04/05/2023) (cit. on p. 1).

[PMN+21] Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. "Vera: Prediction Techniques for Reducing Harmful Misinformation in Consumer Health Search." In: 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2021 (Virtual Event, July 11–15, 2021). Association for Computing Machinery, 2021, pp. 2066–2070. DOI: 10.1145/3404835.3463120 (cit. on pp. 1, 2, 17, 70, 79, 83, 84, 88, 92–94, 97, 98).

[PNL21] Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. "The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models." In: *CoRR* abs/2101.05667 (2021). ISSN: 2331-8422. DOI: 10.48550/arXiv.2101.05667 (cit. on pp. 4, 13, 17, 54, 55, 57, 69, 72, 76, 79, 80, 82, 83, 91–93, 97).

[Por80] Martin F. Porter. "An Algorithm for Suffix Stripping." In: *Program* 14.3 (1980), pp. 130–137. ISSN: 0033-0337. DOI: 10.1108/eb046814 (cit. on pp. 27, 36, 43).

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. "scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830. ISSN: 1533-7928. URL: https://dl.acm.org/doi/10.5555/1953048.2078195 (visited on 04/06/2023) (cit. on pp. 48, 50).

[PWD+12] Daryl Posnett, Eric Warburg, Premkumar T. Devanbu, and Vladimir Filkov. "Mining Stack Exchange: Expertise Is Evident from Initial Contributions." In: International Conference on Social Informatics. SocialInformatics 2012 (Washington, D.C., United States, Dec. 14–16, 2012). IEEE Computer Society, 2012, pp. 199–204. DOI: 10.1109/SocialInformatics.2012.67 (cit. on p. 11).

[RBE+17] Alexander Ratner, Stephen H. Bach, Henry R. Ehrenberg, Jason Alan Fries, Sen Wu, and Christopher Ré. "Snorkel: Rapid Training Data Creation with Weak Supervision." In: *Proceedings of the VLDB Endowment* 11.3 (2017), pp. 269–282. ISSN: 2150-8097. DOI: 10.14778/3157794.3157797 (cit. on pp. 12, 33, 36–38).

[RG19] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In: Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing. EMNLP-IJCNLP 2019 (Hong Kong, China, Nov. 3–7, 2019). Association for Computational Linguistics, 2019, pp. 3980–3990. DOI: 10.18653/v1/D19-1410 (cit. on pp. 16, 48).

[RHD+19] Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. "Training Complex Models with Multi-Task Weak Supervision." In: 33rd AAAI Conference on Artificial Intelligence. AAAI 2019 (Honolulu, Hawaii, United States, Jan. 27–Feb. 1, 2019). Association for the Advancement of Artificial Intelligence, 2019, pp. 4763–4771. DOI: 10.1609/aaai.v33i01.33014763 (cit. on pp. 12, 33, 95).

[Rij79] C. J. van Rijsbergen. *Information Retrieval*. 2nd ed. Butterworth, 1979. 208 pp. ISBN: 978-0-408-70929-3 (cit. on pp. 7, 10, 44, 49).

[RLS+21] Navid Rekabsaz, Oleg Lesota, Markus Schedl, Jon Brassey, and Carsten Eickhoff. "TripClick: The Log Files of a Large Health Web Search Engine." In: 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2021 (Virtual Event, July 11–15, 2021). Association for Computing Machinery, 2021, pp. 2507–2513. DOI: 10.1145/3404835.3463242 (cit. on pp. 13, 98).

[RSR+20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." In: *Journal of Machine Learning Research* 21, 140 (2020), 140:1–140:67. ISSN: 1533-7928. URL: https://jmlr.org/papers/v21/20-074.html (visited on 03/28/2023) (cit. on pp. 11, 16, 17, 48, 55, 72, 80).

[RWC+19] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language Models are Unsupervised Multitask Learners." In: OpenAI, 2019. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (visited on 03/28/2023) (cit. on pp. 11, 48, 56).

[RWJ+94] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. "Okapi at TREC-3." In: 3rd Text REtrieval Conference. TREC 1994 (Gaithersburg, Maryland, United States, Nov. 2–4, 1994). Vol. 500-225. NIST Special Publication. National Institute of Standards and Technology, 1994, pp. 109–126. URL: https://trec.nist.gov/pubs/trec3/papers/city.ps.gz (visited on 03/28/2023) (cit. on pp. 13, 17, 69, 72, 76, 92, 93, 97).

[RZL+16] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. "SQuAD: 100,000+ Questions for Machine Comprehension of Text." In: Conference on Empirical Methods in Natural Language Processing. EMNLP 2016 (Austin, Texas, United States, Nov. 1–4, 2016). Association for Computational Linguistics, 2016, pp. 2383–2392. DOI: 10.18653/v1/d16-1264 (cit. on pp. 11, 25).

[SAT+22] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher

Semturs, Alan Karthikesalingam, and Vivek Natarajan. "Large Language Models Encode Clinical Knowledge." In: *CoRR* abs/2212.13138 (2022). ISSN: 2331-8422. DOI: 10.48550/arXiv.2212.13138 (cit. on pp. 14, 85, 99).

[SBH+22] Ferdinand Schlatt, Dieter Bettin, Matthias Hagen, Benno Stein, and Martin Potthast. "Mining Health-Related Cause-Effect Statements with High Precision at Large Scale." In: 29th International Conference on Computational Linguistics. COLING 2022 (Gyeongju, Republic of Korea, Oct. 12–17, 2022). International Committee on Computational Linguistics, 2022, pp. 1925–1936. URL: https://aclanthology.org/2022.coling-1.167 (visited on 03/28/2023) (cit. on pp. 11, 66).

[SBM+21] Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. "Evidence-Based Fact-Checking of Health-Related Claims." In: *Findings of the Association for Computational Linguistics*. Conference on Empirical Methods in Natural Language Processing. EMNLP 2021 (Punta Cana, Dominican Republic, Nov. 16–20, 2021). Association for Computational Linguistics, 2021, pp. 3499–3512. DOI: 10.18653/v1/2021.findings-emnlp.297 (cit. on p. 15).

[SCF08] Burr Settles, Mark Craven, and Lewis Friedland. "Active Learning with Real Annotation Costs." In: Workshop on Cost-Sensitive Learning co-located with the 22th Annual Conference on Neural Information Processing Systems. NIPS-CSL 2008 (Vancouver, British Columbia, Canada, Dec. 13–13, 2008). NeurIPS.cc, 2008. URL: https://biostat.wisc.edu/~craven/papers/settles.nips08.pdf (visited on 04/06/2023) (cit. on p. 77).

[SCM21] Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. "COVID-Fact: Fact Extraction and Verification of Real-world Claims on COVID-19 Pandemic." In: 59th Annual Meeting of the Association for Computational Linguistics and 11th International Joint Conference on Natural Language Processing. ACL/IJCNLP 2021 (Virtual Event, Aug. 1–6, 2021). Vol. 1, Long Papers. Association for Computational Linguistics, 2021, pp. 2116–2129. DOI: 10.18653/v1/2021.acl-long.165 (cit. on p. 15).

[Sco92] David W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. Wiley, 1992. 317 pp. ISBN: 978-0-471-54770-9. DOI: 10.1002/9780470316849 (cit. on p. 89).

[SCZ08] Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. "Learning to Rank Answers on Large Online QA Collections." In: 46th Annual Meeting of the Association for Computational Linguistics. ACL 2008 (Columbus, Ohio, United States, June 15–20, 2008). Association for Computer Linguistics, 2008, pp. 719–727. URL: https://aclanthology.org/P08-1082 (visited on 09/30/2022) (cit. on pp. 11, 27).

[SPT+22] Ivan Srba, Branislav Pecher, Matús Tomlein, Róbert Móro, Elena Stefancova, Jakub Simko, and Mária Bieliková. "Monant Medical Misinformation Dataset: Mapping Articles to Fact-Checked Claims." In: 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2022 (Madrid, Spain, July 11–15, 2022). Association for Computing Machinery, 2022, pp. 2949–2959. DOI: 10.1145/3477495.3531726 (cit. on pp. 15, 77).

[SSK+22] Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. "One Embedder, Any Task: Instruction-Finetuned Text Embeddings." In: *CoRR* abs/2212.09741 (2022). ISSN: 2331-8422. DOI: 10.48550/arXiv.2212.09741 (cit. on p. 48).

[SSW+17] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. "Fake News Detection on Social Media: A Data Mining Perspective." In: *ACM SIGKDD Explorations Newsletter* 19.1 (2017), pp. 22–36. ISSN: 1931-0153. DOI: 10.1145/3137597.3137600 (cit. on p. 15).

[SWR+22] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, Maruf Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. "Multitask Prompted Training Enables Zero-Shot Task Generalization." In: 10th International Conference on Learning Representations. ICLR 2022 (Virtual Event, Apr. 25–29, 2022). OpenReview.net, 2022. URL: https://openreview.net/forum?id=9Vrb9D0WI4 (visited on 03/28/2023) (cit. on p. 99).

[SYJ+04] Amanda Spink, Yin Yang, Jim Jansen, Pirrko Nykanen, Daniel P. Lorence, Seda Ozmutlu, and H. Cenk Ozmutlu. "A Study of Medical

and Health Queries to Web Search Engines." In: *Health Information and Libraries Journal* 21.1 (Mar. 12, 2004), pp. 44–51. ISSN: 1471-1842. DOI: 10.1111/j.1471-1842.2004.00481.x (cit. on pp. 1, 3, 9, 11, 19).

[TBM+15] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. "An Overview of the BioASQ Large-scale Biomedical Semantic Indexing and Question Answering Competition." In: *BMC Bioinformatics* 16, 138 (2015), 138:1–138:28. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0564-6 (cit. on p. 21).

[TKC+22] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. "Galactica: A Large Language Model for Science." In: *CoRR* abs/2211.09085 (2022). ISSN: 2331-8422. DOI: 10.48550/arXiv.2211.09085 (cit. on pp. 14, 48).

[TSP+12] George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Éric Gaussier, Patrick Gallinari, Thierry Artières, Michael R. Alvers, Matthias Zschunke, and Axel-Cyrille Ngonga Ngomo. "BioASQ: A Challenge on Large-scale Biomedical Semantic Indexing and Question Answering." In: *Information Retrieval and Knowledge Discovery in Biomedical Text.* AAAI 2012 Fall Symposium (Arlington, Virginia, United States, Nov. 2–4, 2012). Vol. FS-12-05. AAAI Technical Report. Association for the Advancement of Artificial Intelligence, 2012. URL: https://bioasq.org/sites/default/files/PublicDocuments/2012-Tsatsaronis-BioASQ.pdf (visited on 03/28/2023) (cit. on pp. 17, 21).

[TVC+18] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. "FEVER: A Large-scale Dataset for Fact Extraction and VERification." In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. NAACL-HLT 2018 (New Orleans, Louisiana, United States, June 1–6, 2018). Vol. 1, Long Papers. Association for Computational Linguistics, 2018, pp. 809–819. DOI: 10.18653/v1/n18-1074 (cit. on pp. 15, 69, 79, 80).

[UFL+14] Christina Unger, Corina Forascu, Vanessa López, Axel-Cyrille Ngonga Ngomo, Elena Cabrio, Philipp Cimiano, and Sebastian Walter. "Question Answering over Linked Data (QALD-4)." In: *Working Notes for CLEF 2014 Conference*. 5th International Conference of the CLEF Initiative. CLEF 2014 (Sheffield, United Kingdom, Aug. 15–18, 2014). Vol. 1180. CEUR Workshop Proceedings. CEUR-WS.org, 2014, pp. 1172–1180. URL: https://ceur-ws.org/Vol-1180/CLEF2014wn-QA-UngerEt2014.pdf (visited on 09/30/2022) (cit. on pp. 10, 20).

[VAB+20] Ellen M. Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. "TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection." In: *SIGIR Forum* 54.1, 1 (2020), 1:1–1:12. ISSN: 0163-5840. DOI: 10.1145/3451964.3451965 (cit. on p. 13).

[VG19] David Vilares and Carlos Gómez-Rodríguez. "HEAD-QA: A Healthcare Dataset for Complex Reasoning." In: 57th Conference of the Association for Computational Linguistics. ACL 2019 (Florence, Italy, July 28–Aug. 2, 2019). Vol. 1, Long Papers. Association for Computational Linguistics, 2019, pp. 960–966. DOI: 10.18653/v1/p19-1092 (cit. on pp. 10, 20, 21).

[VR14] Andreas Vlachos and Sebastian Riedel. "Fact Checking: Task Definition and Dataset Construction." In: Workshop on Language Technologies and Computational Social Science co-located with the 52nd Annual Meeting of the Association for Computational Linguistics. LACSS 2014 (Baltimore, Maryland, United States, June 26–26, 2014). Association for Computational Linguistics, 2014, pp. 18–22. DOI: 10.3115/v1/W14-2508 (cit. on p. 15).

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is All you Need." In: *Advances in Neural Information Processing Systems*. 31st Annual Conference on Neural Information Processing Systems. NIPS 2017 (Long Beach, California, United States, Dec. 4–9, 2017). NeurIPS.cc, 2017, pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (visited on 03/28/2023) (cit. on pp. 11, 52, 55).

[WA14] Ryen W. White and Ahmed Hassan Awadallah. "Content Bias in Online Health Search." In: *ACM Transactions on the Web* 8.4, 25 (2014), 25:1–25:33. ISSN: 1559-114X. DOI: 10.1145/2663355 (cit. on pp. 2, 4, 98).

[WDS+20] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. "Transformers: State-of-the-Art Natural Language Processing." In: *System Demonstrations*. Conference on Empirical Methods in Natural Language Processing. EMNLP 2020 (Virtual Event, Nov. 16–20, 2020). Vol. Demos. Association for Computational Linguistics, 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6 (cit. on pp. 52, 53).

[WH15] Ryen W. White and Eric Horvitz. "Belief Dynamics and Biases in Web Search." In: *ACM Transactions on Information Systems* 33.4, 18 (2015), 18:1–18:46. ISSN: 1558-2868. DOI: 10.1145/2746229 (cit. on pp. 1, 5, 6, 98).

[Whi14] Ryen W. White. "Belief Dynamics in Web Search." In: *Journal of the Association for Information Science and Technology* 65.11 (2014), pp. 2165–2178. ISSN: 2330-1643. DOI: 10.1002/asi.23128 (cit. on pp. 5, 6).

[WKA+22] Somin Wadhwa, Vivek Khetan, Silvio Amir, and Byron C. Wallace. "RedHOT: A Corpus of Annotated Medical Questions, Experiences, and Claims on Social Media." In: *CoRR* abs/2210.06331 (2022). ISSN: 2331-8422. DOI: 10.48550/arXiv.2210.06331 (cit. on p. 99).

[WLC+20] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. "CORD-19: The Covid-19 Open Research Dataset." In: *CoRR* abs/2004.10706 (2020). ISSN: 2331-8422. DOI: 10.48550/arXiv.2004.10706 (cit. on p. 13).

[WLL+20] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. "Fact or Fiction: Verifying Scientific Claims." In: Conference on Empirical Methods in Natural Language Processing. EMNLP 2020 (Virtual Event, Nov. 16–20, 2020). Association for Computational Linguistics, 2020, pp. 7534–7550. DOI: 10.18653/v1/2020.emnlp-main.609 (cit. on p. 15).

[WLW+22] David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. "MultiVerS: Improving Scientific Claim Verification with Weak Supervision and Full-Document Context." In:

Findings of the Association for Computational Linguistics. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. NAACL-HLT 2022 (Seattle, Washington, United States, July 10–15, 2022). Association for Computational Linguistics, 2022, pp. 61–76. DOI: 10.18653/v1/2022.findings-naacl.6 (cit. on pp. 15, 70, 79–81, 88, 92).

[WMR+21] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. "Ethical and Social Risks of Harm from Language Models." In: *CoRR* abs/2112.04359 (2021). ISSN: 2331-8422. DOI: 10.48550/arXiv.2112.04359 (cit. on pp. 2, 90, 93, 97).

[WRY15] Ryen W. White, Matthew Richardson, and Wen-tau Yih. "Questions vs. Queries in Informational Search Tasks." In: 24th International Conference on World Wide Web. WWW 2015 Companion (Florence, Italy, May 18–22, 2015). Association for Computing Machinery, 2015, pp. 135–136. DOI: 10.1145/2740908.2742769 (cit. on pp. 3, 28, 43).

[YLL22] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. "LinkBERT: Pretraining Language Models with Document Links." In: 60th Annual Meeting of the Association for Computational Linguistics. ACL 2022 (Dublin, Ireland, May 22–27, 2022). Vol. 1, Long Papers. Association for Computational Linguistics, 2022, pp. 8003–8016. DOI: 10.18653/v1/2022.acl-long.551 (cit. on pp. 17, 48).

[YOA+19] Tolga Yilmaz, Rifat Ozcan, Ismail Sengor Altingovde, and Özgür Ulusoy. "Improving Educational Web Search for Question-like Queries Through Subject Classification." In: *Information Processing and Management* 56.1 (2019), pp. 228–246. ISSN: 1873-5371. DOI: 10.1016/j.ipm.2018.10.013 (cit. on p. 28).

[YZK+22] Zhenrui Yue, Huimin Zeng, Ziyi Kou, Lanyu Shang, and Dong Wang. "Contrastive Domain Adaptation for Early Misinformation Detection: A Case Study on COVID-19." In: 31st ACM International Conference on Information and Knowledge Management. CIKM 2022 (Atlanta, Georgia, United States, Oct. 17–21, 2022). Association for Computing Machinery, 2022, pp. 2423–2433. DOI: 10.1145/3511808.3557263 (cit. on p. 15).

[YZL+14] Yanshen Yin, Yong Zhang, Xiao Liu, Yan Zhang, Chunxiao Xing, and Hsinchun Chen. "HealthQA: A Chinese QA Summary System for Smart Health." In: Smart Health: International Conference. ICSH 2014 (Beijing, China, July 10–11, 2014). Vol. 8549. Lecture Notes in Computer Science. Springer, 2014, pp. 51–62. DOI: 10.1007/978-3-319-08416-9_6 (cit. on p. 98).

[Zha10] Yan Zhang. "Contextualizing Consumer Health Information Searching: An Analysis of Questions in a Social Q&A Community." In: ACM International Health Informatics Symposium. IHI 2010 (Arlington, Virginia, United States, Nov. 11–12, 2010). Association for Computing Machinery, 2010, pp. 210–219. DOI: 10.1145/1882992.1883023 (cit. on pp. 3, 4, 16).

[ZPH16] Guido Zuccon, João R. M. Palotti, and Allan Hanbury. "Query Variations and their Effect on Comparing Information Retrieval Systems." In: 25th ACM International Conference on Information and Knowledge Management. CIKM 2016 (Indianapolis, Indiana, United States, Oct. 24–28, 2016). Association for Computing Machinery, 2016, pp. 691–700. DOI: 10.1145/2983323.2983723 (cit. on p. 99).

[ZRC+22] George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. "CODER: An Efficient Framework for Improving Retrieval Through COntextual Document Embedding Reranking." In: Conference on Empirical Methods in Natural Language Processing. EMNLP 2022 (Abu Dhabi, United Arab Emirates, Dec. 7–11, 2022). Association for Computational Linguistics, 2022, pp. 10626–10644. URL: https://aclanthology.org/2022.emnlp-main.727 (visited on 03/28/2023) (cit. on p. 13).

[ZRG+22] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. "OPT: Open Pre-trained Transformer Language Models." In: *CoRR* abs/2205.01068 (2022). ISSN: 2331-8422. DOI: 10.48550/arXiv.2205.01068 (cit. on p. 48).

[ZTA+22] Dake Zhang, Amir Vakili Tahami, Mustafa Abualsaud, and Mark D. Smucker. "Learning Trustworthy Web Sources to Derive Correct Answers and Reduce Health Misinformation in Search." In: 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2022 (Madrid, Spain, July 11–15, 2022). Association for Computing Machinery, 2022, pp. 2099–2104. DOI: 10.1145/3477495.3531812 (cit. on pp. 1, 17, 79, 83, 84).

[ZZL15] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. "Character-level Convolutional Networks for Text Classification." In: *Advances in Neural Information Processing Systems*. 29th Annual Conference on Neural Information Processing Systems. NIPS 2015 (Montreal, Quebec, Canada, Dec. 7–12, 2015). NeurIPS.cc, 2015, pp. 649–657. URL: https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html (visited on 09/30/2022) (cit. on p. 27).

Declaration

I hereby confirm that this thesis is entirely my own original work, without contributions from any sources other than those specified.

Braunschweig, April 11, 2023

Jan Heinrich Reimer