A High-Precision Health-Relatedness Score for Phrases to Mine Cause–Effect Statements from the Web

Anonymous ACL submission

Abstract

The measurement of the health-relatedness of a phrase is important when mining the web at scale for health information, e.g., when building a search engine or when carrying out health-sociological analyses. We propose a new termhood scoring scheme that allows for the prediction of the health-relatedness of phrases at high precision. An evaluation on several corpora of cause–effect statements (heuristically and professionally labeled) yields about 60% recall at over 90% precision, outperforming state-of-the-art vocabulary-based approaches and performing on par with BERT while being less resource-demanding. A new resource of over 4 million health-related cause–effect statements is compiled, such as "Studies show that stress induces insomnia.", which explicitly connect symptoms ('stress') as claimed causes for conditions ('insomnia'). It consists of over 4 million sentences from more than 2 million unique web pages and 234,000 unique websites.[1]

1 Introduction

Health sociology investigates society's interaction with health, where an important subject of interest is how consumers obtain and perceive health-related information. The web, as a main source (Sbaffi and Rowley, 2017), has been frequently studied in this regard throughout the past two decades. Three systematic reviews summarize the outcomes of 79, 157, and 165 studies, respectively (see Table 1): The studies typically focus on a single medical domain and range in size from a handpicked single page to up to 1,524 pages, with averages of 100.5, 78.5, and 50.3 pages per study.

Virtually all the aforementioned studies have been carried out manually. In order to enable scaling up such studies, further automation of various prerequisite tasks is required: (1) the discovery and acquisition of websites and web pages with relevance to health, (2) the extraction of specific health-related statements, and (3) the attribution of health-related statements to authoritative sources (e.g., for fact checking). While the first and third step have been and are subject to ongoing research and development, the second step has received much less attention thus far, especially given the requirement of reaching a high precision so as to minimize noise in subsequent analyses.

Rev.	Studies		Websites / Web Pages				
	Year	Count	Min	Max	Mean	Stddev	Sum
a	2001	79	3	1,147	100.5	157.7	7,796
b	2013	165	3	388	78.5	73.4	12,870
c	2017	157	1	1,524	50.3	133.9	7,891

Table 1: Key statistics of the number of websites or web pages analyzed in studies of online health information as reviewed by (a) Eysenbach et al. (2002), (b) Zhang et al. (2015), (c) Daraz et al. (2018). Some studies are part of more than one review; most do not differentiate websites from web pages.

Since a substantial portion of the information need of health consumers relates to causes and effects, be it the etiology of a condition or the effect of a treatment, we focus on this specific case and contribute towards automating the aforementioned second step as follows: (1) A new approach for measuring the health-relatedness of phrases with high precision is introduced (Section 3). (2) Based on our approach, a new resource compiles health-related cause–effect statements at web scale (Section 4). (3) In an in-depth evaluation, the approach is compared to several state-of-the-art approaches, outperforming state-of-the-art baselines for medical entity linking, while performing on par with BERT while requiring significantly less resources (Section 5).

[1] Code and data will be published alongside the paper; an excerpt is found as supplementary material.

2 Related Work

The impact that online information can have on a consumer's health has sparked the interest of the health-sociological research community ever since the web established itself as an information source in society. For example, user surveys investigate consumers' perceptions of online health information (Diaz et al., 2002), e-health services (Andreassen et al., 2007), as well as the criteria by which consumers judge the quality of a website (Sun et al., 2019).

Information quality appears to be the most-investigated characteristic. Numerous studies systematically reviewed the quality of websites with respect to specific topics like orthodontics (Jiang, 2000) and performance-enhancing drugs (Brennan et al., 2013). Apart from specific topics, restrictions to particular portions of the web are also common. Examples include small-scale studies of dietary advice (Cooper et al., 2012) and the misinterpretation (Yavchitz et al., 2012) or exaggeration (Sumner et al., 2014) of clinical trial results in online news. Recent research focused on social media (Suarez-Lledo and Alvarez-Galvez, 2021), particularly health misinformation on Twitter (Broniatowski et al., 2018; Bal et al., 2020). The accuracy of health information in search result snippets has also been investigated (Bondarenko et al., 2021).

Besides the mostly manual analyses, some quality assessment tasks have been automated, such as the detection of websites listing unproven cancer treatments (Aphinyanaphongs and Aliferis, 2007) as well as fake medical websites (Abbasi et al., 2012), and determining if a website conforms to the HON Code (Boyer and Dolamic, 2015; Boyer et al., 2017), a health information quality standard for websites.

In terms of discriminating between health and non-health-related content, most previous work has focused on classifying entire articles or pages. For example, medical vocabularies are used to detect news articles related to health (Watters et al., 2002; Zheng et al., 2002), and convolutional neural networks to detect mental health-related Reddit posts (Gkotsis et al., 2017). Little previous work exists on classifying phrases or terms as health-related, preventing the automation fact-checking. Further afield, keyword extraction and automatic ontology creation are related, where the goal is to extract prototypical words for a particular domain. For example, the C-value/NC-value method extracts multi-word domain terms from a corpus using term frequencies (Frantzi et al., 2000). Its reliance on the syntactic structure of extracted candidate words render it inapplicable to arbitrary phrases.

More straightforwardly applied are the family of contrastive termhood scores. which relate term frequencies from a domain corpus to term frequencies from one or more out-of-domain corpora. These include tf-idf-inspired measures (Basili et al., 2001; Kim et al., 2009), measures estimating how exclusive a term is for a domain (Khurshid et al., 2000; Park et al., 2008), and combinations or extensions thereof (Wong et al., 2007; Bonin et al., 2010). We transfer contrastive termhood scoring to measuring health-relatedness and compare it with the state-of-the-art medical entity linker, QuickUMLS (Soldaini and Goharian, 2016). Unlike classical medical entity linking algorithms, like MetaMap (Aronson, 2001) and cTakes (Savova et al., 2010), QuickUMLS is faster, achieves higher F1 and recall on several benchmarks, and can be tuned to prioritize precision or recall. Whereas entity linkers using neural language models (Neumann et al., 2019; Nejadgholi et al., 2019) are trained on entire abstracts and require additional context to extract entity candidates, rendering them inapplicable to phrases.

3 Measuring Health-Relatedness

Determining if a phrase is health-related is an issue of ambiguity. Homonomy (same surface form, different meaning) and polysemy (same surface form, different sense) render this decision difficult.[2] This section revisits so-called termhood scores, which measure the degree to which a given word is specific to a certain domain (Kageura and Umino, 1996). We introduce a new generalized score for phrases, and show how to tailor it to the health domain. Underlying our generalized termhood score are contrastive weight (CW) (Basili et al., 2001), term domain-specificity (TDS) (Khurshid et al., 2000; Park et al., 2008), and discriminative weight (DW) (Wong et al., 2007).

[2]The word 'cancer' can refer to a clearly health-related malignant tumor, but also to the zodiac sign, which is less likely to appear in a health-related context.

3.1 Contrastive Termhood Scores

All three considered contrastive termhood scores rely on a corpus of domain-specific text and a contrastive corpus of out-of-domain text. Formally, the health corpus H and the contrastive general corpus G are each represented by the multisets of all words in their texts. The corpus frequency $cf_C(c)$ of a word w in a corpus C denotes the absolute number of w's occurrences in C, while the relative corpus frequency $rf_C(w)$ denotes $cf_C(w)/|C|$ and the inverse corpus frequency $icf(w)$ denotes

$$icf(w) = \log\left(\frac{|H| + |G|}{cf_H(w) + cf_G(w)}\right).$$

The contrastive weight CW of a word w is defined as

$$\mathrm{CW}(w) = \log\left(cf_H(w) + 1\right) \cdot icf(w).$$

It is strongly related to $tf{\cdot}idf$, but instead of term and inverse document frequency, it uses corpus and inverse corpora frequency. The term domain-specificity TDS measures the domain exclusivity of a word w:

$$\mathrm{TDS}(w) = \log\left(\frac{rf_H(w) + 1}{rf_G(w) + 1} + 1\right).$$

The discriminative weight DW was originally defined as the product of CW and an unnormalized version of TDS, which used the corpus frequency cf instead of the relative frequency rf. Since varying corpora sizes heavily affect the unnormalized TDS score, we replace it with its normalized version and simply define DW as

$$\mathrm{DW}(w) = \mathrm{CW}(w) \cdot \mathrm{TDS}(w).$$

3.2 Generalized Phrase Termhood

To calculate termhood scores for phrases instead of words, it appears straightforward to average a phrase's individual word termhood scores: However, this does not work well for health-related phrases with many out-of-domain or stop words, like "unnecessary plastic surgery". Even though 'surgery' has a high termhood score, the overall average is rather low, due to the out-of-domain words 'unnecessary' and 'plastic'. We propose two schemes that avoid the issues of the simple average.

The first uses a weighted average to boost a phrase's words with high termhood. The idea is that a single highly health-related word is able to dictate the termhood score of a phrase, thereby increasing recall. Phrases with a high (unweighted) average termhood will still be ranked high so that precision is not affected. We calculate the weighted average of the termhood scores $x_1, \ldots, x_m$ of an m-word phrase as the generalized mean

$$M_p(x_1, \ldots, x_m) = \left(\frac{1}{m}\sum_{i=1}^{m} x_i^p\right)^{\frac{1}{p}}$$

with the non-zero real-valued parameter p. For $p = 1$, the generalized mean corresponds to the arithmetic mean. By increasing p, the mean is biased towards the higher-valued termhood scores; in the extreme case of $p = \infty$, the largest x_i is returned.

As the second scheme, we propose to also compute the weighted average termhood over the n-grams of a phrase. While the unigram 'plastic' is relatively unrelated to health, the bigram 'plastic surgery' certainly is health-related. Though the above generalized mean already increases the bigram's termhood compared to a simple average, the high occurrence frequency of the bigram itself is an even better indicator for its health-relatedness. Due to the sparsity of larger n-grams, especially prevalent in smaller corpora, we average the termhood scores of a phrase over multiple n-grams. Let s denote the phrase $w_1, \ldots, w_m$ and let $s_{i,k}$ denote the subphrase $w_i, \ldots, w_{i+k}$ of s ($0 \leq k \leq m-1$ and $i \in \{1, \ldots, m-k\}$). Let the above termhood scores $t(.)$ (i.e., CW, TDS, and DW) all be pre-calculated up to n-grams. The phrase termhood $\mathrm{PT}_{t,n,p}(s)$ for phrase s is then defined as

$$\mathrm{PT}_{t,n,p}(s) = \frac{1}{n}\sum_{k=0}^{n-1}\left(\frac{1}{m-k}\sum_{i=1}^{m-k} t(s_{i,k})^p\right)^{\frac{1}{p}}.$$

Corpus		Language	Documents	Words
G	Wikipedia	mixed, layp.	12,265,374	$3.0{\cdot}10^9$
H_1	PubMed	scientific	31,847,923	$3.8{\cdot}10^9$
H_2	PubMed Centr.	scientific	3,611,361	$5.4{\cdot}10^9$
H_3	Textbooks	clinical, educational	434	$1.4{\cdot}10^7$
H_4	Encyclopedias	mixed, layperson	67,967	$9.3{\cdot}10^6$

Table 2: Overview of our evaluation corpora.

3.3 Adaptation to the Health Domain

We select Wikipedia[3] as our contrastive corpus G because of its domain variety, relatively uniform language, and accessibility to the general public. As candidates for a health corpus H, we consider and evaluate four alternatives, each with its own (dis)advantages (see Table 2 for an overview).

The first three corpora use documents provided by the National Library of Medicine: a dump of over 30 million abstracts from PubMed[4], a subset of over 3 million full-text publications from PubMed Central[5] and 434 textbooks of the textbook and monograph category from the NCBI Bookshelf[6]. While both PubMed based corpora are large scale, their language is mainly scientific. The textbook corpus contains more clinical language, which we hypothesize to more closely match the expected proficiency level of web language.

Finally, we also crawled the entries of five consumer-oriented medical encyclopedias (Appendix A). Because the encyclopedias are purposefully written in layperson's terms, its joint language distribution is assumed to be most similar to the target language distribution.

3.4 Pilot Experiments

Comparing the three scores, Figures 1a-c show that, for the PubMed corpus H_1, all scores rank out-of-domain words and stop words lower than health-related words. However, the distributions of CW and TDS differ substantially, with the DW striking a balance between both. While the TDS ranks comparably few words as extremely health-related, the CW has a more even distribution with less extreme differences. Especially, 'ward' has a large difference in ranking between both scores. While it occurs frequently within the health domain so that CW attributes a high health-relatedness, it also occurs frequently in the general domain. Its lacking exclusiveness leads to the TDS scoring it comparably low.

To gain an intuition into the effect of using the different health corpora H_1 to H_4 with the termhood scores, Figures 1c and d compare the DW using the PubMed and Encyclopedia corpora. Here, the difference between scientific and non-scientific language between the two corpora is evident. The word 'experiment' has a comparably high termhood score using the PubMed corpus, but is similarly ranked to stop words using the Encyclopedia corpus. Otherwise, the shape of the distributions and location of examples are fairly similar.

[3]Specifically a dump of English Wikipedia articles from June 1st, 2021.

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

[5]https://www.ncbi.nlm.nih.gov/pmc/

[6]https://www.ncbi.nlm.nih.gov/books

Subset	Opt. Measure	Size	HR	P	R
Full	F1	2,968,345	25.6%	0.78	0.83
Full	Prec.	1,623,968	14.0%	0.90	0.73
Support	F1	111,406	61.8%	0.88	0.93
Support	Prec.	103,792	57.6%	0.91	0.89

Table 3: Descriptive statistics of four health-related cause–effect networks extracted from the CauseNet. The number of statements in each dataset, proportion of health-related statements as well as estimated precision and recall are listed.

4 Case Study: Cause–Effect Statements

To evaluate and demonstrate our approach, we apply it to a large graph of cause–effect statements. The CauseNet (Heindorf et al., 2020) is a graph of over 11 million pairs of cause and effect phrases extracted from all sentences found in the web pages of the ClueWeb12 web crawl.[7] It is important to note that these statements are *claimed* cause–effect statements, i.e., statements that have been made on some web page. Therefore, it contains of cause–effect statements for which empirical evidence can be found ("earthquake → tsunami"), but also many for which this is not the case ("jupiter opposing mars → bad luck on the job"). The statements were extracted using a linguistic pattern matching, achieving an estimated precision of 83%. Precision can be further increased to an estimated 93% by only considering statements with high support, i.e. statements which were extracted more than once using different linguistic patterns. The increase in precision of course takes a toll on recall. Only about 1.6% of statements have high support.

We evaluate our termhood approach (see Section 5.3) on several manually labeled subsets of the CauseNet. Based on this evaluation, we extract four different health-related cause–effect networks, one maximizing the F1-

[7]https://www.lemurproject.org/clueweb12/

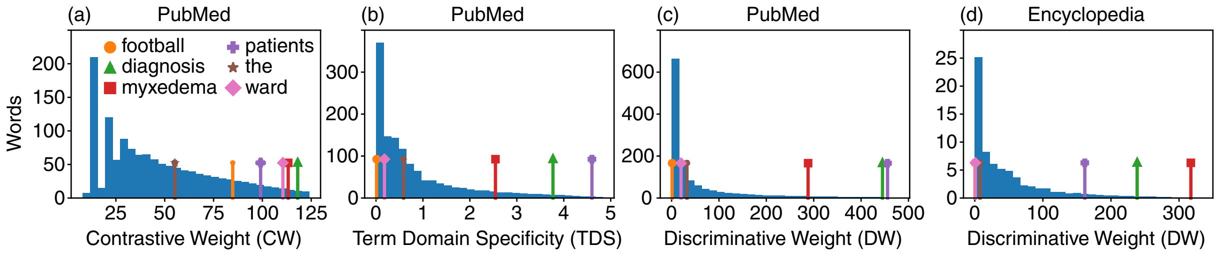

Figure 1: Histograms of termhood frequencies for the (a) CW, (b) TDS, and (c) DW on the PubMed corpus, and for (d) DW on the Encyclopeida corpus. Example words are highlighted.

measure, and one maximizing the F1-measure with at least 90% precision for both the full and the high-precision CauseNet with high support, and release these to the public for further analyses. Table 3 summarizes the descriptive statistics of each resource. Optimizing for high precision on the full CauseNet yields 1,623,968 health-related statements. With high precision, an estimated 14% and an estimated 58% of all statements within the full support subsets of the CauseNet are respectively health-related.

5 Evaluation

This section reports on an in-depth evaluation of our health-relatedness score compared to state-of-the-art entity linking and BERT-based baselines and different parameterizations on four labeled cause–effect statement datasets.

5.1 Baselines

Vocabulary-Based Approach. A precision oriented approach for determining health-relatedness of a phrase is to check if it, or subphrases of it, are part of a medical vocabulary. To judge the performance of our health-relatedness score, we therefore compare it to a vocabulary-based approach based on QuickUMLS (Soldaini and Goharian, 2016), a state-of-the-art medical entity linker. The proportion of words within a phrase which could be matched to UMLS concepts is then used as a health-relatedness score.

In more detail, given a phrase s, first all medical entity mentions E are extracted. Overlapping entity mentions, e.g. 'cancer' is contained in "breast cancer", are handled by taking only the longest mentioned entity, resulting in a subset of non-overlapping entity mentions $\hat{E}$. Next, stop words[8] not contained in any entity mentions are removed from s, yielding $\hat{s}$. The vocabulary health-relatedness score $V(s)$ is then computed as

$$V(s) = \frac{|\hat{s}| - \sum_{e_i \in \hat{E}} |e_i|}{|\hat{s}|}.$$

Entity mentions are linked to the UMLS Metathesaurus (Humphreys and Lindberg, 1993), which is a mix of medical vocabularies of varying specificity. We investigate three decreasingly specific vocabulary subsets in an attempt to increase precision: the MeSH hierarchy[9], the MeSH hierarchy with additional synonyms (MeSH Syn) and the entire UMLS Metathesaurus (see Appendix B for further details). For all three variants, we also consider restricting the set of concepts to a set of medically specific semantic types (ST21pv) as proposed in the MedMentions entity linking dataset (Mohan and Li, 2019). Finally, several different string similarity thresholds (Jaccard similarity was used in this work) were tested to allow for fuzzy string matching and increase recall.

BERT-Based Approach.

As a second baseline, we use a BERT-based sequence classifier which is trained to predict if a sequence of tokens originates from a health-related corpus. Starting from a pretrained SciBERT (Beltagy et al., 2019) model, we fine-tune two different models. One model each is trained to predict if a noun phrase originates from the PubMed H_1 or the Encyclopedia H_4 corpus. Noun phrases from the Wikipedia corpus G serve as negative samples. Further details about the training procedure can be found in Appendix C.

[8]The English nltk stop words list is used.

[9]https://www.nlm.nih.gov/mesh/meshhome.html

5.2 Reference Datasets

We apply three labeling strategies across four reference datasets. The first reference dataset is collected from Wikidata and labeled using a heuristic. With the help of a medical practitioner (a practicing orthopedist and professor), we gather nine general root concepts that include the majority of health-related concepts. Then all 9,317 Wikidata relations with the *has cause* (P828) and/or *has effect* (P1542) predicates are extracted. All relations for which both the cause and effect concepts are direct or indirect children of the root concepts are considered health-related. See Appendix D for a full list of root concepts and further details. As this dataset propagates labels heuristically, we consider it a silver standard.

Next, we manually classified two different sets of CauseNet statements; 1,000 randomly sampled statements from each, the full CauseNet (Full), and the high-support subset (Support). A subset of 100 statements from the Full dataset was labeled by 3 separate annotators, achieving a Cohen's Kappa of 0.77. These datasets are considered as gold standard.

Finally, after evaluating on the aforementioned datasets, we sampled 1,000 statements from the full CauseNet that were closest to the decision threshold of the termhood classifier with the highest F1 score and at least 90% precision on the Full dataset. The aforementioned practitioner labeled the dataset, with the additional option to label statements with unsure. For lack of a better term, we call this dataset a platinum standard, as it is professionally labeled and specifically focuses on a difficult subset of cause–effect statements. The best approaches are evaluated on the full dataset (Practitioner-Full) and the confidence splits (Practitioner-Sure, Practitioner-Unsure).

Table 4 gives an overview of dataset statistics. Interestingly, the proportion of health statements within the CauseNet datasets varies substantially. The higher the support, the more likely a statement is health-related. Additionally, the high proportion of statements marked as unsure by the practitioner shows the difficulty of the task. While some statements were marked unsure because of unknown terminology, most were borderline decisions because of ambiguous concepts (e.g., *poor treatment* → *problems*), or the difficulty to delineate the health domain from other related domains (e.g., biological processes *cold air* → *bronchoconstriction*).

Dataset	Size	Health-related	Length
Wikidata	9317	31.0%	4.90
Full	1000	19.7%	7.21
Support	1000	50.3%	3.40
Practitioner-Full	1000	77.2%	5.99
Practitioner-Sure	594	82.0%	5.84
Practitioner-Unsure	406	70.2%	6.21

Table 4: Overview of the evaluation datasets, including the proportion of true health-related cause–effect statements, and the average number of words per cause/effect phrase.

5.3 Results

To combine the individual termhood scores of the cause and effect phrases we use "and" (both cause and effect scores need to exceed the decision threshold) as an upper bound precision-oriented operator. Following the rationale of using the generalized mean for increasing recall in health-phrase detection, we also test the generalized mean for combining cause and effect phrase termhood. To differentiate between parameters, we denote p for averaging n-gram termhood by p_n and for averaging phrase termhood by p_p. By setting $p_p = \infty$, the maximum phrase termhood score of either cause and effect is used. Thereby, $p_p = \infty$ is the same as using "or" (one of cause or effect phrase termhood scores need to exceed the decision threshold) and acts as the complement to the "and" operator and as a recall-oriented upper bound.

To evaluate the different approaches we run a grid search over the parameters and thresholds on the silver and gold-standard datasets and test for significance using a bootstrap test with 5,000 permutations. See Appendix E for a full description of parameters. Table 5 gives an overview of the best variants for each approach in terms of F1 measure.

While no approach is able to statistically significantly outperform all others, all termhood scores and the BERT-based approach are able to statistically significantly ($p < 0.05$) outperform the vocabulary approaches across all datasets. The vocabularies are unable to achieve high precision. While it is usually possible to tune the decision threshold to achieve perfect precision, the binary classification, i.e.

	Approach	Parameters	Operator	P	R	F1
Wikidata	MeSH	Jacc.=0.9	p_p=1	0.57	0.74	0.65
	MeSH Syn	Jacc.=0.9	p_p=2	0.55	0.78	0.65
	UMLS	Jacc.=1.0	AND	0.49	0.83	0.62
	BERT	PubMed	p_p=∞	0.72	0.82	0.77
	CW	Encyc., n=1, p_n=10	p_p=5	0.70	0.74	0.72
	TDS	Encyc., n=1, p_n=10	p_p=2	0.70	**0.88**	**0.78**
	DW	Encyc., n=1, p_n=2	p_p=2	**0.74**	0.83	0.78
Full	MeSH	Jacc.=1.0	p_p=10	0.44	0.77	0.56
	MeSH Syn	Jacc.=1.0	p_p=1	0.57	0.60	0.59
	UMLS	Jacc.=1.0	AND	0.40	0.69	0.51
	BERT	PubMed	p_p=10	**0.84**	0.79	**0.81**
	CW	Encyc., n=1, p_n=5	p_p=5	0.77	0.79	0.78
	TDS	Encyc., n=3, p_n=1	p_p=2	0.74	**0.86**	0.79
	DW	Encyc., n=3, p_n=1	p_p=1	0.78	0.83	0.80
Support	MeSH	Jacc.=1.0	p_p=1	0.62	0.87	0.72
	MeSH Syn	Jacc.=0.9	p_p=1	0.74	0.76	0.75
	UMLS	Jacc.=1.0	AND	0.60	0.91	0.72
	BERT	PubMed	p_p=10	**0.92**	0.87	0.90
	CW	Encyc., n=3, p_n=10	p_p=2	0.82	0.88	0.85
	TDS	Encyc., n=2, p_n=1	p_p=1	0.88	0.93	**0.90**
	DW	Encyc., n=3, p_n=5	p_p=1	0.86	**0.93**	0.90

Table 5: Parameterizations of each approach optimized for F1 on the silver- and gold-standard evaluation datasets.

	Approach	Parameters	Operator	P	R	F1
Wikidata	BERT			-	-	-
	CW	Encyc., n=1, p_n=10	p_p=5	0.90	0.39	0.54
	TDS	Encyc., n=2, p_n=1	p_p=1	0.90	0.17	0.28
	DW	Encyc., n=1, p_n=5	p_p=1	**0.91**	**0.50**	**0.64**
Full	BERT	PubMed	AND	**0.92**	0.50	**0.65**
	CW	Encyc., n=1, p_n=5	p_p=2	0.91	0.52	0.66
	TDS	Encyc., n=2, p_n=1	p_p=1	0.90	**0.62**	**0.73**
	DW	Encyc., n=3, p_n=2	p_p=1	0.92	0.58	0.71
Support	BERT	PubMed	p_p=10	**0.92**	0.87	0.90
	CW	Encyc., n=1, p_n=10	p_p=5	0.91	0.75	0.82
	TDS	Encyc., n=3, p_n=1	p_p=1	0.91	**0.89**	**0.90**
	DW	Encyc., n=2, p_n=1	p_p=1	0.90	0.87	0.89

Table 6: Parameterizations of each approach with at least 90% precision optimized for F1 on the silver- and gold-standard datasets.

a concept is either health-related or not, creates an upper bound for the vocabulary based approaches. Even setting the threshold such that only fully matched phrases are included leads to some false positive predictions.

Termhood Score Comparison. Compared to the vocabulary approaches, the termhood scores have more granular distributions and the decision threshold can therefore be tuned to achieve high precision. Table 6 lists the best performing approaches with at least 90% precision in terms of F1-measure (none of the vocabulary approaches were able to achieve more than 90% precision). When high precision is required, the term domain-specificity outperforms the contrastive weight on the two CauseNet evaluation datasets, featuring substantially higher recall at similar precision. However, the exact opposite relationship can be seen for the Wikidata dataset. By combining both CW and TDS, the DW achieves the best or only marginally worse F1 scores on all evaluation datasets.

Taking a closer look at the effect the various health corpora have on classification performance shows that the Encyclopedia corpus always leads to the best performance. Irrespective of dataset, every termhood score is able to achieve the highest overall F1 score in both unconstrained (Table 5) and constrained precision scenarios (Table 6) using the Encyclopedia corpus. This effect does not translate to the BERT model. The models trained on the PubMed H_1 corpus outperform the Encyclopedia H_4 corpus trained models on all datasets.

In contrast, the effects of the n-gram and generalized mean parameters on performance are more subtle. The CW and the TDS each prefer high and low p_n values respectively. Especially the contrastive weight profits from the possibility to increase the p_n value and subsequently increase recall. The term domain-specificity is already precision-oriented and therefore performs best with lower p_n values. The DW again strikes a balance between both and prefers higher or lower p_n values depending on the proportion of health-related labels in the dataset.

Finally, the n-gram variants have the smallest impact on performance. When switching to $n = 3$, the CW loses 10% points in recall on the Full dataset, while the TDS and DW have their largest drops at 9% and 5% points for $n = 1$ respectively. Over all other datasets the drop in performance is negligible. This most likely stems from the fact that the Full dataset has by far the longest average event length at 7.21 words per event, which enables the termhood scores to take full advantage of longer n-grams.

Practitioner Evaluation. We finally evaluate the termhood scores and BERT-based approach on the difficult subset of relations labeled by a medical practitioner. As a reminder, based on the results of the evaluation on the silver- and gold-standard datasets, we use the discrim-

	Approach	Parameters	Operator	P	R	F1
Full	BERT	PM	AND	**0.91**	0.18	0.30
	CW	Encyc., n=3, p_n=10	AND	0.91	**0.19**	**0.31**
	TDS	Encyc., n=3, p_n=1	AND	0.90	0.06	0.11
	DW	Encyc., n=2, p_n=1	AND	0.91	0.15	0.25
Sure	BERT	PM	AND	0.90	**0.82**	**0.86**
	CW	Encyc., n=2, p_n=2	AND	**0.90**	0.59	0.72
	TDS	Encyc., n=1, p_n=1	AND	0.90	0.34	0.50
	DW	Encyc., n=3, p_n=1	AND	0.90	**0.66**	0.76
Unsure	BERT	PM	AND	**1.00**	0.02	0.05
	CW	Encyc., n=1, p_n=5	AND	**0.94**	0.05	0.10
	TDS	Textbook, n=1, p_n=2	$p_p=\infty$	0.90	**0.10**	**0.18**
	DW	Textbook, n=2, p_n=2	p_p=5	0.91	0.07	0.14

Table 7: Parameterizations of each statistical approach with at least 90% precision optimized for F1 on the practitioner evaluation datasets.

inative weight (Encyclopedia corpus, $n = 3$, $p_n = 2$, $p_p = 1$) and sample the 1,000 relations closest to the decision threshold (120) tuned for high F1 with precision over 90%.

Table 7 gives an overview of the best paramemeterizations. Again requiring high precision, we find that performance drastically drops on the Practitioner-Full dataset. All scores have to set a high decision threshold and use the "and" operator to reach the high precision and requirement and thereby sacrifice recall. Considering the split of into Practitioner-Sure and Practitioner-Unsure relations however shows that the approaches especially struggle with the relations the practitioner was not confident about. The performance on the Practitioner-Sure subset on the other hand is on par with the best approaches on the Full dataset, with the BERT approach significantly outperforming the termhood scores.

To gain insight into the performance drop of the unsure dataset, we sample the highest scored true negative relations of the best discriminative weight approach, i.e. the relations which the approach considers to likely be health-related that the practitioner labeled as health unrelated. See Table 8 for the top four examples. We find that the two issues, health domain demarcation and handling general concepts mentioned in Section 5.2, reflect themselves in the classification. A "stroke" causing a "reduced cell count", as in the second to last example of Table 8, might have medical relevance, but could just as well be the plain description of a biological process. Second, while it is unclear which "small amounts" are meant

Cause → Effect	**DW**	
	Cause	Effect
cellulite → congested digestive system	155.95	140.32
significant toxin buildup → feeling	164.37	129.70
condition → changes to the brain	179.32	110.98
stroke → reduced cell count	128.57	160.20
small amounts → digestive problems	108.52	176.77

Table 8: Highest scored true negative statements by the best-performing discriminative-weight approach on the Practitioner-Unsure dataset. The first effect contains a spelling error not handled by the web crawl extraction.

in final example of Table 8, the discriminative weight nonetheless considers the concept as likely health-related because of its frequent usage.

6 Conclusions

We develop a novel approach to determine the health-relatedness of arbitrary short phrases with a high precision, developing a new generalized termhood score. To demonstrate and evaluate our approach in a realistic setting, we apply it, among other datasets, to a web-scale graph of cause–effect statements. In comparison to state-of-art entity linking approaches, our approach is the only one capable of achieving the high precision required for practical purposes, outperforming the baseline approaches on all evaluation datasets. Combined with our generalization, the discriminative weight score proves to be most robust, with the term domain-specificity performing slightly better in high-precision scenarios.

We apply the best precision and F1-oriented approaches to the full CauseNet graph, and a precision oriented subset of the graph. The result is a new resource of high-precision health-related statements at an unprecedented scale, suitable for investigating health-sociological questions automatically. At an estimated precision of 0.9 and estimated recall of 0.73, the precision-oriented extraction on the full graph contains 1,623,968 health-related statements from 4,420,897 statements as well as 234,355 and 2,139,563 unique websites and web pages. This opens up new possibilities for the quantitative analysis of health-related information on the web.

Ethical Considerations

Research on health-related tasks can be often sensitive as its haphazard transfer into practice may cause significant harm. Though our research is not aimed at supporting medical treatments, its envisioned application in health-sociological analyses may cause these analyses to include or exclude pieces of information in error. This is why we explicitly aim for a high precision: ensuring that a phrase that achieves a high health-relatedness score is actually health-related protects both the time and effort of health sociologists tasked with analyzing them, as well as the privacy of people whose web content is ambiguous or otherwise close to health, but not quite crossing the line from being subject to critical interpretation by health experts. Nevertheless, for some applications, achieving a high recall, and thus be inclusive of all that is health-related at the expense of false positives might also be important. The limitations of our approach in this regard are clearly outlined, yet we do see potential of shifting its operating point toward that end.

References

Ahmed Abbasi, Fatemeh “Mariam” Zahedi, and Siddharth Kaza. 2012. Detecting Fake Medical Web Sites Using Recursive Trust Labeling. *ACM Transactions on Information Systems* 30, 4 (Nov. 2012), 22:1–22:36. https://doi.org/10.1145/2382438.2382441

Hege K. Andreassen, Maria M. Bujnowska-Fedak, Catherine E. Chronaki, Roxana C. Dumitru, Iveta Pudule, Silvina Santana, Henning Voss, and Rolf Wynn. 2007. European Citizens’ Use of E-Health Services: A Study of Seven Countries. *BMC Public Health* 7, 1 (April 2007), 53. https://doi.org/10.1186/1471-2458-7-53

Yin Aphinyanaphongs and Constantin Aliferis. 2007. Text Categorization Models for Identifying Unproven Cancer Treatments on the Web. *Studies in Health Technology and Informatics* 129, Pt 2 (2007), 968–972.

A. R. Aronson. 2001. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. *Proceedings of the AMIA Symposium* (2001), 17–21.

Rakesh Bal, Sayan Sinha, Swastika Dutta, Rishabh Joshi, Sayan Ghosh, and Ritam Dutt. 2020. Analysing the Extent of Misinformation in Cancer Related Tweets. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 2020), 924–928.

Roberto Basili, Alessandro Moschitti, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. 2001. A Contrastive Approach to Term Extraction. In *Proceedings of the TIA 2001: Terminologie et Intelligence Artificielle*. 119–128.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3615–3620. https://doi.org/10.18653/v1/D19-1371

Alexander Bondarenko, Ekaterina Shirshakova, Marina Driker, Matthias Hagen, and Pavel Braslavski. 2021. Misbeliefs and Biases in Health-Related Searches. In *30th ACM International Conference on Information and Knowledge Management (CIKM 2021)*. ACM. https://doi.org/10.1145/3459637.3482141

Francesca Bonin, Felice Dell’Orletta, Giulia Venturi, and Simonetta Montemagni. 2010. A Contrastive Approach to Multi-Word Extraction from Domain-Specific Corpora. In *Proceedings of the International Conference on Language Resources and Evaluation*. European Language Resources Association, Valletta, Malta, 19–21.

Célia Boyer and Ljiljana Dolamic. 2015. Automated Detection of HONcode Website Conformity Compared to Manual Detection: An Evaluation. *Journal of Medical Internet Research* 17, 6 (June 2015), e3831. https://doi.org/10.2196/jmir.3831

Célia Boyer, Cédric Frossard, Arnaud Gaudinat, Allan Hanbury, and Gilles Falquetd. 2017. How to Sort Trustworthy Health Online Information? Improvements of the Automated Detection of HONcode Criteria. *Procedia Computer Science* 121 (Jan. 2017), 940–949. https://doi.org/10.1016/j.procs.2017.11.122

Brian P. Brennan, Gen Kanayama, and Harrison G. Pope. 2013. Performance-Enhancing Drugs on the Web: A Growing Public-Health Issue. *The American Journal on Addictions* 22, 2 (2013), 158–161. https://doi.org/10.1111/j.1521-0391.2013.00311.x

David A. Broniatowski, Amelia M. Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C. Quinn, and Mark Dredze. 2018. Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. *American Journal of Public Health* 108, 10 (Oct. 2018), 1378–1384. https://doi.org/10.2105/AJPH.2018.304567

Benjamin E. J. Cooper, William E. Lee, Ben M. Goldacre, and Thomas A. B. Sanders. 2012. The Quality of the Evidence for Dietary Advice given in UK National Newspapers. *Public Understanding of Science* 21, 6 (Aug. 2012), 664–673. https://doi.org/10.1177/0963662511401782

Lubna Daraz, Allison S. Morrow, Oscar J. Ponce, Wigdan Farah, Abdulrahman Katabi, Abdul Majzoub, Mohamed O. Seisa, Raed Benkhadra, Mouaz Alsawas, Prokop Larry, and M. Hassan Murad. 2018. Readability of Online Health Information: A Meta-Narrative Systematic Review. *American Journal of Medical Quality: The Official Journal of the American College of Medical Quality* 33, 5 (2018), 487–492. https://doi.org/10.1177/1062860617751639

Joseph A. Diaz, Rebecca A. Griffith, James J. Ng, Steven E. Reinert, Peter D. Friedmann, and Anne W. Moulton. 2002. Patients' Use of the Internet for Medical Information. *Journal of General Internal Medicine* 17, 3 (2002), 180–185. https://doi.org/10.1046/j.1525-1497.2002.10603.x

Gunther Eysenbach, John Powell, Oliver Kuss, and Eun-Ryoung Sa. 2002. Empirical Studies Assessing the Quality of Health Information for Consumers on the World Wide WebA Systematic Review. *JAMA* 287, 20 (May 2002), 2691–2700. https://doi.org/10.1001/jama.287.20.2691

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic Recognition of Multi-Word Terms:. The c-Value/Nc-Value Method. *International journal on digital libraries* 3, 2 (2000), 115–130.

George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim J. P. Hubbard, Richard J. B. Dobson, and Rina Dutta. 2017. Characterisation of Mental Health Conditions in Social Media Using Informed Deep Learning. *Scientific Reports* 7, 1 (March 2017), 45141. https://doi.org/10.1038/srep45141

Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. CauseNet: Towards a Causality Graph Extracted from the Web. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery, New York, NY, USA, 3023–3030.

B L Humphreys and D A Lindberg. 1993. The UMLS Project: Making the Conceptual Connection between Users and the Information They Need. *Bulletin of the Medical Library Association* 81, 2 (April 1993), 170–177.

You-Ling Jiang. 2000. Quality Evaluation of Orthodontic Information on the World Wide Web. *American Journal of Orthodontics and Dentofacial Orthopedics* 118, 1 (July 2000), 4–9. https://doi.org/10.1067/mod.2000.104492

Kyo Kageura and Bin Umino. 1996. Methods of Automatic Term Recognition: A Review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 3, 2 (Jan. 1996), 259–289. https://doi.org/10.1075/term.3.2.03kag

Ahmad Khurshid, Lee Gillman, and Lena Tostevin. 2000. Weirdness Indexing for Logical Document Extrapolation and Retrieval. In *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*. Gaithersburg, USA, 1–8.

Su Nam Kim, Timothy Baldwin, and Min-Yen Kan. 2009. An unsupervised approach to domain-specific term extraction. In *Proceedings of the Australasian Language Technology Association Workshop*. Sydney, Australia, 94–98. https://www.aclweb.org/anthology/U09-1013

Sunil Mohan and Donghui Li. 2019. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. *arXiv:1902.09476 [cs]* (Feb. 2019). arXiv:1902.09476 [cs]

Isar Nejadgholi, Kathleen C. Fraser, Berry De Bruijn, Muqun Li, Astha LaPlante, and Khaldoun Zine El Abidine. 2019. Recognizing UMLS Semantic Types with Deep Learning. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*. Association for Computational Linguistics, Hong Kong, 157–167. https://doi.org/10.18653/v1/D19-6219

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics, Florence, Italy, 319–327. https://doi.org/10.18653/v1/W19-5034 arXiv:arXiv:1902.07669

Youngja Park, Siddharth Patwardhan, Karthik Visweswariah, and Stephen C Gates. 2008. An Empirical Analysis of Word Error Rate and Keyword Error Rate. In *Proceedings of the Ninth Annual Conference of the International Speech Communication Association*. Brisbane, Australia, 2070–2073.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, Component Evaluation and Applications. *Journal of the American Medical Informatics Association* 17, 5 (Sept. 2010), 507–513. https://doi.org/10.1136/jamia.2009.001560

Laura Sbaffi and Jennifer Rowley. 2017. Trust and Credibility in Web-Based Health Information: A Review and Agenda for Future Research. *Journal* of Medical Internet Research 19, 6 (2017), e218. https://doi.org/10.2196/jmir.7579

Luca Soldaini and Nazli Goharian. 2016. Quickumls: A Fast, Unsupervised Approach for Medical Concept Extraction. In *MedIR Workshop, SIGIR*. Association for Computing Machinery, New York, NY, USA, 1–4.

Victor Suarez-Lledo and Javier Alvarez-Galvez. 2021. Prevalence of Health Misinformation on Social Media: Systematic Review. *Journal of Medical Internet Research* 23, 1 (Jan. 2021), e17187. https://doi.org/10.2196/17187

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A. Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, Fred Boy, and Christopher D. Chambers. 2014. The Association between Exaggeration in Health Related Science News and Academic Press Releases: Retrospective Observational Study. *BMJ* 349 (Dec. 2014), g7015. https://doi.org/10.1136/bmj.g7015

Yalin Sun, Yan Zhang, Jacek Gwizdka, and Ciaran B. Trace. 2019. Consumer Evaluation of the Quality of Online Health Information: Systematic Literature Review of Relevant Criteria and Indicators. *Journal of Medical Internet Research* 21, 5 (May 2019), e12522. https://doi.org/10.2196/12522

Carolyn Watters, Wanhong Zheng, and Evangelos Milios. 2002. Filtering for Medical News Items. *Proceedings of the American Society for Information Science and Technology* 39, 1 (2002), 284–291. https://doi.org/10.1002/meet.1450390131

Wilson Wong, Wei Liu, and Mohammed Bennamoun. 2007. Determining Termhood for Learning Domain Ontologies Using Domain Prevalence and Tendency. In *Proceedings of the 6th Australasian Conference on Data Mining*. Citeseer, 47–54.

Amélie Yavchitz, Isabelle Boutron, Aida Bafeta, Ibrahim Marroun, Pierre Charles, Jean Mantz, and Philippe Ravaud. 2012. Misrepresentation of Randomized Controlled Trials in Press Releases and News Coverage: A Cohort Study. *PLOS Medicine* 9, 9 (Sept. 2012), e1001308. https://doi.org/10.1371/journal.pmed.1001308

Yan Zhang, Yalin Sun, and Bo Xie. 2015. Quality of Health Information for Consumers on the Web: A Systematic Review of Indicators, Criteria, Tools, and Evaluation Results. *Journal of the Association for Information Science and Technology* 66, 10 (2015), 2071–2084. https://doi.org/10.1002/asi.23311

Wanhong Zheng, Evangelos Milios, and Carolyn Watters. 2002. Filtering for Medical News Items Using a Machine Learning Approach. *Proceedings of the AMIA Symposium* (2002), 949–953.

A Encyclopedia Links

http://health.am/encyclopedia
https://medlineplus.gov/encyclopedia.html
https://merriam-webster.com/medical
https://ucsfhealth.org (var. sub pages)
https://www.rxlist.com/drug-medical-dictionary/article.htm

B UMLS Vocabularies

For all vocabulary subsets we use the 2020AB revision of the UMLS Metathesaurus. For the MeSH subset we gather all concepts contained in the MeSH vocabulary and filter out all atoms contained in MeSH. The MeSH Syn. subset also includes all MeSH concepts, but keeps all atoms linked to those concepts irrespective of the vocabulary that atom is from. For the full UMLS subset we use all concepts and atoms from every Category 0 (no additional restrictions or license terms apply) vocabulary.

C BERT Training

The BERT approach was trained by fine-tuning the huggingface[10] transformers `allenai/scibert_scivocab_uncased` checkpoint. PyTorch[11] and PyTorchLightning[12] were used to train the model using a batch size of 32 and learning rate of 0.000005. The input text was split into sentences using nltk[13] and noun phrases extracted using spacy.[14]. Due to the large corpora sizes fine-tuning converged before a single complete epoch was reached. Therefore, training was halted after no decrease in training loss was reached for 15 consecutive training loss samples, where a sample was taken every 1,000 steps.

D Wikidata Details

Root wikidata concepts: *fungus* (Q764), *protein* (Q8054), *microorganism* (Q39833), *biogenic substance* (Q289472), *medical procedure* (Q796194), *disease causative agent* (Q2826767), *etiology* (Q5850078), *physiological condition* (Q7189713) and *medicinal product* (Q86746756).

[10] https://huggingface.co/
[11] https://pytorch.org/
[12] https://www.pytorchlightning.ai/
[13] https://www.nltk.org/
[14] https://spacy.io/

All relations with a chain of predicates starting at one of the root concepts, and consisting of only *subclass of* (P279), *parent taxon* (P171), *risk factor* (P5642), and optionally ending with *instance of* (P31) to both the cause and effect concepts are considered as related to health.

In total, 11,160 relations were extracted from Wikidata. We removed 799 invalid relations with missing concept labels. We additionally removed all relations pertaining to COVID-19 (1,044 in total), because these are severely overrepresented and COVID-19 was not yet found in the Textbook corpus, nor CauseNet.

E Grid Search Parameters

For the three vocabulary approaches (MeSH, MeSH Syn. and UMLS) seven Jaccard distance thresholds (0.4, 0.5, . . . , 0.9, 1.0) were tested. For the three termhood scores (CW, TDS and DW) four health corpora (PubMed, PubMed Central, Textbook, Encyclopedias), three n-gram sizes n (uni-, bi- and trigrams) and five values for p_n (1, 2, 5, 10, ∞) for averaging n-gram termhood scores using the generalized mean are tested. Finally, for the final relation classification, the same set of values for p_p are tested for averaging cause and effect scores and in addition to the "and" operator.