Efficiently Scoring the Health-relatedness of Web Pages

Ferdinand Schlatt[1]

[1]Friedrich-Schiller Universität Jena

Abstract

This paper describes our software submission for determining the health-relatedness of web pages for the 1st International Workshop on Open Web Search [1] (WOWS'24). Determining to what degree a web page is health-related is crucial in health information retrieval, and scaling this task to the web requires efficient methods. We present a simple approach based on termhood scores, which determine how strongly a term is associated with a specific domain. In previous work, we have shown that termhood scores can effectively determine the health-relatedness of phrases and sentences. In this work, we apply them to web pages. Our software submission efficiently computes an effective and nuanced health / medical relatedness score for a given web page.

Keywords

termhood score, health information retrieval, web search

1. Introduction[1]

Consumers frequently turn to the web for health information [3]. The availability of online health information has democratized access to health knowledge but also harbors risks. Spreading misinformation and lacking quality of health information has raised concern among researchers and practitioners [4, 5].

Providing consumers with high-quality health information is a complex task. One of the first steps for building a search engine for health information is determining whether a web page contains high-quality health-related information. Previous work has focused on classifying the quality of health information [6, 7, 8, 9], but assumes that the web page is health-related. We focus on the prior task of determining the health-relatedness of a web page.

Determining health-relatedness at web scale requires an efficient and effective method. We propose an efficient approach based on termhood scores. Termhood scores determine how strongly a term is associated with a specific domain [10]. They were initially developed to automatically extract terms for building ontologies. Previously, we have shown that termhood scores can also effectively determine the health-relatedness of phrases and sentences [2]. In this work, we apply them to web pages and provide an easy-to-use software submission for the 1st International Workshop on Open Web Search to determine the health-relatedness of web pages.

2. Related Work

The impact of online health information on consumers has attracted the interest of the health sociology research community. Information quality appears to be the most studied characteristic from a health sociology perspective. Numerous studies systematically analyzed the quality of websites related to specific topics such as orthodontics [11] or performance-enhancing drugs [12], but more general studies with limitations to specific parts of the web are also common. Examples include studies of dietary advice [13] or the misinterpretation [14] and exaggeration [15] of clinical trial results in online news. Recent studies also targeted web search snippets [16] and social media [17], particularly health misinformation on Twitter [18, 19].

WOWS'24: 1st International Workshop on Open Web Search, March 28, 2024, Glasgow, Scotland

ferdinand.schlatt@uni-jena.de (F. Schlatt)

0000-0002-6032-909X (F. Schlatt)

[1]Several parts of this paper (especially in Section 3) are copied verbatim from our previous work [2].

Experts have defined different standards for the quality of online health information [20, 21, 22]. However, a large majority of websites do not follow these criteria, and the overall quality of health information is considered problematic by several surveys [4, 5, 23]. To aid consumers in finding high-quality information, previous work has automated quality assessment tasks [6, 7, 8, 9]. However, most approaches assume the web page is already health-related, and the task of determining health-relatedness is not addressed. Some previous work has investigated classifying if a web page is health-related. For example, medical vocabularies have been used to classify news articles [24, 25] or convolutional neural networks to classify Reddit posts [26]. These models, however, focus on specific types of web pages and are expensive to scale for web-scale analysis.

Instead, we use contrastive termhood scores to determine the health-relatedness of arbitrary web pages. Contrastive termhood scores relate term frequencies from a domain corpus to frequencies from one or more out-of-domain corpora. These include $tf \cdot idf$-inspired measures [27, 28], measures estimating how exclusive a term is for a domain [29, 30], and combinations or extensions thereof [31, 32]. They are particularly well-suited for web-scale analysis, as they are efficient to compute and do not require labeled training data. Contrastive termhood scores are usually applied to a set of extracted terms. We have previously shown that they can also be effectively applied to phrases and sentences [2]. We build on our previous work by applying contrastive termhood scores to web pages.

3. Measuring Health Relatedness

Termhood scores assess to what degree a term is exclusive to a particular domain [10]. Usually, termhood scores define a term as a nominal phrase consisting of one or multiple words, e.g., 'breast cancer.' However, obtaining nominal phrases requires expensive syntactic parsing. Instead, we consider a single word as a term. Assessing whether a word is health-related can be difficult, particularly for homonymous (same surface form, different meaning) or polysemous (same surface form, different sense) words. For instance, 'cancer' may refer to a health-related malignant tumor and the zodiac sign. The latter is unlikely to appear in a health-related context.

Contrastive termhood scores compare the occurrence frequencies of a word from an in-domain corpus to one or multiple out-of-domain corpora. Thereby, they determine a word's domain exclusivity. This section discusses three existing contrastive termhood scores and then explains how we apply them to assessing the health-relatedness of web pages.

3.1. Existing Contrastive Termhood Scores

The termhood scores contrastive weight (CW) [27], term domain specificity (TDS) [30], and discriminative weight (DW) [31] rely on a corpus H of domain-specific texts (in our case: health-related texts) and at least one contrastive corpus G of general or out-of-domain texts (in our case: Wikipedia). To score a term t, CW, TDS, and DW use occurrence frequencies: the absolute corpus occurrence frequency $freq_C(t)$ (i.e., the absolute number of occurrences of t in corpus C), the relative corpus occurrence frequency $rel_C(t) = freq_C(t)/|C|$ (where $|C|$ denotes the number of words in a corpus), and the inverse corpora frequency $icf(t)$ defined for H and G together as

$$icf(t) = \log\left(\frac{|H| + |G|}{freq_H(t) + freq_G(t)}\right).$$

The contrastive weight CW of a term t is similar to $tf \cdot idf$ but uses the corpus-oriented frequencies:

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

For the term domain specificity TDS, we unify the slightly different definitions of Ahmad et al. [29], Park et al. [30], and Wong et al. [31] as

$$\mathrm{TDS}(t) = \log\left(\frac{rel_H(t) + 1}{rel_G(t) + 1} + 1\right).$$

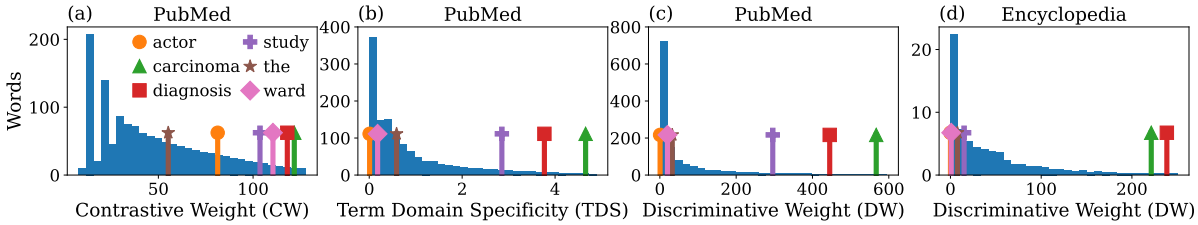

Figure 1: Termhood score frequencies of (a) CW, (b) TDS, and (c) DW on the PubMed corpus, and of (d) DW on the Encyclopedia corpus. Number of words in thousands. Example words are highlighted.

Finally, the discriminative weight DW was originally defined as the product of CW and a version of TDS that uses $freq_C(t)$ instead of $rel_C(t)$. Since the values of such an "unnormalized" TDS depend on corpus size, we use our above corpus-agnostic normalized version but still compute DW as

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

3.2. Contrastive and Health Domain Corpora

We select Wikipedia[2] as our contrastive general corpus G since it covers a wide variety of domains and is easily accessible. For health corpora H, we provide two alternatives, each with its own (dis)advantages. The first is a dump of over 30 million MEDLINE abstracts from PubMed[3]. The PubMed corpus is large-scale, reducing noise in the termhood scores. However, it mainly contains scientific language, which may not match the language of arbitrary consumer-oriented health-related web pages. For the second health corpus, we crawled the entries of five consumer-oriented medical online encyclopedias.[4–8] Since they are written in layperson's terms, we assume their language is similar to the target language distribution. However, the number of words in the Encyclopedia corpus is about three orders of magnitude smaller than in the PubMed corpus.

3.3. Pilot Inspection and Comparison

To get an impression of the scores and the corpus impact, we inspect the unigram termhood score distributions in general and for some example words. Figures 1 (a–c) show the score distributions of CW, TDS, and DW with PubMed as the domain-specific corpus H. All scores rank the depicted example out-of-domain words and the stop word lower than the health-related words. However, the assessment of CW and TDS can differ substantially for specific terms. For example, 'ward' occurs frequently within texts from the PubMed corpus, so CW attributes a rather high health relatedness. At the same time, 'ward' also occurs frequently in the general domain. The lacking "exclusiveness" leads TDS to score it relatively low. Unsurprisingly, the product score DW amplifies the extremes of both scores.

As for the effect of different health corpora, Figures 1 (c) and (d) contrast the DW scores using the PubMed corpus (rather scientific language) to using the Encyclopedia corpus (rather layperson language). As an example result, the word 'study' has a comparably high termhood score using the PubMed corpus, but is ranked like a non-health-related word using the Encyclopedia corpus.

3.4. Effectiveness and Efficiency Analyses

In previous work, we evaluated the effectiveness and efficiency of the termhood scores on a set of 1,000 manually labeled sentences [2]. We compared the termhood scores to two types of baseline approaches.

[2]Dump of all English Wikipedia articles from July 1, 2021.
[3]https://pubmed.ncbi.nlm.nih.gov/
[4]http://health.am/encyclopedia
[5]https://medlineplus.gov/encyclopedia.html
[6]https://merriam-webster.com/medical
[7]https://ucsfhealth.org (various subpages)
[8]https://www.rxlist.com/drug-medical-dictionary/article.htm

Table 1
Effectiveness and time efficiency comparison of several baseline approaches and our termhood scores on a set of 1,000 manually labeled sentences.

(a) Effectiveness on the test set as precision (P), recall (R), F1, or Matthews correlation coefficient (M). The best approach per evaluation metric is highlighted in bold.

Approach	P	R	F1	M
cTakes	0.57	0.46	0.51	0.42
MetaMap	0.42	0.49	0.45	0.33
QuickUMLS	0.49	0.49	0.49	0.38
ScispaCy	0.42	0.60	0.49	0.37
BERT	0.76	0.74	**0.75**	**0.70**
SciBERT	0.64	0.66	0.65	0.57
PubMedBERT	**0.87**	0.57	0.69	0.66
CW	0.67	0.63	0.65	0.58
TDS	0.69	0.63	0.66	0.59
DW	0.71	**0.77**	0.74	0.68

(b) Run time efficiency of the different approaches. Time per instance averaged over 10 runs, speedup computed against BERT as the most effective approach from Table 1a.

Approach	ms	Speedup
cTakes	212.12	0.2
MetaMap	120.28	0.4
QuickUMLS	8.98	5.3
ScispaCy	15.96	3.0
(PubMed / Sci) BERT	47.77	1.0
CW / TDS / DW	**1.02**	**46.8**

The first type uses a medical entity linker to find mentions of medical entities. We tested four different entity linkers: cTakes [33], MetaMap [34], QuickUMLS [35], and ScispaCy [36]. A sentence is considered health-related if the proportion of words in a sentence that are medical entities exceeds a threshold. The second type uses a pre-trained language model to predict the health-relatedness of a sentence. We fine-tuned several BERT variants [37, 38, 39] to predict if a sentence came from an in-domain (PubMed or Encyclopedia) corpus or the contrastive Wikipedia corpus. We fine-tuned hyperparameters for each approach on a development set. The effectiveness results are shown in Table 1a. We additionally evaluated the time efficiency of the different approaches. Results are shown in Table 1b.

Overall, we found the medical entity linkers to have both poor effectiveness and time efficiency. The pre-trained language models were the most effective overall, with BERT achieving the highest F1 and Matthews correlation coefficient. However, the discriminative weight termhood score was only slightly worse and about 47 times faster. The termhood scores are substantially more time efficient because inference mostly consists of looking up precomputed occurrence frequencies in a hash table and computing simple aggregations over these frequencies. The memory footprint of both models is similar at around 400 MB for the model weights and occurrence frequencies, respectively. We refer to our previous work for a more detailed discussion of the results [2].

3.5. Measuring Health Relatedness of Queries and Web Pages

Our contribution to the 1st International Workshop on Open Web Search is integrating the termhood scores into a software submission to determine the health-relatedness of queries and web pages. Our submission efficiently computes the discriminative weight for each word in a given text (e.g., a query or web page) using both the PubMed and the Encyclopedia corpus. The mean and median over all words are computed for each corpus, resulting in four nuanced termhood scores for an input text. These scores can then be used downstream in a search engine to filter or rank health-related content.

While we validated our approach on sentences (from web pages), we expect it to generalize to any arbitrary text, because the termhood scores are computed by aggregating word-level scores. The distribution of words for different texts may differ, meaning the thresholds for health-relatedness may need to be adjusted. For example, queries will usually contain fewer stop words and more nouns, which may lead to higher average termhood scores. In addition, since the discriminative weight is sensitive to corpus size (compare the y-axes of Figures 1 (c) and (d)), the thresholds may need to be adjusted when using different health corpora.

3.6. Pilot Study

Table 2: Example queries from the TREC Web tracks.

Text	Rank	DW
sore throat	1	411.7
forearm pain	2	354.0
joints	3	331.5
...		
lower heart rate	15	168.4
ct jobs	16	167.2
...		
angular cheilitis	33	76.90
getting organized	34	74.77
...		

Table 3: nDCG@10 for health and other queries.

Model	HQ	OQ
Dirichlet	0.27	0.25
BM25	0.24	0.19
MonoT5 3b	0.18	0.20
Splade	0.17	0.19

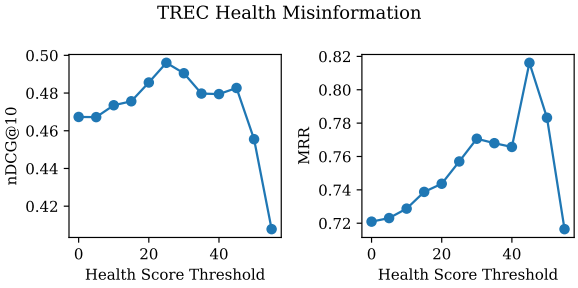

Figure 2: Effect of filtering for health-relatedness on the TREC Health Misinformation 2019 task.

To exemplify use cases of our approach and to validate its effectiveness, we conducted a small pilot study on a sample of tasks from the TIREx benchmark [40, 41]. We note that our experiments and methodology are by no means scientifically rigorous and only serve to give a general understanding of how to apply our software submission. Specifically, we consider the four TREC Web tracks from 2009 to 2012 [42, 43, 44, 45] and investigate (1) what proportion of the queries are health-related and (2) how the effectiveness of various retrieval systems differs between health-related and non-health-related queries. We also consider the TREC Health Misinformation 2019 [46] task and investigate what effect filtering for health-related documents has on the effectiveness of ranking models.

Table 2 shows example queries from the TREC Web tracks sorted by discriminative weight using the Encyclopedia corpus. The top queries are certainly health-related. Queries further down in the ranking become less health-related, with 'ct jobs' being the first we deem is not health-related. We assume that in this case, 'ct' stands for Connecticut and not computed tomography. We then divide the queries into two groups, the top 15 queries are grouped into a health category, and the remaining 185 are grouped into an other category. Some false positives and negatives are present, e.g., angular cheilitis at rank 33. However, a manual inspection of the queries showed that most are classified correctly.

Table 3 shows the nDCG@10 for two lexical and two neural-based ranking models for the two query groups. We observe that the lexical models both achieve higher effectiveness on the health queries compared to the other queries. The neural models, on the other hand, exhibit the exact opposite behavior and are more effective on the other queries than the health queries. This suggests it may be beneficial to apply different retrieval models for health-related queries.

Finally, Figure 2 shows the effect of filtering documents for health-relatedness on the TREC Health Misinformation 2019 task. The underlying idea is to remove documents that are not related to health and thereby improve ranking effectiveness. We took the ranking of BM25 and removed documents below a certain mean discriminative weight threshold. Both nDCG@10 and MRR increase with an increasing threshold up to a certain point. The optimal threshold for nDCG@10 lies at 25, improving the effectiveness by 0.03 compared to the baseline unfiltered ranking. The effect on MRR is more pronounced. The optimal threshold seems to lie at 30, increasing effectiveness by 0.05. An outlier at a threshold of 45 pushes the MRR even further. This suggests that filtering documents for health-relatedness can improve ranking effectiveness.

4. Conclusion

We presented an efficient method for determining the health-relatedness of web pages. Our method is based on contrastive termhood scores, which compare the occurrence frequencies of a word from a

health domain corpus to one or multiple out-of-domain corpora. In previous work, we found termhood scores can effectively determine the health-relatedness of phrases and sentences. We build on this work and apply termhood scores to web pages. Our software submission is an easy-to-use docker container that efficiently computes an effective and nuanced health-relatedness score for a given web page. The score is useful for various downstream tasks, such as filtering or ranking web pages in a search engine.

References

[1] S. Farzana, M. Fröbe, M. Granitzer, G. Hendriksen, D. Hiemstra, M. Potthast, S. Zerhoudi, 1st International Workshop on Open Web Search (WOWS), in: Advances in Information Retrieval. 46th European Conference on IR Research (ECIR 2024), Lecture Notes in Computer Science, Springer, 2024.

[2] F. Schlatt, D. Bettin, M. Hagen, B. Stein, M. Potthast, Mining Health-related Cause-Effect Statements with High Precision at Large Scale, in: N. Calzolari, C. Huang, H. Kim, J. Pustejovsky, L. Wanner, K. Choi, P. Ryu, H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S. Na (Eds.), 29th International Conference on Computational Linguistics (COLING 2022), International Committee on Computational Linguistics, 2022, pp. 1925–1936. URL: https://aclanthology.org/2022.coling-1.167.

[3] J. A. Diaz, R. A. Griffith, J. J. Ng, S. E. Reinert, P. D. Friedmann, A. W. Moulton, Patients' use of the internet for medical information, Journal of General Internal Medicine 17 (2002) 180–185. URL: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1525-1497.2002.10603.x. doi:10.1046/j.1525-1497.2002.10603.x.

[4] G. Eysenbach, J. Powell, O. Kuss, E.-R. Sa, Empirical studies assessing the quality of health information for consumers on the world wide web: A systematic review, JAMA 287 (2002) 2691–2700. URL: https://doi.org/10.1001/jama.287.20.2691. doi:10.1001/jama.287.20.2691.

[5] Y. Zhang, Y. Sun, B. Xie, 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 (2015) 2071–2084. URL: https://doi.org/10.1002/asi.23311. doi:10.1002/asi.23311.

[6] Y. Aphinyanaphongs, C. Aliferis, Text Categorization Models for Identifying Unproven Cancer Treatments on the Web, Studies in Health Technology and Informatics 129 (2007) 968–972.

[7] A. Abbasi, F. M. Zahedi, S. Kaza, Detecting Fake Medical Web Sites Using Recursive Trust Labeling, ACM Transactions on Information Systems 30 (2012) 22:1–22:36. doi:10.1145/2382438.2382441.

[8] C. Boyer, L. Dolamic, Automated Detection of HONcode Website Conformity Compared to Manual Detection: An Evaluation, Journal of Medical Internet Research 17 (2015) e3831. doi:10.2196/jmir.3831.

[9] C. Boyer, C. Frossard, A. Gaudinat, A. Hanbury, G. Falquetd, How to sort trustworthy health online information? improvements of the automated detection of HONcode criteria, Procedia Computer Science 121 (2017) 940–949. doi:10.1016/j.procs.2017.11.122.

[10] K. Kageura, B. Umino, Methods of automatic term recognition: A review, Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 3 (1996) 259–289. URL: https://www.jbe-platform.com/content/journals/10.1075/term.3.2.03kag. doi:10.1075/term.3.2.03kag.

[11] Y.-L. Jiang, Quality evaluation of orthodontic information on the world wide web, American Journal of Orthodontics and Dentofacial Orthopedics 118 (2000) 4–9. URL: https://www.sciencedirect.com/science/article/pii/S0889540600337490. doi:10.1067/mod.2000.104492.

[12] B. Brennan, G. Kanayama, H. Pope, Performance-enhancing drugs on the web: A growing public-health issue, The American Journal on Addictions 22 (2013) 158–161. URL:

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1521-0391.2013.00311.x. doi:10.1111/j.1521-0391.2013.00311.x.

[13] B. E. J. Cooper, W. E. Lee, B. M. Goldacre, T. A. B. Sanders, The quality of the evidence for dietary advice given in UK national newspapers, Public Understanding of Science 21 (2012) 664–673. URL: https://doi.org/10.1177/0963662511401782. doi:10.1177/0963662511401782.

[14] A. Yavchitz, I. Boutron, A. Bafeta, I. Marroun, P. Charles, J. Mantz, P. Ravaud, Misrepresentation of randomized controlled trials in press releases and news coverage: A cohort study, PLOS Medicine 9 (2012) e1001308. URL: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001308. doi:10.1371/journal.pmed.1001308.

[15] P. Sumner, S. Vivian-Griffiths, J. Boivin, A. Williams, C. A. Venetis, A. Davies, J. Ogden, L. Whelan, B. Hughes, B. Dalton, F. Boy, C. D. Chambers, The association between exaggeration in health related science news and academic press releases: Retrospective observational study, BMJ 349 (2014) g7015. doi:10.1136/bmj.g7015.

[16] A. Bondarenko, E. Shirshakova, M. Driker, M. Hagen, P. Braslavski, Misbeliefs and biases in health-related searches, in: Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM 2021), Virtual Event, Queensland, Australia, November 1–5, 2021, 2021, pp. 2894–2899. URL: https://doi.org/10.1145/3459637.3482141. doi:10.1145/3459637.3482141.

[17] V. Suarez-Lledo, J. Alvarez-Galvez, Prevalence of health misinformation on social media: Systematic review, Journal of Medical Internet Research 23 (2021) e17187. URL: https://www.jmir.org/2021/1/e17187. doi:10.2196/17187.

[18] D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. Quinn, M. Dredze, Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate, American Journal of Public Health 108 (2018) 1378–1384. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6137759/. doi:10.2105/AJPH.2018.304567.

[19] R. Bal, S. Sinha, S. Dutta, R. Joshi, S. Ghosh, R. Dutt, Analysing the extent of misinformation in cancer related tweets, in: Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020), Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8–11, 2020, 2020, pp. 924–928. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/7359.

[20] W. M. Silberg, G. D. Lundberg, R. A. Musacchio, Assessing, Controlling, and Assuring the Quality of Medical Information on the Internet: Caveant Lector et Viewor—Let the Reader and Viewer Beware, JAMA 277 (1997) 1244–1245. doi:10.1001/jama.1997.03540390074039.

[21] D. Charnock, S. Shepperd, G. Needham, R. Gann, DISCERN: An instrument for judging the quality of written consumer health information on treatment choices., Journal of Epidemiology & Community Health 53 (1999) 105–111. doi:10.1136/jech.53.2.105.

[22] C. Boyer, V. Baujard, A. Geissbuhler, Evolution of health web certification through the HONcode experience, Studies in Health Technology and Informatics 169 (2011) 53. doi:10.3233/978-1-60750-806-9-53.

[23] L. Daraz, A. S. Morrow, O. J. Ponce, W. Farah, A. Katabi, A. Majzoub, M. O. Seisa, R. Benkhadra, M. Alsawas, P. Larry, M. H. Murad, Readability of online health information: A meta-narrative systematic review, American Journal of Medical Quality 33 (2018) 487–492. doi:10.1177/1062860617751639.

[24] C. R. Watters, W. Zheng, E. E. Milios, Filtering for medical news items, in: Proceedings of the 65th ASIS&T Annual Meeting (ASIST 2002), Philadelphia, PA, USA, November 18–21, 2002, volume 39, 2002, pp. 284–291. URL: https://doi.org/10.1002/meet.1450390131. doi:10.1002/meet.1450390131.

[25] W. Zheng, E. E. Milios, C. R. Watters, Filtering for medical news items using a machine learning approach, in: Proceedings of the American Medical Informatics Association Annual Symposium (AMIA 2002), San Antonio, TX, USA, November 9–13, 2002, 2002. URL: https://knowledge.amia.org/amia-55142-a2002a-1.610020/t-001-1.612667/f-001-1.612668/a-191-1.612675/a-192-1.612672.

[26] G. Gkotsis, A. Oellrich, S. Velupillai, M. Liakata, T. J. P. Hubbard, R. J. B. Dobson, R. Dutta,

Characterisation of mental health conditions in social media using informed deep learning, Scientific Reports 7 (2017) 45141. URL: https://www.nature.com/articles/srep45141. doi:10.1038/srep45141.

[27] R. Basili, A. Moschitti, M. T. Pazienza, F. M. Zanzotto, A contrastive approach to term extraction, in: Proceedings of Terminologie et Intelligence Artificielle (TIA 2001), Nancy, France, May 3–4, 2001, 2001, pp. 119–128. URL: http://disi.unitn.it/~moschitti/articles/TIA2000.pdf.

[28] S. N. Kim, T. Baldwin, M. Kan, An unsupervised approach to domain-specific term extraction, in: Proceedings of the Australasian Language Technology Association Workshop (ALTA 2009), Sydney, Australia, December 3–4, 2009, 2009, pp. 94–98. URL: https://aclanthology.org/U09-1013/.

[29] K. Ahmad, L. Gillam, L. Tostevin, University of Surrey participation in TREC8: Weirdness indexing for logical document extrapolation and retrieval (WILDER), in: Proceedings of The Eighth Text REtrieval Conference (TREC 1999), Gaithersburg, Maryland, USA, November 17–19, 1999, 1999. URL: http://trec.nist.gov/pubs/trec8/papers/surrey2.pdf.

[30] Y. Park, S. Patwardhan, K. Visweswariah, S. C. Gates, An empirical analysis of word error rate and keyword error rate, in: Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008), Brisbane, Australia, September 22–26, 2008, 2008, pp. 2070–2073. URL: http://www.isca-speech.org/archive/interspeech_2008/i08_2070.html.

[31] W. Wong, W. Liu, M. Bennamoun, Determining termhood for learning domain ontologies using domain prevalence and tendency, in: Proceedings of the Sixth Australasian Data Mining Conference (AusDM 2007), Gold Coast, Queensland, Australia, December 3–4, 2007, 2007, pp. 47–54. URL: http://crpit.scem.westernsydney.edu.au/abstracts/CRPITV70Wong.html.

[32] F. Bonin, F. Dell'Orletta, S. Montemagni, G. Venturi, A contrastive approach to multi-word extraction from domain-specific corpora, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta, May 17–23, 2010, 2010. URL: http://www.lrec-conf.org/proceedings/lrec2010/summaries/553.html.

[33] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, C. G. Chute, Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications, Journal of the American Medical Informatics Association 17 (2010) 507–513. URL: https://doi.org/10.1136/jamia.2009.001560. doi:10.1136/jamia.2009.001560.

[34] A. R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program, in: Proceedings of the American Medical Informatics Association Annual Symposium (AMIA 2001), Washington, DC, USA, November 3–7, 2001, 2001. URL: https://knowledge.amia.org/amia-55142-a2001a-1.597057/t-001-1.599654/f-001-1.599655/a-003-1.600128/a-004-1.600125.

[35] L. Soldaini, N. Goharian, QuickUMLS: A fast, unsupervised approach for medical concept extraction, in: Proceedings of the 2nd SIGIR Workshop on Medical Information Retrieval (MedIR 2016), Pisa, Italy, July 21, 2016, 2016. URL: http://medir2016.imag.fr/data/MEDIR_2016_paper_16.pdf.

[36] M. Neumann, D. King, I. Beltagy, W. Ammar, ScispaCy: Fast and robust models for biomedical natural language processing, in: Proceedings of the 18th BioNLP Workshop and Shared Task (BioNLP@ACL 2019), Florence, Italy, August 1, 2019, 2019, pp. 319–327. URL: https://doi.org/10.18653/v1/w19-5034. doi:10.18653/v1/w19-5034.

[37] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019), Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[38] I. Beltagy, K. Lo, A. Cohan, 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 2019), Hong Kong, China, November 3–7, 2019, 2019, pp. 3613–3618. URL: https://doi.org/10.18653/v1/D19-1371. doi:10.18653/v1/D19-1371.

[39] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon,

Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare 3 (2022) 2:1–2:23. URL: https://doi.org/10.1145/3458754. doi:10.1145/3458754.

[40] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. URL: https://link.springer.com/chapter/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.

[41] M. Fröbe, J. Reimer, S. MacAvaney, N. Deckers, S. Reich, J. Bevendorff, B. Stein, M. Hagen, M. Potthast, The Information Retrieval Experiment Platform, in: H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe, B. Poblete (Eds.), 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023), ACM, 2023, pp. 2826–2836. URL: https://dl.acm.org/doi/10.1145/3539618.3591888. doi:10.1145/3539618.3591888.

[42] C. L. A. Clarke, N. Craswell, I. Soboroff, Overview of the TREC 2009 Web Track., in: Proceedings of TREC 2009, volume 500–278 of *NIST Special Publication*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 2009, p. 9. URL: http://trec.nist.gov/pubs/trec18/papers/WEB09.OVERVIEW.pdf.

[43] C. L. A. Clarke, N. Craswell, I. Soboroff, G. V. Cormack, Overview of the TREC 2010 Web Track., in: Proceedings of TREC 2010, volume 500–294 of *NIST Special Publication*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 2010, p. 9. URL: https://trec.nist.gov/pubs/trec19/papers/WEB.OVERVIEW.pdf.

[44] C. L. A. Clarke, N. Craswell, I. Soboroff, E. M. Voorhees, Overview of the TREC 2011 Web Track., in: Proceedings of TREC 2011, volume 500–296 of *NIST Special Publication*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 2011, p. 9. URL: http://trec.nist.gov/pubs/trec20/papers/WEB.OVERVIEW.pdf.

[45] C. L. A. Clarke, N. Craswell, E. M. Voorhees, Overview of the TREC 2012 Web Track., in: Proceedings of TREC 2012, volume 500–298 of *NIST Special Publication*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 2012, p. 8. URL: http://trec.nist.gov/pubs/trec21/papers/WEB12.overview.pdf.

[46] M. Abualsaud, C. Lioma, M. Maistro, M. Smucker, G. Zuccon, Overview of the TREC 2019 Decision Track., in: Proceedings of TREC 2019, volume 500–331 of *NIST Special Publication*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, 2019, p. 19. URL: https://trec.nist.gov/pubs/trec28/papers/OVERVIEW.D.pdf.