

DS4DH at TREC Health Misinformation 2021: Multi-Dimensional Ranking Models with Transfer Learning and Rank Fusion

Boya Zhang¹, Nona Naderi^{2,3}, Fernando Jaume-Santero^{1,2}, and Douglas Teodoro^{1,2,3}

¹Department of Radiology and Medical Informatics, University of Geneva
<first>.<last>@unige.ch

²HES-SO University of Applied Sciences and Arts of Western Switzerland
<first>.<last>@hesge.ch

³Swiss Institute of Bioinformatics

Abstract

This paper describes the work of the Data Science for Digital Health (DS4DH) group at the TREC Health Misinformation Track 2021. The TREC Health Misinformation track focused on the development of retrieval methods that provide relevant, correct and credible information for health related searches on the Web. In our methodology, we used a two-step ranking approach that includes *i*) a standard retrieval phase, based on BM25 model, and *ii*) a re-ranking phase, with a pipeline of models focused on the usefulness, supportiveness and credibility dimensions of the retrieved documents. To estimate the usefulness, we classified the initial rank list using pre-trained language models based on the transformers architecture fine-tuned on the MS MARCO corpus. To assess the supportiveness, we utilized BERT-based models fine-tuned on scientific and Wikipedia corpora. Finally, to evaluate the credibility of the documents, we employed a random forest model trained on the Microsoft Credibility dataset combined with a list of credible sites. The resulting ranked lists were then combined using the Reciprocal Rank Fusion algorithm to obtain the final list of useful, supporting and credible documents. Our approach achieved competitive results, being top-2 in the compatibility measurement for the automatic runs. Our findings suggest that integrating automatic ranking models created for each information quality dimension with transfer learning can increase the effectiveness of health-related information retrieval.

1 Introduction

The purpose of the TREC Health Misinformation Track¹ is to develop retrieval systems that provide relevant and correct information for health-related Web searches. The challenge provides 50 health-related topics, among which only 35 topics are

¹<https://trec-health-misinfo.github.io/>

evaluated. As illustrated in Table 1, each topic contains a query, a description, a narrative, a stance and an evidence. For automatic runs, only the query and the description are used. The TREC Health Misinformation corpus contains one billion English documents extracted from the April 2019 snapshot of Common Crawl (Raffel et al., 2020).

For a document to be relevant and correct, the TREC Health Misinformation challenge considers three levels of information quality: (1) *usefulness*, that is, whether a document contains relevant information to answer a topic’s question; (2) *supportiveness*, that is, whether a document contains supportive information for the descriptions marked as helpful or dissuasive information for the descriptions marked as unhelpful; and (3) *credibility*, that is, whether an information source document is considered credible in the field of knowledge.

This paper describes submissions of DS4DH to the TREC Health Misinformation 2021, which achieved top-2 performance in the compatibility assessment for the automatic runs. Section 2 introduces our methodology based on a two-step approach of document retrieval for the multi-dimensional retrieval task and the specific evaluation criteria used to assess these dimensions with a single ranked list. Section 3 presents and discusses our results. Finally, Section 4 concludes this paper and proposes future studies.

2 Methods

The overview of our pipeline is presented in Figure 1. First, in the retrieval phase, we extracted 10,000 documents using a BM25 model (Robertson and Zaragoza, 2009). Second, in the re-ranking phase, we estimated relevance of the documents according to the usefulness, supportiveness and credibility scores. At the end, 7 runs with different model combinations were submitted.

Topic 101	
Query	ankle brace achilles tendonitis
Description	Will wearing an ankle brace help heal achilles tendonitis?
Narrative	Achilles tendonitis is a condition where one experiences pain in the Achilles tendon located near the heel. An ankle brace is usually worn around the ankles to protect and limit movement. A very useful document would discuss the effectiveness of using ankle braces to help heal Achilles tendonitis. A useful document would help a user make a decision about the use of ankle braces for treating tendonitis by providing information about recommended treatments for Achilles tendonitis, ankle braces, or both.
Stance	unhelpful
Evidence	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3134723/

Table 1: Topic Example.

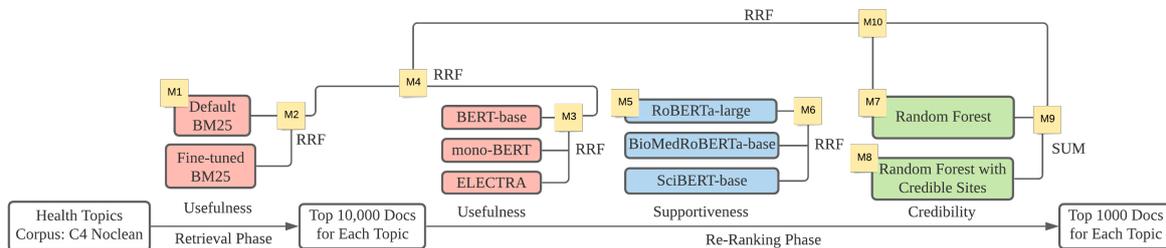


Figure 1: Retrieval Pipeline. M1: BM25 baseline. M2: M1 combined with a fine-tuned index using a silver standard query relevance. M3: a re-ranked list using masked language models (MLMs) fine-tuned on the MS MARCO dataset. M4: a combination BM25 and MLMs. M5: a RoBERTa model fine-tuned on the FEVER+SciFact fact-checking datasets. M6: A combination of three MLMs trained on the FEVER+SciFact fact-checking corpora. M7: a random forest model trained on the Microsoft Credibility dataset to predict a site’s credibility. M8: a list of credible websites scrapped from the Health-on-Net search engine for the challenge’s queries. M9: a linear combination of M7 and M8 models. M10: a combination of model M4, M6 and M9. All ranking combinations, apart from M9, were created using the reciprocal ranking fusion (RRF) algorithm with the k parameter set to 60.

2.1 Retrieval Phase

We extracted 10,000 documents using a BM25 model with standard (Robertson and Zaragoza, 2009) and fine-tuned parameters. For the fine-tuning version, as the query relevance was not available, we created topics, that is, query + description, for a set of indexed documents using the a keyword2query and doc2query approaches proposed by Bennani-Smires *et al.* (Bennani-Smires *et al.*, 2018) and Nogueira *et al.* (Nogueira *et al.*, 2019b), respectively. Then, a known-item search approach was applied using the silver topics and the BM25 parameters was fine-tuned using a grid search. This resulted in two initial ranking lists.

2.2 Re-Ranking Phase

For re-ranking the retrieved documents, we use a set of machine learning models trained to classify documents according to the usefulness, supportive-

ness, and credibility criteria.

Usefulness To improve the usefulness dimension of the retrieved documents, we implemented re-ranking models based pre-trained language models fine-tuned on the MS MARCO dataset: BERT-base (Li *et al.*, 2020), mono-BERT-large (Nogueira *et al.*, 2019a) and ELECTRA (Clark *et al.*, 2020). While BM25 provides a strong baseline for usefulness, it does not consider the relation and context of words. Thus, the pre-trained language models are used to enhance the quality of the original ranking (Teodoro *et al.*, 2021) as similarly shown to improve other natural language processing tasks like named entity recognition (Naderi *et al.*, 2021). Given a topic and a document, the language model infers whether the document is relevant or not to the topic.

Supportiveness In this information quality dimension, documents are identified under three levels: 1) *supportive* - the document supports the treatment; 2) *dissuades* - the document refutes the treatment; 3) *neutral* - the document does not contain enough information to make the decision. We want documents that are either supportive or dissuasive on the top of the ranking list, which means that correct or factual documents are boosted and misinforming documents should be downgraded.

The supportiveness dimension shares similarities with claim-checking models, which take a claim and a document as the information source, and *validate* or *refute* the claim based on document content (Stambach et al., 2021). Their main difference is that for claim-checking models, we assume that the information source is always correct. Thus, for the supportiveness criterion, we add a further classification step, which evaluates the documents as correct or incorrect. To do so, we used a k-nearest neighbors algorithm (Teodoro et al., 2010) based on the top-k assignments provided by the claim-checking models, that is, a majority vote is used to decide whether the treatment should be supported or dissuaded. Then, higher rank is given to the correct supportive/dissuasive documents, medium rank is given to the neutral documents and lower rank is given to the incorrect supportive/dissuasive documents. The details are shown in Figure 2 and Table 2.

We used three models from the Scientific Claim Verification task (Wadden et al., 2020) to classify the treatments: RoBERTa-Large (Liu et al., 2019), BioMedRoBERTa-base (Gururangan et al., 2020) and SciBERT-base (Beltagy et al., 2019). These models were trained on either scientific or large English corpora and fine-tuned on the FEVER (Thorne et al., 2018) and SciFact (Wadden et al., 2020) datasets. The information that these models learned from previous corpora benefits our ranking task through transfer learning.

Credibility For estimating credibility, we develop a random forest classifier trained on the Microsoft Credibility dataset (Schwarz and Morris, 2011) with a set of features, such as readability, openpage rank² and number of CSS style sheets. The dataset consists of 1,000 Web pages on five topics of Health, Politics, Finance, Environmental Science, and Celebrity News. The Web pages are manually rated with credibility scores between 1

("very non-credible") and 5 ("very credible").³ We convert these scores for a binary classification setting – that is, the scores of 4 and 5 are considered as 1 or *credible* and scores of 1, 2, and 3 are considered as 0 or *non-credible*. For the readability score, we rely on SMOG index, which estimates the years of education an average person needs to understand a piece of writing. Following (Schwarz and Morris, 2011), we retrieve a Web page’s PageRank and use it as a feature to train the classifier. We further use the number of CSS style definitions for estimating the effort for the design of a Web page (Olteanu et al. (2013) showed the effectiveness of this feature). Furthermore, a list of credible websites scrapped from the Health On the Net search engine⁴ for the challenge’s queries is combined with the baseline model to explore better performance. The result of the classifier was added with a unitary value for the Health on the Net credible sites.

2.3 Submitted Runs

Run 1: Baseline run. A combination of M1, M5 and M7 in Figure 1. This automatic run was created using a model based on Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) of three models: i) usefulness, created using a default BM25, ii) supportiveness, created using a RoBERTa large model fine-tuned on the FEVER and SciFact corpus, and iii) credibility, created using a credibility random forest classifier as described in Section 2.

Run 2: A combination of M2, M6 and M9 in Figure 1. This automatic run was created using a rank fusion based on RRF of three models: i) usefulness, created using a combined default BM25 with a fine-tuned BM25 model using known item search with query and description been generated using transfer learning from language models, ii) supportiveness, created using a combined rank of three transformer-based models fine-tuned on the FEVER and SciFact corpus, and iii) credibility, created using a random forest classifier trained on the Microsoft Credibility dataset combined with a list of credible sites.

Run 3: A combination of M3, M6 and M9 in Figure 1. This automatic run was created using a rank fusion based on RRF of three models: i)

³A credible Web page is defined as "a page whose information one can accept as the truth without needing to look elsewhere."

⁴www.hon.ch

²<https://www.domcop.com/openpagerank/documentation>

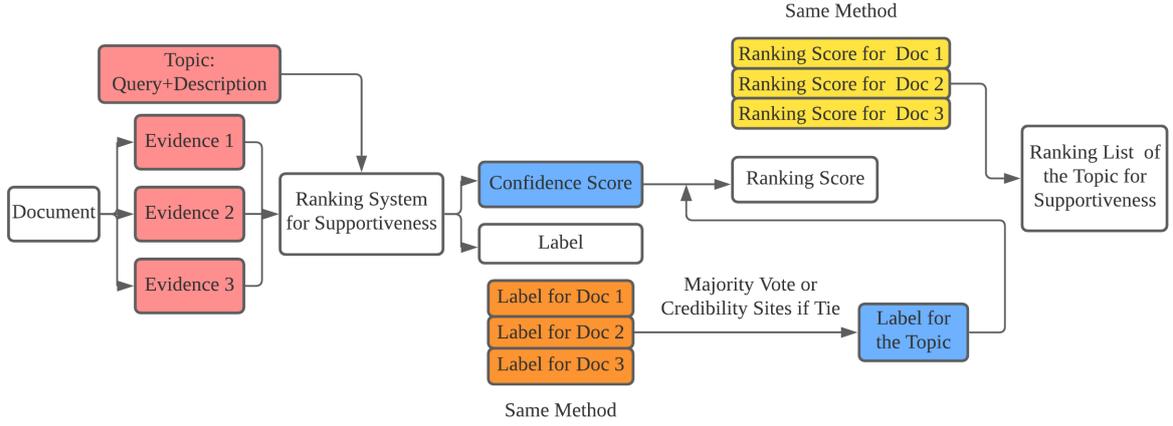


Figure 2: Re-Ranking Pipeline for Supportiveness. When the supportive and dissuasive documents are for one topic, we use the majority vote to decide the correct label. The document ranking score of the incorrect side becomes negative. We bring in the credibility sites when tie.

Evidences (E_1, \dots, E_i)			Document (D_j)		
Supports	Dissuades	Neutral	Label	Confidence Score	Ranking Score
✓	✗		Supports	S_{max}	S_{max} or $-S_{max}$
✗	✓		Dissuades	S_{max}	S_{max} or $-S_{max}$
✗	✗	✓		$1 - \frac{1}{N} \sum_{i=1}^N S_i$	$1 - \frac{1}{N} \sum_{i=1}^N S_i$
✗	✗	✗		-3	-3
✓	✓			-2	-2

Table 2: Ranking Score for Supportiveness. When the document only includes neutral evidence, the document confidence score is $1 - \frac{1}{N} \sum_{i=1}^N S_i$, where S_1, S_2, \dots, S_i is the confidence score of each evidence in the document. When the document contains both supports and dissuades evidences without considering the neutral cases, the document confidence score is -2. When evidences are either support or dissuades without considering the neutral cases, the document confidence score is S_{max} , where S_{max} is highest among the confidence score of each evidence in the document. When the qualified evidence is void, the document confidence score is -3.

usefulness, created using a combination of three transformed-based language models trained on the MS MARCO corpus⁵, ii) supportiveness, created using a combined rank of three transformer-based models fine-tuned on the FEVER and SciFact corpus, and iii) credibility, created using a random forest model trained on the Microsoft Credibility dataset combined with a list of credible sites.

Run 4: A combination of M4, M5 and M9 in Figure 1. This automatic run was created using a rank fusion based on RRF of three models: i) usefulness, created using a combined BoW model with three transformed-based language models trained on the MS MARCO corpus, ii) supportiveness, created using a RoBERTa large model fine-tuned on the FEVER and SciFact corpus, and iii) credibility, created using a random forest model trained on the

Microsoft Credibility dataset combined with a list of credible sites.

Run 5: A combination of M4, M6 and M7 in Figure 1. This automatic run was created using a rank fusion based on RRF of three models: i) usefulness, created using a combined BoW model with three transformed-based language models trained on the MS MARCO corpus, ii) supportiveness, created using a combined rank of three transformer-based models fine-tuned on the FEVER and SciFact corpus, and iii) credibility, create using a random forest model trained on the Microsoft Credibility dataset.

Run 6: A combination of M4, M6 and M9 in Figure 1. This automatic run was created using a rank fusion based on RRF of three models: i) usefulness, created using a combined BoW model with three transformed-based language models trained on the

⁵<https://microsoft.github.io/msmarco/>

MS MARCO corpus, ii) supportiveness, created using a combined rank of three transformer-based models fine-tuned on the FEVER and SciFact corpus, and iii) credibility, created using a random forest classifier combined with a list of credible sites.

Run 7: A combination of all the individual models in Figure 1. This automatic run was created using a rank fusion based on RRF of the individual models used to create the i) usefulness (5 individual models), ii) supportiveness (3 individual models), and iii) credibility (2 individual models).

2.4 Evaluation Methods

The assessments of this year’s TREC Health Misinformation were divided into “compatibility with helpful” (help) and “compatibility with harmful” (harm) metrics. In order to evaluate the system’s ability to have higher levels for helpful information and lower levels for harmful information, the “compatibility with harmful” results were subtracted from the “compatibility with helpful” results, which is marked as “help-harm”. This orders systems correctly with similar helpful compatibility and lower harmful compatibility. For more details, please see (Clarke et al., 2021).

3 Results and Discussion

The official challenge results are shown in Table 3 and Figure 3. The ‘Best Help’, ‘Best Harm’ and ‘Best Hp-Hm’ are the highest “compatibility with helpful” (help), lowest “compatibility with harmful” (harm) and highest “help-harm” from the list of top-3 automatic runs from each group. Note that ‘Run 1’ to ‘Run 7’ are runs submitted by the DS4DH group. ‘BM25 Baseline’ is the baseline model by the TREC Health Misinformation.

In our submitted runs, the best harm is in ‘Run 4’, the best help and help-harm are both in ‘Run 7’. This is as expected since Run 7 combines all the individual models to maximize the system ranking ability. The helpful compatibility is 0.136 which is comparable to the ‘Best Help’, and the harmful compatibility 0.095 is significantly low, which is similar to the ‘Best Harm’. Therefore, we obtained help-harm at 0.041, which is right after the overall ‘Best Hp-Hm’, together with superior performance on lower ranking harmful information.

The ‘Best Help’ run achieved in the challenge gives the helpful compatibility at 0.203. However, for this run, the harmful compatibility is also as

Run	Help	Harm	Help-Harm
Best Help	0.203	0.168	0.034
Best Harm	0.006	0.022	-0.016
Best Hp-Hm	0.195	0.153	0.043
Run 1	0.107	0.101	0.006
Run 2	0.103	0.089	0.014
Run 3	0.101	0.086	0.015
Run 4	0.108	0.076	0.032
Run 5	0.093	0.094	-0.001
Run 6	0.098	0.089	0.009
Run 7	0.136	0.095	0.041
BM25 Baseline	0.122	0.144	-0.022

Table 3: Averaged Compatibility of Submitted Runs.

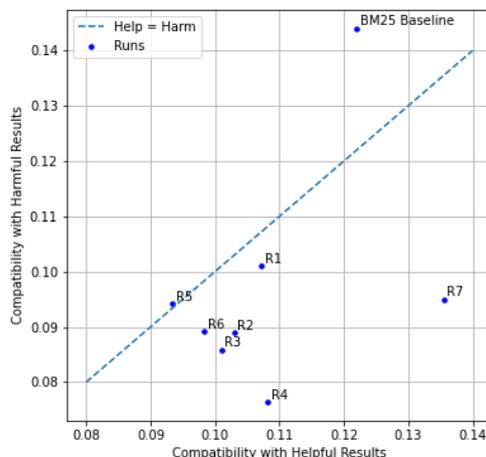


Figure 3: Relations between ‘Help’ and ‘Harm’ for Submitted Runs.

high as 0.168. Therefore, the help-harm is 0.034, which can distinguish the helpful and harmful information but needs improvement by lower ranking the harmful one. The ‘Best Harm’ gives the harmful compatibility at 0.022, which is significantly lower than other runs. However, the helpful compatibility for this run is only 0.006, resulting in a help-harm of -0.016, in a sense that neither helpful nor harmful information is retrieved. On the other hand, the ‘Best Hp-Hm’ gives relatively higher helpful compatibility at 0.195 and relatively lower compatibility at 0.153. In the end, the best help-harm 0.043 is obtained.

3.1 Models Analysis

To understand the performance of each component, we study the compatibility of individual models. The results are depicted in Table 4 and Figure 4. The BM25 model for usefulness in the retrieval phase gives highest helpful compatibil-

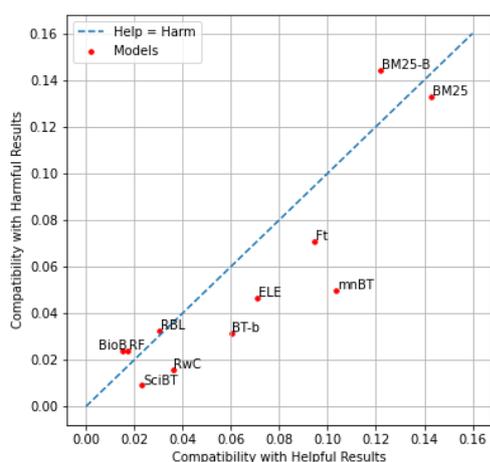


Figure 4: Relations between ‘Help’ and ‘Harm’ for Individual Models. ELE: ELECTRA. RWC: Random Forest with Credibility Sites. mnBT: monoBERT. RF: Random Forest. BioB: BioMedRoBERTa-base. RBL: RoBERTa-Large. SciBT: SciBERT. BM25: Standard BM25. Ft: Fine-tuned BM25. BM25-B: BM25 Baseline by TREC Health Misinformation

ity with 0.143. This is beneficial since at this phase, overall compatibility is more critical. The monoBERT model for usefulness in the re-ranking phase gives highest help-harm compatibility at 0.053, which is the top-1 result compared to the automatic runs submitted by the participants. This indicates that the model can effectively differentiate between helpful and harmful information. The SciBERT-base model for supportiveness gives lowest harmful compatibility at 0.009 with satisfactory helpful compatibility. This demonstrates the model’s ability to identify misinformation.

As a result of the foregoing analysis, further experiments could be explored: 1) in the document retrieval phase, BM25 should be the prior model. 2) in the re-ranking phase: prioritising the monoBERT for usefulness, the SciBERT-base model for supportiveness, the ‘Random Forest with Credibility Sites’ model for credibility.

4 Conclusion

In the average compatibility measurement of automatic runs, our contributions came out on top-2. Our findings imply that combining automatic ranking models for each information quality dimension with transfer learning can improve the quality of health-related information retrieval by allowing the proper documents to be retrieved while discarding

the incorrect ones. The RRF algorithm is a robust alternative for combining ranks when no trained set is available. Further empirical approaches could be to fine-tune our models based on topic content and generate manual runs and to re-rank the top 10% of the retrieved documents to reduce the possibility of bringing the harmful documents from the bottom of the ranking list to the top.

References

- 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)*, pages 3615–3620.
- Kamil Bennani-Smires, Claudiu Musat, Andreea Hossman, Michael Baeriswyl, and Martin Jaggi. 2018. Simple unsupervised keyphrase extraction using sentence embeddings. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 221–229.
- Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.
- Charles LA Clarke, Alexandra Vtyurina, and Mark D Smucker. 2021. Assessing top-preferences. *ACM Transactions on Information Systems (TOIS)*, 39(3):1–21.
- Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.
- Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. 2020. Parade: Passage representation aggregation for document reranking. *arXiv preprint arXiv:2008.09093*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Nona Naderi, Julien Knafou, Jenny Copara, Patrick Ruch, and Douglas Teodoro. 2021. Ensemble of deep masked language models for effective named entity

	Individual Model	Help	Harm	Help-Harm
Usefulness (Retrieval)	BM25	0.143	0.133	0.010
	Fine-tuned BM25	0.095	0.071	0.024
Usefulness (Re-Ranking)	BERT-base	0.060	0.031	0.029
	monoBERT	0.103	0.050	0.053
	ELECTRA	0.071	0.047	0.024
Supportiveness	BioMedRoBERTa-base	0.015	0.024	-0.009
	RoBERTa-large	0.031	0.033	-0.002
	SciBERT-base	0.023	0.009	0.014
Credibility	RandomForest	0.017	0.024	-0.007
	RFwithCredibilitySites	0.036	0.016	0.020
TREC HM	BM25 Baseline	0.122	0.144	-0.022

Table 4: Averaged Compatibility of Individual Models.

recognition in health and life science corpora. *Frontiers in Research Metrics and Analytics*, 6:689803–689803.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019a. Multi-stage document ranking with bert. *arXiv e-prints*, pages arXiv–1910.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019b. Document expansion by query prediction. *arXiv e-prints*, pages arXiv–1904.

Alexandra Olteanu, Stanislav Peshterliev, Xin Liu, and Karl Aberer. 2013. Web credibility: Features exploration and credibility prediction. In *European conference on information retrieval*, pages 557–568. Springer.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Julia Schwarz and Meredith Morris. 2011. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1245–1254.

Dominik Stammbach, Boya Zhang, and Elliott Ash. 2021. The choice of knowledge base in automated claim checking. *arXiv*, pages 2111–07795.

Douglas Teodoro, Sohrab Ferdowsi, Nikolay Borissov, Elham Kashani, David Vicente Alvarez, Jenny Copara, Racha Gouareb, Nona Naderi, Poorya Amini, et al. 2021. Information retrieval in an infodemic: the case of covid-19 publications. *Journal of medical Internet research*, 23(9):e30161.

Douglas Teodoro, Julien Gobeill, Emilie Pasche, P Ruch, D Vishnyakova, and Christian Lovis. 2010. Automatic ipc encoding and novelty tracking for effective patent mining. In *The 8th NTCIR Workshop*

Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550.

A Additional Illustration for Compatibility Measurement



Figure A.1: Runs Evaluated with Compatibility.

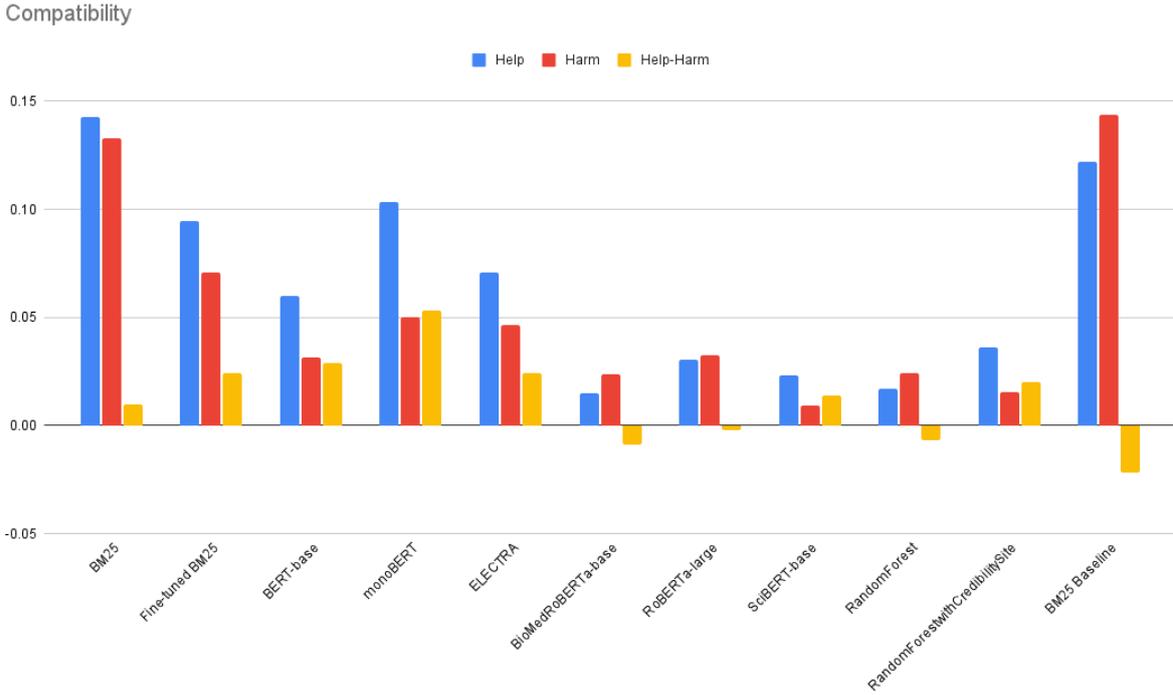


Figure A.2: Individual Models Evaluated with Compatibility.