NLM at TREC 2020 Health Misinformation and Deep Learning Tracks

Yassine Mrabet, Mourad Sarrouti, Asma Ben Abacha, Soumya Gayen, Travis Goodwin, Alastair Rae, Will Rogers, and Dina Demner-Fushman

National Library of Medicine
8600 Rockville Pike, Bethesda, MD, 20894

Abstract. This paper describes the participation of the National Library of Medicine to TREC 2020. Our main focus was the health misinformation track. We also participated to the Deep Learning track to both evaluate and enhance our deep re-ranking baselines for information retrieval. Our methods include a wide variety of approaches, ranging from conventional Information Retrieval (IR) models, neural re-ranking models, Natural Language Inference (NLI) models, Claim-Truth models, hyperlinks-based scores such as Page Rank and HITS, and ensemble methods.

1 Health Misinformation Track

With the fast pace of online content publication, misinformation about COVID-19 and the new coronavirus proved difficult to track and debunk at scale. The health misinformation track at TREC 2020 tackles this issue through an international challenge on the automatic recognition of misinformation from the web using a crawl of new articles published between January and April 2020[1] as a reference dataset.

The challenge relies on a set of 46 questions about COVID-19 and their reference yes/no answer. Two tasks are considered. The first Total Recall task focuses on misinformation and requires participating systems to rank documents promulgating misinformation first. The second Ad-hoc task tackles the retrieval of relevant, correct, and credible information first.

For our participation, we first parsed the target Common Crawl News collection and used a combination of the Optimaize language detector[2] and an ASCII character ratio threshold to keep only documents written in English.

We indexed the filtered documents at two different levels of granularity: (1) document-level indexing and (2) sentence level indexing. We applied different conventional information retrieval models to retrieve either the top 10000 or top 1000 documents, as well as relevance-based T5 and BERT re-ranking models, and rank-based ensembles with the different approaches. Figure 1 presents an overview of our data pipeline, approaches and workflow.

[1] Common Crawl News: https://github.com/commoncrawl/news-crawl
[2] https://github.com/optimaize/language-detector

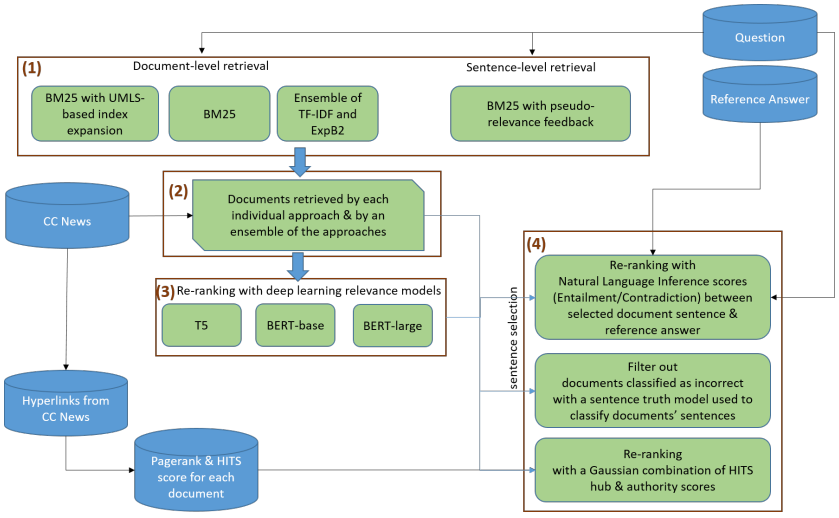

Fig. 1. Methods Overview

1.1 Retrieval Performance

We analyzed retrieval performance as it is critical for both the ad-hoc and total recall tasks. Table 1 presents a summary of our both our backend retrieval approaches and some of our first runs. All submitted runs are described in more details in section 1.2

We used the derived qrels for the useful (relevant) aspect to evaluate each of our approaches. We computed the values for ndcg @1000, ndcg @10, reciprocal rank (rr), and recall @1000 (cf. table 2.

The sentence-level indexing approaches (BNU, and TME) under-performed substantially document-level indexing approaches. Which is likely due in part to the *very low overlap between the lists of documents* retrieved by the document-based and sentence-based methods (cf. figure 2) and the *high correlation between the ratio of retrieved documents annotated by NIST assessors and the NDCG values* (cf. table 2.

To investigate this hypothesis, we performed a manual evaluation of one of the sentence-based approaches to analyze further the error cases.

We pooled the top 20 documents for each query from the BM25 (BNU) - T5 sentence-based method on all 46 topics and used the specific set of sentences returned by the method for each document as our textual evidence to assess its relevance for the topic. Table 3 shows the number of annotated documents, the number of documents in common between our annotations and the official useful qrels from NIST, and the agreement between our annotations and the official qrels on the common documents.

We present a summary of all annotation disgreement cases in table 4.

Model	Search Engine	Description
BM25	Terrier [9]	BM25 baseline
BM25 - T5	Terrier	BM25 baseline followed by a re-ranking using relevance scores from a T5 model [10] trained on MS-MARCO [8]
BM25 QE	Terrier	BM25 baseline with query expansion
Essie R [5]	Essie	Specialised biomedical search engine developed at the NLM with query relaxation
Essie L [5]	Essie	Specialised biomedical search engine developed at the NLM with extended query relaxation
TF-IDF	Terrier	TF-IDF baseline
InExpB2	Terrier	Inverse Expected Document Frequency model with Bernoulli after-effect and normalization
CombSUM [1]	Terrier	Ensemble of TF-IDF and InExpB2
CombSUM 2 - T5	Terrier	The CombSUM method applied to automatice summaries of the documents and re-ranking with T5
BM25 titles only	SolR	BM25 retrieval using only the title field
BM25 (BNU)	SolR	Sentence-level index with n-gram boosted retrieval. The sentences are grouped by their document ID and the sentences scores is used to produce a list of ranked documents.
BM25 (BNU) - T5	SolR	BM25 (BNU) followed by a re-ranking using relevance scores from a T5 model trained on MS-MARCO
BM25 (BNU) - BL	SolR	BM25 (BNU) followed by a re-ranking using relevance scores from a BERT-Large model [3] trained on MS-MARCO
E3	–	Ranking based ensemble
E4	–	Ranking-based ensemble
TME	–	Average-rank-based ensemble including BM25, BM25-T5, Essie-L, CombSUM, BM25 (BNU), and BM25 (BNU) - T5/BL.
TME_GH	–	TME results re-ranked with the Gaussian HITS scores
TME_NLIR	–	TME results re-ranked with a NLI model

Table 1. Summary of Information Retrieval and Re-ranking methods

method	**NDCG**	**NDCG@10**	**Recall**	**RR**	**% assessed (p+n)**
BM25	51.74	35.27	52.11	48.28	22.62
BM25 QE	46.17	21.17	48.73	31.96	21.85
BM25 - T5	54.89	**53.24**	54.24	**69.92**	21.83
ESSIE R	26.74	33.72	25.44	55.98	7.41
ESSIE L	34.00	32.29	26.99	55.33	14.39
InExpB2	54.43	47.14	57.02	59.15	**22.57**
TF-IDF	**56.12**	50.41	**57.98**	65.75	22.38
CombSUM	55.38	48.94	57.73	65.07	22.54
CombSUM 2	44.91	37.26	34.05	53.90	22.54
BM25 - titles	11.83	20.4	7.9	36.96	5.33
BM25 (BNU)	27.10	31.14	22.79	48.73	12.34
BM25 (BNU) - T5	32.96	48.18	27.50	65.60	12.34
BM25 (BNU) - BL	32.13	41.60	28.55	58.92	12.34
E3	27.55	32.84	34.36	50.7	10.06
E4	31.59	32.65	41.25	50.41	11.12
TME	33.07	27.92	49.36	47.94	11.8
TME_GH	36.53	28.56	49.38	48.69	11.81
TME_NLIR	19.23	25.46	24.54	40.32	7.53
Pearson correlation w/ % assessed	**0.96**	0.56	0.77	0.38	1

Table 2. Evaluation of our retrieval and re-ranking approaches for relevance.

Assessed docs	% in common with official qrels	Agreement (p+n)
881	30.6%	70.2%

Table 3. Statistics on manually annotated documents for error analysis.

Case %	Disagreement Type and Examples
1.6%	Relevant document annotated as not relevant.
13.1%	Borderline and clear cases where we reverted back to the official annotations.
85%	Divided between: – **The sentences answer the question but the document is not specific to the topic**, e.g.: "*He said, while social distancing and staying at home are effective prevention methods for Covid-19, it is very difficult for detainees to be granted such precautions.*" – **The answer can be inferred from the sentences but is not directly stated**, e.g.: "*Vaccinologist X told Y, a health website in the Asian country: "Quitting or staying away from passive smoking can indirectly prevent covid-19 infection.*" – **The answer is a part of a larger statement**, e.g.: "*Some of the tree species, including Khaya senegalensis, turmeric, ginger, garlic, iyere, onion, and so on contain good phyto-medicinal ingredients capable of curing COVID-19.*" or "*The drugs chloroquine, hydroxychloroquine and a combination of lopinavir and ritonavir can be used to treat other ailments and for treating Covid-19.*" – **The sentences are relevant for the question but do not fit the narrative or provide a definitive answer**, e.g.: "*Though some studies have shown ARBs increase ACE 2 activity in animal models, it must be emphasized that the results have been inconsistent, the researchers said.*" – **The sentences answer both the question and the narrative**, e.g.: "*Don't take the findings to mean those with type O blood are immune or are less likely to get COVID-19.*"

Table 4. Analysis and summary of all annotation disagreement cases.

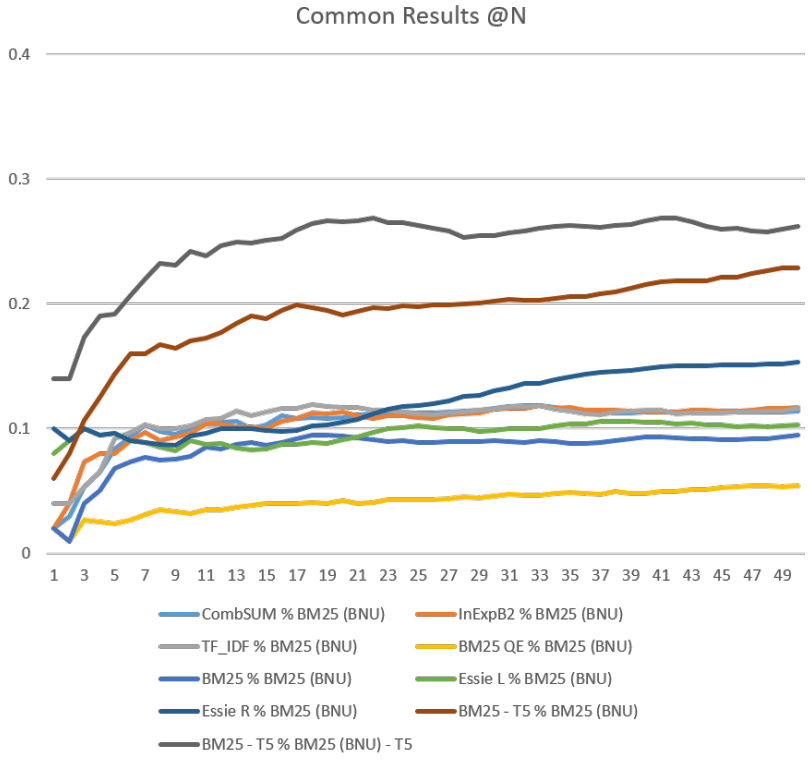

Fig. 2. Correlation between sentence- vs. document-level indexing approaches. Computed as number of common results at rank N.

To evaluate the impact of low intersection between the retrieval methods and the relatively low number of assessed documents for the sentence-based retrieval approaches, we merged our additional annotations with the official challenge annotations for the useful label (following the disagreement resolution described on table 4), and evaluated the relevance scores of the same methods on the expanded annotations (cf. table 5).

Our findings show that a substantial part of the lower performance of the sentence-based retrieval methods were not due to retrieval errors but rather to the incompleteness of the collection, with an NDCG@10 value improving from 48.18% to 81.41% for BM25 (BNU) - T5, and a relative improvement of 35% for NDCG.

All other sentence-based methods also benefited substantially from the added annotations. Approaches that relied on document-level retrieval with T5-based re-ranking also had substantial improvements in NDCG@10 and RR despite the low overlap in retrieved documents (cf. figure 2). Improvements reached up to 20% for BM25 - T5 in NDCG@10 and 15% in RR.

method	**NDCG**	**NDCG@10**	**Recall**	**RR**	**% assessed**
BM25	52.72	37.42	51.96	51.91	22.98
BM25 QE	47.05	22.69	49.34	33.32	22.24
BM25 - T5	**58.62**	63.93 (+20%)	55.91	80.48 (+15%)	22.22
Essie R	27.35	35.73	26.08	58.75	7.65
Essie L	34.43	33.22	27.26	55.48	14.74
InexpB2	55.48	50.12	56.76	64.95	22.95
TF-IDF	57.12	53.42	**57.57**	69.41	22.77
CombSUM	56.37	52.21	57.18	70.41	22.93
CombSUM 2 - T5	47.73	47.51 (+28%)	35.99	67.75 (+25%)	22.93
BM25 titles	12.23	21.4	40.4	7.86	5.45
BM25 (BNU)	31.78 (+17%)	37.53 (+20%)	27.9 (+22%)	55.98 (+15%)	12.92
BM25 (BNU) - T5	44.39 (+35%)	**81.41** (+69%)	34.64 (+25%)	**91.68** (+40%)	12.92
BM25 (BNU) - BL	41.05 (+27%)	58.49 (+40%)	35.12 (+23%)	74.47 (+26%)	12.92
E3	32.99 (+20%)	43.5 (+32%)	40.24 (+17%)	56.75 (+12%)	10.54
E4	36.49 (+15%)	41.62 (+27%)	46.22 (+12%)	56.53 (+12%)	11.59
TME	39.72 (+20%)	34.9 (+25%)	51.66	57.21 (+19%)	12.19
TME_GH	39.81	35.74 (+25%)	51.67	57.99 (+19%)	12.2
TME_NLIR	20.16	26.94	25.07	44.25	7.65
Pearson correlation w/ %assessed	**0.91** (-0.04)	0.32 (-0.23)	0.63 (-0.13)	0.37	1

Table 5. Evaluation of retrieval relevance with the additional annotations (deltas of more than 10% from original are indicated between parentheses).

The correlation between the performance measures of all approaches and their ratio of assessed documents decreased, with a 4 points drop in correlation for NDCG, despite the ratio of assessed documents increasing by only 1.12% for the pooled BM25 (BNU) - T5 approach.

This suggests that the distribution of false negatives in the collection was biased towards a restricted ranking perspective favoring document-level indexing and retrieval.

Considering that documents that were assessed as "not useful" were not annotated for correctness and credibility, this distribution bias has likely impacted the evaluation of our more advanced inference, re-ranking, and ensemble approaches that relied on the sentence-level retrieval methods.

1.2 Ad-hoc Runs

We submitted 10 runs to the Ad-hoc task.

BNU_ENS_NLI. Based on a sentence-level index of the CCN Covid collection, we first retrieved the top-30000 sentences with BM25 (BNU). The parent documents were then scored incrementally with each relevant sentence, then re-ranked with an ensemble of models (T5, BERT-base, BERT-large). Each question from the topics description was then transformed to an affirmative sentence automatically using syntactic rules and the Roberta [7] model for Natural Language Inference model trained on MultiNLI [12] was used to infer whether the most relevant sentence from the documents had an entailment/neutral/contradiction relation with the affirmative form of the topic. The final ranking was performed by putting the documents validating the reference answer first (in order of relevance) then the remaining documents (still ranked by relevance).

TME_NLIR. Ensemble retrieval method combining 4 conventional IR models, and 4 deep-learning based re-ranking models. The results were filtered to keep only documents with sentences mentioning both the subject and object entities in the questions. The most relevant sentence of each document was used to detect entailment, neutrality, or contradiction with the affirmative form of the topic. The results were then re-ranked according to their contradiction/entailment scores.

BNU_T5_CTM. Based on a sentence-level index of the CCN Covid collection, we first retrieved the top-30000 sentences with BM25 (BNU). The parent documents were then scored incrementally with each relevant sentence, then re-ranked with a T5 relevance-based ranking model. Only the most relevant sentence from the IR-based search was then classified as containing "false" or "true" claims using a Claim-Truth Model (CTM). The CTM model was built using T5 and a manually created training dataset derived from fact-checking websites. We used the dataset from [11] for validation. The model is fine-tuned to produce the tokens "true" or "false" depending on whether the claim is misinformation or not.

CTM_R1. Top-1000 documents retrieved with (BM25), re-ranked with a T5 relevance-based re-ranking model applied to the first 250 words in each document. Each sentence in the re-ranked documents was then classified as containing

"false" or "true" claims using the CTM model. A voting method was applied to derive a document-level classification from the sentences classification into a document-level classification.

CTM_R2. Top-1000 documents retrieved with the CombSum IR method. Each document was then automatically summarized using a pointer-generator model and the summaries were re-ranked with a T5 model. Each summary was then classified as containing "false" or "true" claims using a voting approach similar to NLM_CTM_R1.

TME_GH. We used the Hypertext Induced Topic Selection (HITS) algorithm [4, 6] to rank all domain names according to their authority and hub scores. Following several manual evaluations of different combinations of both scores, we selected a dependency based combination where the authority score is more significant at lower hub scores: i.e., when the source is a good authority without being a large hub. We wrap up the score in a Gaussian function to better distinguish between otherwise close scores, with:

$$GH = exp^{-(\frac{auth_s}{1+10hub_s}-1)^2} \quad (1)$$

Figure 3 shows the top 30 domains ranked with GH. For the NLM_TME_GH run, we first retrieved the relevant document with the TME retrieval approach (cf. table 2), then re-ranked the documents by using the GH value of their domain name as a score boost.

BNU_E_GH. Re-ranking with Gaussian HITS (hub/authority scores) and page rank scores based on the BNU_E retrieval ensemble.

E3. Average rank-based ensemble of 4 Ad-hoc runs based on different methods: BNU_T5_CTM, CTM_R2, BNU_ENS_NLI, TME_NLIR.

E4. Average rank-based ensemble of 5 Ad-hoc runs based on different methods: BNU_T5_CTM, CTM_R2, BNU_ENS_NLI, TME_NLIR, CTM_R1.

CTM_R1 was our best performing run for the binary measures and MAP using the official qrels. An important part of the performance is likely due to both a good NDCG value for its underlying retrieval approach (BM25-T5) and the high ratio of assessed documents among its top-1000 results. This aspect is better shown when comparing it with BNU_T5_CTM that used the same CTM model but a different sentence-based retrieval approach.

Our second best run for the binary measures was TME_GH, using an ensemble of retrieval methods including the better performing CombSUM and BM25-T5 approaches, and re-ranking with our Gaussian HITS approach GH.

The run with highest compatibility with helpful content was CTM_R1, and the run with the less compatibility with harmful content was TME_NLIR, which also had the best ratio of helpful vs. harmful compatibility.

On a more general note, the compatibility scores with harmful content for the submitted runs were correlated with their ratio of assessed documents (%A), with a Pearson correlation value of 0.49. This might be due to a lower presence in the collection for useful, correct (and credible) labels when compared with the combinations of useful, not correct, and (not) credible, which were used for the compatibility scores with harmful content.

	NDCG				cam_map			Compatibility[2]			Ret
	U.	U.C.	U.B.	U.C.B.	C.B.	U.B.	U.C.B.	Help.	Harm.	Ratio	%A
CTM_R1	**38.56**	**33.80**	**38.62**	**31.55**	**10.03**	**13.40**	**11.85**	**31.53**	6.58	4.79	10.46
CTM_R2	23.77	20.20	23.32	19.03	4.38	5.87	5.11	22.28	3.64	6.12	7.16
E3	27.55	24.63	25.78	22.38	4.13	5.18	5.07	19.69	3.00	6.56	10.06
E4	31.59	28.71	29.94	26.50	5.21	6.76	6.06	20.50	3.30	6.21	11.12
TME	36.51	29.79	33.21	28.31	6.78	10.05	8.63	15.03	6.86	2.19	11.8
TME_GH	36.53	29.84	33.29	28.50	6.84	10.01	8.61	15.32	6.80	2.25	11.81
$_{T}ME\_NLIR$	19.23	17.52	17.15	16.48	3.15	3.62	3.30	16.70	**1.30**	**12.84**	7.53
BNU_ENS_GH	29.25	23.58	27.91	22.81	7.43	10.27	8.94	24.59	8.04	3.03	8.61
BNU_ENS_NLI	14.07	14.58	13.12	13.31	4.11	3.73	3.66	19.20	1.89	10.15	4.02
BNU_T5_CTM	25.22	20.84	26.39	19.60	6.28	9.79	8.33	26.56	6.09	4.36	6.83
Correl. w/ Ret. %A	**0.91**	**0.91**	0.84	0.90	0.47	0.54	0.54	-0.16	0.48	-0.40	1

Table 6. Ad-hoc runs results

1. who.int	11. reuters.com	21. thelancet.com
2. cdc.gov	12. nbcnews.com	22. medrxiv.org
3. arcgis.com	13. amazon.com	23. fda.gov
4. bloomberg.com	14. nejm.org	24. scmp.com
5. apple.com	15. sciencemag.org	25. jamanetwork.com
6. cnn.com	16. whitehouse.gov	26. washingtonpost.com
7. nih.gov	17. bbc.com	27. cnbc.com
8. apnews.com	18. nytimes.com	28. theguardian.com
9. wsj.com	19. change.org	29. zoom.us
10. jhu.edu	20. nature.com	30. harvard.edu

Fig. 3. Top 30 domains as ranked by the Gaussian HITS method (GH).

As could be expected from our initial retrieval experiments, the relevance scores were highly correlated with the ratio of assessed documents in the top-1000 results of each approach, with Pearson correlation values ranging from .47 to .54 for MAP, and from .84 to .91 for NDCG.

1.3 Total Recall Runs

Following the challenge instructions, we submitted only three runs for the total recall task.

CTM_R1_C. We retrieved the top-10000 documents with (BM25-T5). Each sentence in the documents was then classified as containing "false" or "true" claims using the CTM model. A voting method was applied to generate a document-level class from turn the sentences classification. Only documents classified as "false" were selected for the Total Recall task.

BNU_E_NLI_C. Based on a sentence-level index of the CCN Covid collection, we first retrieved the top-30000 sentences with BM25 (BNU). The parent documents were then scored incrementally with each relevant sentence, then re-ranked with an ensemble of models (T5, BERT-base, BERT-large). Each question from the topics description was then transformed to an affirmative sentence automatically using syntactic rules and the Roberta [7] model for Natural Language

Inference model trained on MultiNLI [12] was used to infer whether the most relevant sentence from the documents had an entailment/neutral/contradiction relation with the affirmative form of the topic. The final ranking was performed by putting the documents contradicting the reference answer first (in order of relevance) then the remaining documents (still ranked by relevance).
TME_NLIR_C. Ensemble retrieval method combining 4 conventional IR models, and 4 deep-learning based re-ranking models. The results were filtered to keep only documents with sentences mentioning both the subject and object entities in the questions. The most relevant sentence of each document was used to detect entailment, neutrality, or contradiction with the affirmative form of the topic. The results were then re-ranked according to their contradiction/entailment scores.

Run	R-precision
TME_NLIR_C	3.06
BNU_E_NLI_C	<u>6.30</u>
CTM_R1_C	**9.76**

Table 7. Total Recall Runs results

2 Deep Learning Track

2.1 Passage Ranking

We submitted 4 runs for the passage ranking task.

nlm-bert-rr. a re-ranking run based on the official top 1000 passages provided by the challenge. We trained a BERT-base classifier on the MS-MARCO training data and re-ranked the documents using their relevance score from prediction.

nlm-prfun-bert. a full-ranking run where We indexed the sentences (UNits) of each passage then retrieved the top 1000 passages using a BM25 model with Pseudo-Relevance Feedback (PRF). We first retrieved the top 3000 sentences then scored the parent passages incrementally with each retrieved sentence and ranked the passages based on their final score. Pseudo-relevance feedback was applied at the sentence-level retrieval. We tested different negative sampling strategies to train a BERT-base model. We picked the best strategy from our tests on the dev set with 2:1 negative to positive ratio with examples randomly selected from the top 1000 passages.

nlm-ens-bst-2. Following the observation that a passage-level index provided substantially different results that the sentence-based index, we combined the results of the same BERT-base model when trained separately on each retrieval method using a downstream rank-based boost, where the score of each passage is boosted by $(1 + 1/r_2)$, and r_2 is the rank of the passage in the other method if it exists. When a passage was common to both methods, we selected the average of the boosted scores as the final passage score. The goal of this method was to keep the boosted scores comparable with the original scores of a given method and limit the range of re-ranking for any given passage.

nlm-ens-bst-3. In this run we followed a similar strategy to nlm-ens-bst-2 and added nlm-bert-rr as a third component of the ensemble.

Run	NDCG @ 10	NDCG @1000	RR	MAP
nlm-bert-rr	.672	.647	.778	.434
nlm-prfun-bert	.664	.643	**.860**	.426
nlm-ens-bst-2	**.693**	.667	.820	**.459**
nlm-ens-bst-3	.680	**.685**	.849	.452

Table 8. Passage Ranking Runs results

Our best reciprocal rank value was obtained by the nlm-prfun-bert run relying on sentence-level indexing and retrieval, whole the best ndcg@10 value was obtained by our ensemble method combining different sources for the top 1000 passages and a BERT model for re-ranking.

2.2 Document Retrieval

We submitted 2 baselines for the document ranking task.
nlm-bm25-prf-1. This is a full ranking run produced by a BM25 model based on a document-level index with a word limit set to 10 for query expansion with pseudo-relevance feedback.
nlm-bm25-prf-2. This is a full ranking run produced by a BM25 model based on a document-level index with a word limit set to 20 for query expansion with pseudo-relevance feedback.

Run	NDCG @ 10	NDCG @100	RR	MAP
nlm-bm25-prf-1	.467	.467	.808	.272
nlm-bm25-prf-2	**.470**	**.482**	**.809**	**.291**

Table 9. Document Ranking Runs results

3 Conclusion

We have described our submissions to the Health Misinformation and Deep Learning tracks of TREC 2020. Our methods have shown that point-wise re-ranking with neural language models fine-tuned for query-document relevance and claim-truth T5 models can outperform substantially conventional retrieval methods and NLI-based re-ranking. In the misinformation track, our approaches and runs followed an exploratory strategy where we used multiple top-K retrieval methods with low pairwise overlap in result sets. This exploration along with the error analysis that we presented in this paper highlighted a potential bias in the distribution of false negatives in the collection, favoring document-level indexing to retrieve the initial top-K texts for re-ranking in the ad-hoc an total recall tasks. Moving forward, a more relevant use of the test collection could be to rely on top-K retrieval methods that have high result set overlap with the list of assessed documents, and to focus on improvements of the downstream misinformation-based re-ranking task. It could also be relevant to study novel evaluation metrics that would take into account the amount of evidence that was used to evaluate the results of each approach.

Acknowledgments

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

References

1. Abacha, A.B., Khelifi, S.: List at trec 2015 clinical decision support track: Question analysis and unsupervised result fusion. In: TREC (2015)
2. Clarke, C.L., Smucker, M.D., Vtyurina, A.: Offline evaluation by maximum similarity to an ideal ranking. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. pp. 225–234 (2020)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Gibson, D., Kleinberg, J., Raghavan, P.: Inferring web communities from link topology. In: Proceedings of the ninth ACM conference on Hypertext and hypermedia: links, objects, time and space—structure in hypermedia systems: links, objects, time and space—structure in hypermedia systems. pp. 225–234 (1998)
5. Ide, N.C., Loane, R.F., Demner-Fushman, D.: Essie: a concept-based search engine for structured biomedical text. Journal of the American Medical Informatics Association **14**(3), 253–263 (2007)
6. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM) **46**(5), 604–632 (1999)
7. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
8. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: Ms marco: A human-generated machine reading comprehension dataset (2016)
9. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Johnson, D.: Terrier information retrieval platform. In: European Conference on Information Retrieval. pp. 517–519. Springer (2005)
10. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020)
11. Song, X., Petrak, J., Jiang, Y., Singh, I., Maynard, D., Bontcheva, K.: Classification aware neural topic model and its application on a new covid-19 disinformation corpus (2020)
12. Williams, A., Nangia, N., Bowman, S.R.: A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426 (2017)