Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Search like an expert: Reducing expertise disparity using a hybrid neural index for COVID-19 queries

Vincent Nguyen\textsuperscript{a,b,}\textsuperscript{*}, Maciej Rybinski\textsuperscript{b}, Sarvnaz Karimi\textsuperscript{b}, Zhenchang Xing\textsuperscript{a}

\textsuperscript{a} The Australian National University, Canberra, Australia
\textsuperscript{b} CSIRO Data61, Sydney, NSW, Australia

**ABSTRACT**

Consumers from non-medical backgrounds often look for information regarding a specific medical information need; however, they are limited by their lack of medical knowledge and may not be able to find reputable resources. As a case study, we investigate reducing this knowledge barrier to allow consumers to achieve search effectiveness comparable to that of an expert, or a medical professional, for COVID-19 related questions. We introduce and evaluate a hybrid index model that allows a consumer to formulate queries using consumer language to find relevant answers to COVID-19 questions. Our aim is to reduce performance degradation between medical professional queries and those of a consumer. We use a universal sentence embedding model to project consumer queries into the same semantic space as professional queries. We then incorporate sentence embeddings into a search framework alongside an inverted index. Documents from this index are retrieved using a novel scoring function that considers sentence embeddings and BM25 scoring. We find that our framework alleviates the expertise disparity, which we validate using an additional set of crowdsourced—consumer—queries even in an unsupervised setting. We also propose an extension of our method, where the sentence encoder is optimised in a supervised setup. Our framework allows for a consumer to search using consumer queries to match the search performance with that of a professional.

1. Introduction

As COVID-19—an infectious disease caused by a coronavirus—led the world to a pandemic, a large number of scientific articles shedding light on the new disease appeared in journals and other venues. Many scientific communities joined the effort of tackling the coronavirus crisis. In the computational linguistics and information retrieval community these efforts can be divided into the following: (1) \textit{applications}, including national epidemic surveillance\textsuperscript{[1]}, Atlas visualisation\textsuperscript{[2]}, information extraction\textsuperscript{[3]}, pandemic information retrieval\textsuperscript{[4]}, mental health\textsuperscript{[5]}, and gender-specific pandemic response\textsuperscript{[6];} and (2) \textit{datasets}, such as CORD-19\textsuperscript{[4]}, CODA-19\textsuperscript{[7]}, COVID-Q\textsuperscript{[8]}, LitCovid\textsuperscript{[9]}, and EPICQA\textsuperscript{[10]}. In the applications space, we present a novel approach to pandemic information retrieval that improves biomedical academic literature search on COVID-19 for experts and more importantly for \textit{consumers}. The consumer demographic includes anyone who does not have a high degree of expertise in the disease of interest. It encompasses policymakers, patients, the general public\textsuperscript{[11]}, or professionals from other fields. This demographic has a broad impact on society, and thus, meeting their information needs is important. However, due to an expertise gap, members of this group often experience difficulty in formulating queries to effectively search peer-reviewed medical literature to meet their information need\textsuperscript{[12]}. Thus, consumers may instead rely on less complete or, more worryingly, less credible resources, such as social media\textsuperscript{[13]}. However, a system allowing consumers to search for medical literature by what they mean (semantics) rather than what they know (vocabulary), can lead them to find answers to their questions via reputable peer-reviewed medical articles. Allowing users to search this way is critical during a pandemic, as there is a copious amount of false health-related information online\textsuperscript{[14]}, which can lead, for example, to reduced compliance with health guidelines and vaccine hesitancy\textsuperscript{[15]}.

Similarly, reducing the expertise gap allows professionals from other...
fields, who may be working on COVID-related tasks, faster access to documents meeting their information need. This reduction is important as even a professional seeking answers in biomedical literature often requires at least 30 minutes and usually has less than a 50% success rate [16].

The novelty at the beginning of the COVID-19 pandemic led to an unusual situation where medical professionals and consumers could have similar information needs. However, meeting the demands of both these demographics is challenging due to the expertise gap [17] and the lack of search systems designed for zero-day health pandemics [4].

Therefore, we investigate search effectiveness degradation in the COVID-19 pandemic scenario between professionals and consumers. We demonstrate that the expertise gap can be reduced through the use of contextual language representations in our hybrid neural index framework [18]. We also propose the extension to our approach, where we leverage the relevance feedback captured from professional users to refine (finetune) the neural indexing model to achieve increased retrieval effectiveness for professional and consumer users alike.

We empirically evaluate the effectiveness of our method using expert-formulated queries and two parallel sets of consumer queries: one collected via crowdsourcing, where users were asked to paraphrase an information need; and the second one compiled from queries sourced directly from MedlinePlus. Our results show that our system works well for both a consumer and a professional without the need to alter the system characteristics for either demographic. Moreover, the system works effectively in both the zero-shot and supervised settings.

The contributions of the paper can be summarised as follows:

- A novel application of a hybrid neural index that allows a consumer to query by what they mean, rather than what they know;
- An evaluation that shows a reduction in the effectiveness difference when searching as a consumer compared to an expert, as well as improving search effectiveness in general;
- A supervised learning (finetuning) extension to the neural indexing framework;
- A crowdsourced dataset of consumer query paraphrases parallel to TREC COVID topics [4].

2. Background

We provide an overview of the Sentence Transformers that allow for universal sentence vector representations as used in our study (An overview of the Transformers model is provided in Appendix B for reference). A review of the most relevant studies in the literature is also provided in this section.

**Sentence Transformer.** Transformers and their respective encoders do not have any inductive bias towards the final embeddings’ dimensions. In situations where task-specific labelled data is available, one could compute similarities by training a similarity measure over the corpus. However, this technique results in generally non-intuitive measurements as these similarity spaces tend to be non-linear and expensive to compute in real-time [19].

Sentence Transformer [19] rectifies this problem by introducing an inductive bias onto the transformer model via finetuning. The model in architecture mimics a Siamese network, where pairs of embeddings are computed simultaneously and trained to minimise the distance between similar embeddings and maximise the distance between dissimilar embeddings. In practice, for pairwise comparison, two embeddings are computed sequentially with the same network before any back-propagation occurs given by the equation:

\[ V_j = T_{sent}(S_j), j \in \{1, 2\}, \] (1)

where \( V_j \) is the \( j \)-th embedding vector which is generated by the \( T_{sent} \), Sentence Transformer, which has an input of a sentence pair, \( S_1, S_2 \), which can be denoted by an ordered collection of one-hot encoded words with dimensionality equivalent to the size of the vocabulary set, \( |V| \):

\[ S = \{w_1, w_2, ..., w_i\}, w \in \mathbb{N}^{|V|}. \] (2)

The sentence transformer is trained on a regression task with a mean squared error loss:

\[ \ell(y, V_1, V_2) = \|y - \cos(V_1, V_2)\|^2, \] (3)

where \( \ell(y, V_1, V_2) \) is a loss function with the inputs: ground truth label, \( y \) and sentence vectors \( V_1, V_2 \) and \( \cos(V_2, V_1) \) cosine similarity function given by:

\[ \cos(A, B) = \frac{A \cdot B}{|A| \times |B|}. \] (4)

**BM25.** Okapi Best Matching 25 (BM25) [20] is a bag-of-words featured-based ranking function used to estimate the relevance of a document, \( D \) for a given query \( Q \):

\[ BM25(D, Q) = \sum_{q_i \in Q} IDF(q_i) \cdot f(q_i, D) \cdot (k_1 + 1)/f(q_i) + k_2 \cdot (1 - b + b \cdot |D|/dl_{avg}) \] (5)

where \( IDF(q_i) \) represents the inverse document frequency of a particular query term \( q_i \), \( f(q_i, D) \) denotes the frequency of \( q_i \) in \( D \), \( f(q_i) \) denotes the frequency of \( q_i \) in the corpus and \( |D|/dl_{avg} \) denotes the document length divided by the average document length. \( k_1 \) and \( b \) are dimensionless hyperparameters, which can be interpreted as term saturation and document length penalty respectively.

BM25 is a strong baseline that machine learning models, including strong neural models, have struggled to beat [21,22] historically. In the context of the TREC COVID dataset, complex machine learning systems tended to fare worse in earlier rounds. This is reflected in earlier TREC evaluations (so, true zero- or few-shot settings), where it was discovered that simpler models, such as keyword-based statistical models that do not rely on machine learning, performed better than commercially deployed systems [23].

2.1. Related work

In this section, we discuss consumer health systems and also systems in the context of pandemic information retrieval.

**Consumer health search systems.** We define consumer health search systems as systems where the primary queries are health-related and may relate to a specific disease or medication and the system’s intent is to help the information need of a consumer.

An example of a consumer health system is ChiQA proposed by [24]. It is an end-to-end hybrid retrieval system where consumers type in their medical questions into the system which automatically extracts answers from reputable resources. The system performs well with questions that are similar to those in a known database, however, other missing questions use automatic synthesis from PubMed documents. Arguably, a limitation with ChiQA is that the system automatically synthesises answers given to them. This requirement restricts its users from seeing where the answers come from and furthermore, users tend to seek their information using keywords rather than the natural form (question) [25] which potentially reduces the effectiveness of the database which expects questions.

[26] propose using precise health card information on consumer search engines such as Google to inform the consumer’s decision-making process. These health cards have been shown to be effective for
accurate answers from a database.

questions, or multi-disease queries. Like ChiQA, it can only provide
architecture is then successfully applied to ad hoc re-ranking ([33
achieve state-of-the-art for several NLP tasks [32]. The
tecture — specifically BERT, which uses bi-directional transformer archi
of course, the T5 neural model in addition to the Anserini toolkit [29]. It
uses paragraph-level granularity at index time.

The use of neural networks in search has mostly been limited to re-ranking top results retrieved by a traditional ranking mechanism, such as Okapi BM25 [20]. That is, only a portion of top results is re-scored with a neural architecture [30]. Since the most successful neural re-ranking models depend on joint modelling of both documents and the query, re-scoring the entire collection becomes costly. Moreover, the effectiveness gains achieved with neural re-ranking are debated [22] until recently [31].

Since late 2018, large neural models pre-trained on language modeling—specifically BERT, which uses bi-directional transformer architecture—achieve state-of-the-art for several NLP tasks [32]. The architecture is then successfully applied to ad hoc re-ranking ([33–35]).

The existing pandemic search models are tuned for a specific task and may not generalise to other COVID-19 tasks.

### Datasets

**Documents.** CORD-19 (The Covid-19 Open Research Dataset) [36] is a dataset of research articles on coronaviruses (COVID-19, SARS, and MERS). It is compiled from three sources: (1) PubMed Central (PMC); (2) articles by the World Health Organisation (WHO); and (3) bioRxiv and medRxiv. This collection grew to over 68,000 articles by mid-June 2020 and 191,000 by mid-July 2020. The growth of CORD-19 continues with weekly updates [4]. For the experiments in this paper, we use the July 16, 2020 snapshot of the CORD-19 collection as this was the last dataset used in the TREC COVID challenge. Some statistics about this snapshot are shown in Table 1.

**TREC COVID Professional Topics.** As part of the TREC COVID search challenge, NIST provides a set of important COVID-related topics. Over multiple rounds, the topic set is augmented. Round 1 has 30 topics, with five new topics added per subsequent round. A sample topic is shown in Fig. 1. Each topic consists of three parts: query, question, and narrative.

In our experiments, we use TREC COVID topics as a proxy for professional-curated queries. To simplify the experiments we use the full set of topics (1–50), without the round-based setup.

**EPICQA Consumer Topics.** As a proxy for consumer search, we use

### Table 1

| Dataset Topic | Number of Sentences | Number of Tokens | Number of Documents |
|---------------|---------------------|------------------|---------------------|
| (1) Original | How does the coronavirus respond to changes in the weather effect on coronavirus |
| Crowdsourced | How COVID-19 respond to change |
| Crowdsourced | how does weather affect covid |
| Crowdsourced | coronavirus response weather changes |
| (2) Original | What kinds of complications related to COVID-19 are associated with diabetes |
| Crowdsourced | COVID diabetic complications |
| Crowdsourced | Complications to diabetics with coronavirus |
| Crowdsourced | COVID-19 are associated with diabetes |
| Crowdsourced | covid-19 complications diabetes |
| (3) Original | Which biomarkers predict the severe clinical course of 2019-ncov infection? |
| Crowdsourced | biomarkers predict course 2019-ncov infection |
| Crowdsourced | Severe covid and biomarkers |
| Crowdsourced | biomarker predictors covid clinical course |
| (4) Original | How much impact do masks have on preventing the spread of the COVID-19? |
| Crowdsourced | Do masks prevent spreading corona virus? |
| Crowdsourced | masks to prevent covid19 |
| Crowdsourced | masksCOVID-19 |
| Crowdsourced | masks prevent spread of COVID-19 |
| (5) Original | What is the mechanism of inflammatory response and pathogenesis of COVID-19 cases? |
| Crowdsourced | mechanism for inflammatory response in covid cases? |
| Crowdsourced | inflammatory response and pathogenesis of COVID-19 |
| Crowdsourced | inflammatory response and pathogenesis covid |
| Crowdsourced | inflammatory response pathogenesis mechanism COVID-19 cases |

**Examples of crowdsourced (query) and EPICQA (narrative facet) topics.**

A sample topic from the COVID search task.

**Fig. 1.** A sample topic from the COVID search task.
general public queries supplied in the preliminary Epidemic Question Answering challenge (EPICQA [10]). These queries are consumer variations sourced from MedlinePlus and matched to the TREC COVID counterparts by the EPICQA organisers, who refer to these queries as consumer-friendly queries. The dataset contains 42 queries, as parallel queries for TREC COVID topics 35–37 and 46–50 are not included.

Crowdsourced Queries We collected consumer variations of questions sourced from the TREC COVID search challenge using crowdsourcing, via Amazon Mechanical Turk. We used the following inclusion criteria for participation:

- Resident of the following English speaking countries: Great Britain, Australia, New Zealand, United States;

and the following exclusion criteria:

- Participants who may be traumatised from partaking in a COVID-19 study (e.g., relatives of COVID-19 patients).

We presented thirty-four participants with an interface containing the Google search engine with personalised search disabled. Participants were asked to search for relevant websites to a particular information need. Each participant was given a maximum of 10 minutes per topic. Participants were encouraged to ignore or remove terms they did not understand in the query as they would not have searched using these terms otherwise. We collected the query variations they used to search for the information we limit their query’s maximum length to prevent participants from copying the entire question. The maximum reformulated query length \( l \) (in words) is given by:

\[
l = \max(5, |Q|^{0.1\frac{1}{2}}),
\]

where \(|Q|\) denotes the length of the original topic’s question field. We use these query variations as a second proxy for consumer search. We collected between 3 and 5 query variations per topic, given a total of 217 query variations. The user interface used for this study is shown in Appendix A.6. Examples of the crowdsourced queries (together with original questions) presented to the participants are shown in Table 2.3

On average we sourced 4.34 (±0.72) reformulations per query. The average length of the TREC COVID questions presented to the participants was 10.2 (±3.04) words, while the mean reformulation length is 4.45 (±1.24). It is worth noting that these reformulations are on average longer than official (expert-formulated) queries (as opposed to questions) supplied with the TREC COVID topics (average length of 3.24, ±1.22). This context is important for experiments that contrast retrieval effectiveness on short queries and reformulations presented further in this paper.

On average a reformulated query would contain 3.3 (±1.54) terms of the original question. While this finding can be largely explained with our crowdsourcing set-up, it seems to point to a user behaviour where queries are created by combining a reduction of known phrases (i.e., picking keywords) and adding some additional terms from the users’ vocabularies. This can be illustrated with an example (3) from Table 2— all reformulations select some key terms (‘biomarkers’, ‘clinical course’, ‘predict’), but two of them also replace ‘2019-ncov’ infection with (now) more idiomatic ‘covid’. We analyse this user behaviour further by calculating word differences between original questions and crowdsourced reformulation.

In particular, for each term \( t \) present in the entire set of original questions we calculate all reformulations omitting \( t \) (where it was present in the original query). We normalise this raw count by dividing it by the total number of reformulations sourced for original questions containing \( t \). We refer to this value as removal frequency. We also count terms added to reformulations (i.e., terms that appear in reformulations but do not appear in the original questions).

Some of the removal frequency results are intuitive—participants tend to remove stopwords (‘a’, ‘as’, ‘at’ all have removal frequency of 1) and semantic sugar (e.g., ‘develop’ in ‘develop symptoms’, also has removal frequency of 1). Interestingly, some specific concepts were also unanimously removed from user queries, e.g., ‘t-cell’ and ‘antibody’. This can be better understood in context, as both these terms appear in question from topic 49 (Do individuals who recover from COVID-19 show sufficient immune response, including antibody levels and t-cell mediated immunity, to prevent re-infection?), which means that they are arguably not central to the information need (reinfection is).

Among the most frequently added terms alternative names for the condition or virus are common. Term ‘covid’ was added to 63 reformulations (3 times as a part of ‘covid 19’ collocation), followed by ‘covid19’ (8), ‘corona’ (5), ‘covid-19’ (5), and ‘coronavirus’ (3). Another term added frequently is ‘and’ (10), often used to join keywords (e.g., ‘severe covid and biomarkers’). Some participants added terms like does (5), how (4), and ‘can’ (3) to formulate their queries as questions. A relatively large proportion of infrequent or unique added terms are changes in grammar for keyword queries (e.g., ‘spreading’ in reformulation 1 of Example 4 from Table 2).

Another prominent class of unique added terms is misspellings (e.g., ‘violence’, ‘proteins’, ‘impacts’, ‘comlications’). A separate subclass of misspellings that can be observed frequently (18 occurrences) is wrongly merged words (for example see Table 2: ‘masksCOVID-19’ in reformulation 3 of example 3 and ‘forinflammatory’ in reformulation 1 of example 5). While these merges can be an artifact of our crowdsourcing setup (specifically, word limit), we observe several cases of wrongly merged words with reformulations that would have been under the word limit with correct spacing. This seems to suggest that wrongly merged words account for an important class of errors in consumer queries.

Relevance Judgements. TREC COVID introduced a round-based, manual, incremental evaluation setup, where documents already judged in the previous round did not count towards evaluation in subsequent rounds. Document-topic pairs were selected for manual judgment based on a pooling method over a sample of the submitted runs. Here, we use cumulative judgment sets. In other words, to evaluate a topic, we use a set of judgments compiled from a sum of judgment pools pertaining to all individual rounds where this topic was present in TREC COVID.

For each topic, a document is judged as irrelevant (0), partially relevant (1), and relevant (2) by a biomedical professional. Table 3 presents an overview of the relevance judgments pool used in our experiments.

We use the original relevance judgements from the TREC COVID challenge for evaluation in all experiments. In particular, we evaluate EPICQA and crowdsourced consumer queries against TREC COVID relevance judgements for their professional-formulated TREC COVID counterparts.

4. Methods

Our approach uses a hybrid scoring function between BM25 and cosine similarity of universal sentence embeddings generated at index-time.

BM25. We tune the BM25 model, similar to [37] on recall. We do a search on the parameters for \( k_1 \in [0.1, 3.7] \) and \( b \in [0.0, 1.0] \) in 0.1
To evaluate, we used a subset of relevance judgements to determine the optimal values based on either a precision-based metric or a recall-based metric.

**Neural Index.** To bridge the gap between medical professionals and a consumer, we utilise a neural embedding space where both inputs can be directly compared. We apply our model which we proposed in the TREC COVID challenge [18]. It uses pretrained transformer encoders fine-tuned with a Natural Language Inference (NLI) objective [19] that can produce universal sentence embeddings. These embeddings can be used for direct comparison using cosine similarity, without additional training, in a neural semantic index.

In our model, we use a hybrid neural index, where we have a neural index in conjunction with a traditional inverted index. Formally, the relevance score $\psi$ for the $i^{th}$ topic $T_i$ and document $d \in D$ is given by:

$$\psi(T_i, d) = \log \left( \sum_{f \in d} BM25(f, t) \sum_{t \in T_i} cos(V_f, V_t) + \alpha \right),$$

where $z$ is a normalisation hyper-parameter, $t \in T_i$ represents possible fields of the topic (i.e., query, narrative and question), $f \in d$ represents possible facets of the document (i.e., abstract, title, body), BM25 denotes the BM25 scoring function (Eq. 5), $V_f$ denotes the universal neural representation of the topic field, $V_t$ denotes the universal neural representation of the document facet, and $\alpha$ denotes cosine similarity (Eq. 4) and $\alpha$ denotes an offset (which we set to 1), such that the function’s range be non-negative. An overview of the system is shown in Fig. 2.

Scoring is done for all documents, although a caveat is that log zero (for a BM25 of zero) is undefined which will exclude these documents from the result set, however the cosine calculation is still done for these documents. As such, our method is considered as dense retrieval rather than re-ranking.

The $z$ hyper-parameter normalises each query dynamically and is determined analytically by:

$$z = \sqrt{\max (BM(f, t))},$$

where $R_{\text{cos}}$ is the upper range of the summed cosine similarity function:

$$R_{\text{cos}} = \max \left( \sum_{f \in d} \sum_{t \in T_i} \cos(V_f, V_t) \right).$$

### 4.1. Finetuning a sentence similarity ranking model

We extend our method in Section 4 by training the model with the concept of ranking. To do this, we perform additional training of the models with relevance judgments from the TREC COVID challenge, which we normalise by dividing by two to give a range of $[0, 1]$. We use the Sentence Transformer training procedure. Specifically, a regression objective is trained by taking the cosine similarity between embeddings (Eq. 4) and then a cosine similarity regression loss objective (Eq. 3).

### 5. Experiments

We perform a series of experiments focused on (1) showing that a retrieval effectiveness gap between professionals and consumers exists, and (2) validating our hypothesis that semantic indexing can effectively bridge this gap.

As a traditional keyword matching information retrieval model, we use a BM25 baseline optimised on the TREC COVID dataset (professional queries). We compare its retrieval effectiveness on TREC COVID and respective sets of parallel consumer queries (EPICQA and the crowdsourced queries). The effectiveness of the professional-optimised BM25 model is then compared to that of untuned Neural Information Retrieval (NIR) on all three sets of queries (professional, EPICQA, and crowdsourced). In particular, we experiment on EPICQA consumer queries and contrast these results with those obtained for a corresponding subset of TREC COVID topics. When comparing results on crowdsourced queries, we compare against experiments on the full TREC COVID dataset. In all systems using neural indexing we use the BM25 parameters tuned on professional queries. This allows us to replicate a scenario in which a seed of expert knowledge (TREC COVID topics and judgments) is used to increase retrieval effectiveness in a system used by a wider public. The same approach is reflected in our supervised experiments, described
subset of the TREC COVID expert topics. When comparing the crowd topics, as EPICQA only provided consumer-friendly narratives for a remaining 20% of the professional queries and corresponding EPICQA e. NIR and variants), the same BM25 hyper-parameters are used. More the query field was used. For both comparisons and across all models (i. Nielsen experiments: NIR-finetuned and Neural Finetuned. For each test fold (on which the model is evaluated), both models, professional, and consumer, are finetuned on the training subset (remaining 4 folds) of professional queries (with corresponding relevance judgments); evaluation is limited to a subset of parallel queries (present in both datasets). We use a paired t-test to evaluate the statistical significance of improvements over the Expert BM25 baseline. † denotes a significant difference with a p-value of less than 0.05, while a ‡ denotes a statistically significant difference with a p-value less than 0.01.

| Query Type | Model Type | NDCG@10 | NDCG@20 | P@10 | P@5 | P@100 | Recall@100 | bpref@1000 |
|------------|------------|---------|---------|------|-----|-------|------------|------------|
| Consumer   | BM25       | 0.642   | 0.603   | 0.717 | 0.515 | 0.729 | 0.109      | 0.315      |
| Professional | BM25     | 0.737   | 0.716   | 0.794 | 0.631 | 0.812 | 0.133      | 0.360      |
| Consumer   | Neural Only | 0.501   | 0.481   | 0.536 | 0.390 | 0.562 | 0.080      | 0.274      |
| Professional | Neural Only| 0.508   | 0.476   | 0.540 | 0.397 | 0.516 | 0.084      | 0.284      |
| Consumer   | Neural finetuned | 0.738 | 0.711   | 0.747 | 0.588 | 0.769 | 0.128      | 0.418‡     |
| Professional | Neural finetuned | 0.791   | 0.768   | 0.796 | 0.628 | 0.806 | 0.141      | 0.441†     |
| Consumer   | NIR        | 0.711   | 0.707   | 0.779 | 0.607 | 0.814 | 0.127      | 0.362      |
| Professional | NIR     | 0.787†  | 0.763| 0.852| 0.680 | 0.876| 0.144| 0.400‡    |
| Consumer   | NIR-finetuned | 0.830| 0.811    | 0.879| 0.714| 0.903| 0.152‡| 0.469‡     |
| Professional | NIR-finetuned | 0.882| 0.861| 0.919| 0.770| 0.928| 0.167| 0.510‡     |

Table 5
Comparison for Consumer (Crowdsourced) and Professional (TREC COVID) queries using different retrieval models. 5-fold cross-validation is used for finetuned neural models experiments: NIR-finetuned and Neural Finetuned. For each test fold (on which the model is evaluated), both models, professional, and consumer, are finetuned on the training subset (remaining 4 folds) of professional queries (with corresponding relevance judgments); evaluation is limited to the query field in the topics. The same professional finetuned neural models are used for the consumer queries. We use a paired t-test to evaluate the statistical significance of improvements over the Expert BM25 baseline. † denotes a significant difference with a p-value of less than 0.05, while a ‡ denotes a statistically significant difference with a p-value less than 0.01.

| Query Type | Model Type | NDCG@10 | NDCG@20 | P@10 | P@5 | P@100 | Recall@100 | bpref@1000 |
|------------|------------|---------|---------|------|-----|-------|------------|------------|
| Consumer   | BM25       | 0.609   | 0.565 | 0.635 | 0.453 | 0.676 | 0.096      | 0.302      |
| Professional | BM25     | 0.642   | 0.623   | 0.682 | 0.546 | 0.704 | 0.117      | 0.349      |
| Consumer   | Neural Only | 0.475   | 0.429   | 0.462 | 0.336 | 0.482 | 0.074      | 0.253      |
| Professional | Neural Only| 0.481   | 0.440   | 0.464 | 0.326 | 0.444 | 0.069      | 0.243      |
| Consumer   | Neural Only NIR-finetuned | 0.765| 0.730| 0.739| 0.570| 0.762| 0.134| 0.420      |
| Professional | Neural Only NIR-finetuned | 0.733| 0.703| 0.706| 0.560| 0.730| 0.126| 0.404      |
| Consumer   | NIR        | 0.642   | 0.594   | 0.667 | 0.481 | 0.682 | 0.101      | 0.309      |
| Professional | NIR     | 0.684†  | 0.632   | 0.728| 0.561| 0.736| 0.120| 0.348      |
| Consumer   | NIR-finetuned | 0.753| 0.710| 0.778| 0.592| 0.808| 0.128| 0.385      |
| Professional | NIR-finetuned | 0.751| 0.722| 0.799| 0.647| 0.826| 0.139| 0.406‡     |

Further in this section. Moreover, tuning the parameters on professional queries means we compare a tuned baseline (‘professional BM25’) to warm-tuned systems (with the BM25 component optimised for professional, not consumer, queries). When comparing EPICQA and the TREC COVID expert topics, the all topic fields (query, question, narrative) from both topic sets were used. Only parallel topics were used between EPICQA and TREC COVID expert topics, as EPICQA only provided consumer-friendly narratives for a subset of the TREC COVID expert topics. When comparing the crowdsourced query reformulations against TREC COVID expert topics, only the query field was used. For both comparisons and across all models (i.e. e. NIR and variants), the same BM25 hyper-parameters are used. More specifically, the BM25 ranker used in the NIR model (unfinetuned and finetuned) is the same one used for the baseline.

We also measure the effectiveness of the NIR-finetuned model. We evaluate NIR-finetuned in 5-fold cross-validation settings, which means we are finetuning on 80% of the professional queries, evaluating on the remaining 20% of the professional queries and corresponding EPICQA topics such that each validation (20%) set in the 5 folds is disjoint. We also measure the effectiveness of the NIR-finetuned model trained on the entire professional query dataset on the crowdsourced queries. This experiment gives us additional insight into how the professional search (queries, relevance judgments) can contribute to improving the search effectiveness for consumers (the general public).

As an ablation study, we isolate the untuned neural component (Neural Only), and finetuned neural component (Neural Finetuned) to give insight on the contribution of the semantic embeddings in improving the search effectiveness of professionals and consumers.

Finally, we show the generalisability of our method by optimising the BM25 hyperparameters on: (1) a subset (30 topics) of the TREC COVID datasets and (2) on the full (50 topics) TREC COVID datasets. We validate that our semantic indexing improves over a fully tuned BM25 baseline.

6. Experimental setup

All experiments and indexing were performed on a V100 Nvidia GPU with 8 CPU cores and 64 GB RAM at a rate of 2100 documents per hour. We used Elasticsearch as our index backend and bert-as-service5 for encoding sentence embeddings.

In the initial retrieval step using BM25, we retrieve a thousand documents per topic to calculate z in Eq. 7 that is used for the hybrid neural retrieval.

Evaluation metrics. We use a retrieval size of 1000 documents per topic and evaluate using three precision-focused metrics and recall:

5 https://github.com/hanxiao/bert-as-service.
Fig. 3. Hyperparameter tuning of the $k$ (term saturation) and $b$ (document length penalty) values in consumer (crowdsourced) queries. Recall is calculated at cutoff 1000 using round 1 judgements (left), and round 5 judgements (right).

Fig. 4. Hyperparameter tuning of the $k$ (term saturation) and $b$ (document length penalty) values in professional (TREC COVID) topics. We evaluate with recall using round one judgements (left) and round 5 judgements (right).

Fig. 5. Precision at 100 (P@100) against the hyperparameter, $z$. A lower value of $z$ gives more weight to BM25, whereas a higher value gives more weight to neural cosine similarity. A static value of $z$ is used for all Query Types except Automatic Tuning. Automatic tuning uses the upper range of BM25 to determine $Z$ for the professional topics (Eq. 7). The query only setting, where only the query facet of the professional topic is used, allows for fair comparison between consumer crowdsourced topics and professional topics.
1. NDCG at cut-off rank $k$ (NDCG@$k$) [38], a ranking metric that logarithmically penalises relevant documents appearing later in the ranked list of size $k$ (i.e., if relevant documents are ranked below irrelevant documents);
2. Precision at cutoff rank $k$ (P@$k$), which measures the number of retrieved relevant documents at a cutoff rank;
3. Mean Average Precision (MAP), the average precision (precision averaged at all recall levels) averaged over all queries;
4. B-pref, which measures how often relevant judged documents are placed above irrelevant judged documents; B-pref has the benefit of limiting penalty resulting from incomplete judgment sets (specifically, from the presence of unjudged query-document pairs);
5. Recall at cut-off rank $k$ (Recall@$k$), which measures the number of relevant documents at a cut-off rank $k$ proportional to all relevant documents for that given topic.

For all experiments (runs), we use a paired t-test to evaluate the statistical significance of improvements over the professional BM25 baseline. $\dagger$ denotes a significant difference with a $p$-value of less than 0.05, while $\ddagger$ denotes a statistically significant difference with a $p$-value less than 0.01.

7. Results

Our experiments are designed to answer the following: (1) Is there a gap between consumer and professional retrieval effectiveness? And (2) can neural semantic indexing effectively bridge that gap?

Our first experiment (Table 4), uses the EPICQA dataset as a proxy for consumer topics compared against TREC COVID (professional) topics. This experiment highlights a considerable gap in search effectiveness between the consumer and professional BM25 baselines (rows 1–2). This gap can be observed for both models, NIR (rows 3–4) and BM25 (1–2). An important observation is that using the NIR model combined with the BM25 signal tuned on professional queries improves the consumer retrieval effectiveness to a level comparable to that of a tuned BM25 baseline for professional queries (row 2). We further found that by finetuning the neural component (NIR-finetuned), a considerable and statistically significant improvement is observed across NDCG@$k$, P@$k$, Recall@$k$, and bpref@$k$, such that the retrieval effectiveness with the finetuned model and consumer queries (EPICQA) surpasses that of BM25 professional baseline. This is an important result, as neural models are often outperformed by a tuned BM25 baseline [22,23].

We also experiment with the neural ranking component in isolation (neural only; rows 7–8). Comparisons show that the combination of Neural and BM25 (NIR and NIR finetuned) yields better effectiveness than the respective individual components, suggesting that neural and BM25 models combine complementarily.

For the second experiment, we use consumer crowdsourced consumer queries against professional (TREC COVID) queries (so, only the short representation professional-formulated representation of the topic), with results shown in Table 5. Though the scale of the metrics is lower as only the query portion (a few keywords) of the topic is used, we found a similar consumer/professional gap to the first experiment. We also found a similar observation where the neural finetuned models are stronger than the BM25 counterparts and the Consumer NIR model closing the gap with the BM25 professional model.

We found overall that both recall and precision are lower with shorter topics when compared to using full topics (as in Experiment 1) highlighting the importance of precise medical terminology. Moreover, for long-tail metrics with cut-offs higher than 100 (such as bpref), we found that consumer queries were significantly worse than the professional BM25 baseline, even with the neural components.

However, interestingly, the consumer queries perform better than professional queries when using the finetuned neural component (row 9–10). This can be explained by the fact that consumer queries tend to be more verbose in comparison to professional queries; they resemble complete sentences more, which aids the neural component.

For both experiments (Table 4 and 5), the precision metrics indicate that most documents in the top-10 and top-100 (over 80% with the best model) were relevant to the given information need. This is an important result that shows that reliable relevant information is found with our method.

We found that our neural index improves retrieval in both experiments. It also reduces the gap (or the delta) between professional and consumer queries. This is true for: (1) baseline comparison (i.e. Professional BM25 as the baseline against Consumer NIR) where the Consumer models matched (Table 5) or was not significantly worse (Table 4) or (2) pairwise comparison (i.e. Professional NIR against Consumer NIR) where the absolute performance difference between Professional BM25 and Consumer BM25 is reduced when using the neural index.

We report on our hyperparameter tuning for BM25 tuning (Figs. 3 and 4) which we use to determine the settings for the experiments. Similar to [37], we use Recall@100 with round 1 judgements from TREC COVID to find optimal values for parameters $k_1$ and $b$ of Okapi BM25 [20]. Interestingly, we found that for professional and consumer queries, the optimal value is $k_1 = 3.6$ and $b = 0.9$, although the distributions of the two heatmaps are entirely different. We do additional experiments where we verify if these choices hold for the final round of TREC COVID and found that the distributions shift significantly. We also performed BM25 hyperparameter tuning on round 5, or the full set of relevance judgements which gave the values of $k_1 = 1.7$ and $b = 0.9$. However, we found that although BM25 improves slightly, all hybrid approaches performed worse (see Appendix Table D.6 (EPICQA vs professional) and Table D.7 (Crowdsourced vs professional)).

We also report tuning for the hyperparameter $\pi$ in Fig. 5 where automatic tuning outperforms all static values of $\pi$. This result shows that automatic tuning is required for better performance, however, it also shows that the BM25 scoring function and the neural scoring function are closely related as the peaks of all the query types are aligned.

8. Discussion

By applying our neural framework, we found that the gap between medical professionals and individuals without medical expertise is reduced in EPICQA (Table 4) and crowdsourced queries (Table 5). In particular, for the EPICQA dataset, the NIR model, performs better than the BM25 baseline across all metrics (Table 4). Importantly, the ranking effectiveness increased substantially from 0.603 to 0.707 in terms of NDCG@20.

Consumer vs Professional. However, this constructed query setting may not accurately reflect real-world query scenarios. As a proxy, we conduct a user study (Table 5), where users are shown a COVID-19 question, drawn from the TREC COVID topic set, and were asked to find websites with the relevant information which would answer that question. We collected the queries submitted by the users, with each question having four variations. Similar to the previous results (see
9. Conclusions

We introduce a novel application allowing consumers to search for coronavirus-related information using medical terms they know. Our method allows consumers to retrieve similar document sets that a professional would receive in the same search session. This effectiveness is reflected by a decrease in performance degradation between professional topics with BM25 and consumer queries using our method. Our application works in a zero-day pandemic retrieval setting, where labelled data is not readily available, but also in a continual learning setting. In the future, given the general nature of the application, expanding the scope of the search task from coronavirus information needs to general-purpose medical questions is a suitable avenue for future research.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Vincent Nguyen is supported by the Australian Research Training Program and the CSIRO Postgraduate Scholarship. This work is also funded by the CSIRO Precision Health Future Science Platform.

6 For instance, if the TREC COVID expert topic’s query, question and narrative fields are used, this would give a total of 9 cosine calculations.
Appendix A. User study instructions

We presented users with the following interface during the study (Fig. A.6).

Appendix B. Transformer

The Transformer is a transduction architecture that relies entirely on self-attention to compute representations, without using a Recurrent Neural Network (RNN) or convolution operations [40].

The architecture is centered on an encoder-decoder model where the encoder blocks have special Multi-Head attention passed through a Feed-Forward Neural Network (FFNN). The decoder architecture features Masked Multi-Head attention. These blocks are stacked $N$ times, where $N$ is a hyper-parameter for the number of layers in the network. Bidirectional Encoder Representations from Transformers (BERT) [32] uses the encoder half of a Transformer for transfer learning.

Self-attention. A key mechanism in how transformers work is the use of self-attention [40]. Suppose there is a sequence $S = (w_1, w_2, ..., w_{|S|})$, $w_i \in \mathbb{R}^w$. The self-attention for a word $w_j$ is

$$A_{w_j} = \text{softmax} \left( \frac{Q_{w_j} \cdot K^T}{\sqrt{d_k}} \right) V,$$

where $K^T = (k_{w_1}, k_{w_2}, ..., k_{w_{|S|}})$, and $k_{w_i}$ is a learned key vector for each word $w_i$.

Intuitively, self-attention can be viewed as a graph. The nodes (words) in the graph can pool information from other nodes (words) in each attention step in order to optimise towards their own goals. Such a goal could be anaphora resolution (resolving the subject of a pronoun throughout a sentence): The surgeon said he would do the procedure.
Appendix C. Number of Relevance Documents per Topic

An overview of the relevant documents per professional topic is given in Fig. D.7.

Fig. D.7. Number of relevant documents for each topic in the TREC COVID challenge.
Appendix D. Additional results

We show additional results (Table D.6) that are derived from the main results in Table 4, however, the BM25 model used in NIR and as a baseline is tuned using the full set of TREC COVID relevance judgements and topics rather than a subset as in Table 4. There is an improvement in BM25 for both query types, however all neural models performed worse. This is most likely due to overfitting to the professional topics as the model generalised less effectively which the Z parameter could not overcome.

Similarly (Table D.7) shows results derived from Table 5.

Table D.6

Comparison for Consumer (Crowdsourced Queries) and Professional (TREC COVID) using different retrieval models, similar to Table 4 which used a subset for BM25 tuning. The BM25 model here is tuned using the full set of TREC COVID relevance judgements and topics.

| Query Type        | Model Type | NDCC@10 | NDCC@20 | P@10  | P@5   | P@100 | Recall@100 | bpref@100 |
|-------------------|------------|---------|---------|-------|-------|-------|------------|-----------|
| Consumer          | BM25       | 0.634   | 0.601   | 0.710 | 0.508 | 0.762 | 0.108      | 0.319     |
| Professional      | BM25       | 0.743   | 0.724   | 0.807 | 0.637 | 0.833 | 0.125      | 0.368     |
| Consumer          | NIR        | 0.651   | 0.630   | 0.731 | 0.577 | 0.738 | 0.121      | 0.376     |
| Professional      | NIR        | 0.712   | 0.701   | 0.781 | 0.669 | 0.795 | 0.132      | 0.416     |
| Consumer          | NIR-Finetuned | 0.833 | 0.814   | 0.881 | 0.719 | 0.905 | 0.153      | 0.474     |
| Professional      | NIR-Finetuned | 0.867 | 0.858   | 0.907 | 0.782 | 0.916 | 0.159      | 0.503     |
| Consumer          | Neural Only | 0.501   | 0.481   | 0.536 | 0.390 | 0.562 | 0.083      | 0.274     |
| Professional      | Neural Only | 0.508   | 0.476   | 0.540 | 0.397 | 0.516 | 0.084      | 0.284     |
| Consumer          | Neural Finetuned | 0.717 | 0.694   | 0.747 | 0.588 | 0.769 | 0.127      | 0.418     |
| Professional      | Neural Finetuned | 0.770 | 0.753   | 0.796 | 0.627 | 0.806 | 0.131      | 0.424     |

Table D.7

Similar to Table 5 which used a subset for BM25 tuning. The BM25 model here is tuned using the full set of TREC COVID relevance judgements and topics.

| Query Type        | Model Type | NDCC@10 | NDCC@20 | P@10  | P@5   | P@100 | Recall@100 | bpref@100 |
|-------------------|------------|---------|---------|-------|-------|-------|------------|-----------|
| Consumer (Crowdsourced) | BM25       | 0.496   | 0.467   | 0.547 | 0.384 | 0.578 | 0.078      | 0.253     |
| Professional (TREC COVID) | BM25       | 0.653   | 0.624   | 0.714 | 0.544 | 0.724 | 0.117      | 0.355     |
| Consumer (Crowdsourced) | NIR        | 0.533   | 0.501   | 0.582 | 0.419 | 0.593 | 0.085      | 0.267     |
| Professional (TREC COVID) | NIR        | 0.685   | 0.646   | 0.736 | 0.564 | 0.768 | 0.121      | 0.352     |
| Consumer (Crowdsourced) | NIR-Finetuned | 0.752 | 0.727   | 0.808 | 0.663 | 0.828 | 0.145      | 0.672     |
| Professional (TREC COVID) | NIR-Finetuned | 0.770 | 0.743   | 0.817 | 0.706 | 0.858 | 0.157      | 0.711     |
| Consumer          | Neural Only | 0.396   | 0.381   | 0.430 | 0.320 | 0.455 | 0.068      | 0.238     |
| Professional      | Neural Only | 0.423   | 0.413   | 0.464 | 0.326 | 0.444 | 0.069      | 0.243     |
| Consumer (Crowdsourced) | Neural Only NIR-Finetuned | 0.728 | 0.710   | 0.756 | 0.627 | 0.772 | 0.144      | 0.733     |
| Professional      | Neural Only NIR-Finetuned | 0.595 | 0.581   | 0.641 | 0.554 | 0.648 | 0.125      | 0.664     |

Ethics

The user-based component (the crowdsourced query reformulation) in this paper was approved by the Australian National University Human Ethics Committee (Protocol 2020/726).

References

[1] Nina Van Goethem, Aline Vilain, Chloé Wynadham-Thomas, Jessika Deblonde, Nathalie Bosuyt, Tinne Lernout, Javiere Rebello Gonzalez, Sophie Quoilin, Vincent Melis, Dominique Van Beckhoven, Rapid establishment of a national surveillance of covid-19 hospitalizations in belgium, Arch. Public Health 78 (1) (2020) 121, https://doi.org/10.1186/s13012-020-00555-z, ISSN 2049-2258.
[2] Tinne Tuytelaars, Matthew B. Blaschko, Dusan Grujicic, Gorjan Radevski, Self-supervised context-aware covid-19 document exploration through atlas grounding, Proceedings of the NLP COVID-19 Workshop at ACL 2020, Online, 2020. Association for Computational Linguistics. URL https://openreview.net/pdf?id=v8ioFR4fgr.
[3] Jana Verma, Shashank Dubey, Aakash Deep Singh, Kushagra Agarwal, Sourojit Bhaduri, Rajesh Kumar Ranjan, Debasmita Das, Yatin Katyal, Information retrieval and extraction on covid-19 clinical articles using graph community detection and bio-emb embeddings, Proceedings of the NLP COVID-19 Workshop at ACL 2020, Online, 2020. Association for Computational Linguistics. https://openreview.net/pdf?id=W3Dzia1klp.
[4] Kirk Roberts, Tasmeen Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, William Hersh, TREC-COVID: Rationale and structure of an information retrieval shared task for COVID-19, J. Am. Medical Inform. Assoc. 27 (9) (2020) 1431–1436.
[5] John Wolohan, Estimating the effect of covid-19 on mental health: Linguistic indicators of depression during a global pandemic, Proceedings of the NLP COVID-19 Workshop at ACL 2020, Online, 2020. Association for Computational Linguistics. URL https://openreview.net/pdf?id=–2f70OXlGQMd.
[6] Suzanne Stevenson, Jai Agarwal, Alla Rahkovich, Exploration of gender differences in covid-19 discourse on reddit, Proceedings of the NLP COVID-19 Workshop at ACL 2020, Online, 2020. Association for Computational Linguistics. URL https://openreview.net/pdf?id=–v8ioFR4fgr.
[7] Ting-Hao Huang, Chieh-Yang Huang, Chien-Kuang Ding, Yin-Chia Hsu, Lee Giles, Coda-19: Using a non-expert crowd to annotate research aspects on 10,000 abstracts in the covid-19 open research dataset. Proceedings of the NLP COVID-19 Workshop at ACL 2020, Online, 2020. Association for Computational Linguistics. URL https://openreview.net/pdf?id=–X0kK8vdxzS.
[8] Sorosh Vosoughi, Jason Wei, Jerry Wei, Chengyu Huang, What are people asking about covid-19? A question classification dataset, Proceedings of the NLP COVID-19 Workshop at ACL 2020, Online, 2020. Association for Computational Linguistics. URL https://arxiv.org/pdf/2005.12522.pdf.
[9] Zhiyong Lu, Qingyu Chen, Alexis Allot. Keep up with the latest coronavirus research, Nature, 193 (2020). https://www.nature.com/articles/d41586-020-00694-1.
[10] Travis Goodwin, Dina Demner-Fushman, Kyle Lo, Lucy Lu Wang, William Hersh, Hoa Dang, Ian M Soboroff, Overview of the 2020 epidemic question answering track, in: Text Analysis Conference, 2020.
[11] Carolyn M. Clancy, Sherry A. Gled, Nicole Lurie, From research to health policy impact, Health Services Res. 47 (Feb 2012) 337–343, https://doi.org/10.1111/j.1475-6773.2011.01374.x.
[12] Andrea C. Tricco, Wasifa Zarin, Patricia Rios, Vera Nincic, Paul A. Khan, Marco Ghassami, Sanobor Dzber, Ba Pham, Sharon E. Strauss, Eienne V. Langlois, Engaging policy-makers, health system managers, and policy analysts in the knowledge synthesis process: a scoping review, Implement. Sci. 13 (1) (2018) 31, https://doi.org/10.1186/s13012-018-0717-x. ISSN 1748-5908.
and biased information about prostate cancer on youtube, Eur. Urol. 75 (4) (2019) 564–567, https://doi.org/10.1016/j.eururo.2018.10.056.

[14] Azhar Hussain, Syed Ali, Madhia Ahmed, Sheharyar Hussain, The anti-vaccination movement: A regression in modern cucum, Cureus 10 (7) (2018), https://doi.org/10.7759/cureus.2919 e2919, ISSN 2168-8184. URL https://pubmed.ncbi.nlm.nih.gov/30180724/.

[15] Rosenbeek Jon, Schneider Claudia, Dryhurst Sarah, Kerr John, Freeman Alexandra, Recchia Gabriel, van der Bles Anne Marthe, van der Linden Sander. Susceptibility to misinformation about covid-19 around the world, R. Soc. Open. Sci. 7 (2020).

[16] William R. Hersh, et al., Factors associated with success in searching Medline and applying evidence to answer clinical questions, J. Am. Med. Informa Assoc. 9 (2002) 283–295.

[17] Peifan Liu, Lamont B. Antieau, Yu. Hong, Toward automated consumer question answering: Separating consumer questions from professional questions in the healthcare domain, J. Biomed. Inform. 44 (6) (2011) 1032–1038, https://doi.org/10.1016/j.jbi.2011.08.008, ISSN 1532-0464. URL http://www.sciencedirect.com/science/article/pii/S1532046411001353.

[18] Vincent Nguyen, Maciej Rybinski, Sarvnaz Karimi, Zhenchang Xing, Pandemic literature search: Finding information on COVID-19, in: Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association, December 2020, pp. 92–97.

[19] Nils Reimers, Iryna Gurevych, Sentence-BERT: Sentence embeddings using Siamese Networks, in: EMNLP, Hong Kong, China, November 2019, pp. 3982–3992. URL https://www.aclweb.org/anthology/D19-1410.pdf.

[20] Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, Mike Gastford. Okapi at TREC-3, in: TREC, Gaithersburg, MD, US, 01 1995. https://trec.nist.gov/pubs/trec3/t3.proceedings.html.

[21] T. Armstrong, A. Moffat, W. Webber, J. Zobel. Improvements that don’t add up: Ad-hoc retrieval results since 1998, in: CIKM, Hong Kong, China, 2009, pp. 601–610.

[22] Wei Yang, Kuang Lu, Peilin Yang, Jimmy Lin, Critically examining the “neural hype”: Weak baselines and the additivity of effectiveness gains from neural ranking models, in: SIGIR, Paris, France, 2019, pp. 1129–1132. URL https://dl.acm.org/doi/10.1145/3331184.3331340.

[23] Sarvesh Soni, Kirk Roberts, An evaluation of two commercial deep learning-based information retrieval systems for COVID-19 literature, 2020. https://arxiv.org/abs/2007.07846.

[24] Rodrigo Nogueira, Kyunghyun Cho, Passage Re-ranking with BERT. arXiv:2007.07846, 2020. https://arxiv.org/abs/2007.07846.

[25] Jimmy Lin, Neural hype, justified! A recantation. ACM SIGIR Forum, 53, 2019. http://sigir.org/wp-content/uploads/2019/12/p088.pdf.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina 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, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, 2019, pp. 4171–4186.

[27] Peilin Yang, Hui Fang, Jimmy Lin, Amerin: Enabling the use of Lucene for information retrieval research, in: SIGIR, Tokyo, Japan, 2017, pp. 1253–1256.

[28] Ryan McDonald, George Brokos, Ion Androutsopoulos, Deep Relevance Ranking Using Enhanced Document-Query Interactions, in: EMNLP, Brussels, Belgium, 2018, pp. 1849–1860.

[29] Jimmy Lin, Neural hype, justified! A recantation. ACM SIGIR Forum, 53, 2019. http://sigir.org/wp-content/uploads/2019/12/p088.pdf.

[30] Lucy Lo Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jianning Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyan Liu, William Merrill, Paul Mooney, Dewey Murdock, Devvret Rishi, Jerry Sheehan, Zhilong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, Sebastian Kohlmeier. CORVID-19: The Covid-19 Open Research Dataset, in: ACL NLP-COVID Workshop, Online, 2020. https://arxiv.org/pdf/2004.10706.pdf.

[31] Sean MacAvaney, Arman Cohant, Nazi Gobarian, Sledge: A simple yet effective baseline for covid-19 scientific knowledge search, 2020.

[32] Galervo Jarvelin, Jaana Kekalainen, Cumulated gain-based evaluation of ir techniques, ACM Trans. Inf. Syst. 20 (4) (2002) 422–446, https://doi.org/10.1145/592415.592418, ISSN 1046-8188.

[33] Diego Molla, Christopher Jones, Vincent Nguyen, Pandemic literature search: Finding information on COVID-19, in: Working Notes of CLEF 2020, Thessaloniki, Greece, September 2020. CLEF 2020. http://cerc.univ-paris10.fr/WoC-2020/paper_119.pdf.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, Illia Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems vol. 30, Curran Associates Inc, 2017, pp. 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdfs.