S_Covid: An Engine to Explore COVID-19 Scientific Literature

Genoveva Vargas-Solar, Mehrdad Farokhnejad, Ratn Pranesh Raj, Davoud Amiri Mehr

S_Covid: An Engine to Explore COVID-19 Scientific Literature

Mehrdad Farokhnejad
Univ. Grenoble Alpes, Grenoble INP, CNRS, LIG
Grenoble, France
Mehrdad.Farokhnejad@univ-grenoble-alpes.fr

Raj Ratn Pranesh
Birla Institute of Technology
Mesra, India
raj.ratn18@gmail.com

Genoveva Vargas-Solar
Univ. Grenoble Alpes, Grenoble INP, CNRS, LIG-LAFMIA
Grenoble, France
genoveva.vargas@imag.fr

Davoud Amiri Mehr
National Institute of Genetic Engineering and Biotechnology
Tehran, Iran
davoudamirimehr@gmail.com

ABSTRACT

This paper introduces S_Covid, an end-to-end unsupervised learning-based question-answering engine for exploring COVID-19 scientific literature collections. S_Covid enables documents exploration for finding relevant research literature that most possibly contains information that can answer a user query. Thus, S_Covid pinpoints sentences out of research papers that can be possible answers to complex COVID-19 related user queries. We conducted experiments on 80,000 COVID-19 related papers collection. The paper shows statistically how the model performs but also through the feedback of real users. It also compares S_Covid with existing search engines addressing information retrieval of COVID-19 scientific literature.

1 INTRODUCTION

COVID-19 (CoronaVirus Disease 2019), announced by the World Health Organization (WHO) in March 2020, has become a global pandemic in a short period. Scientists around the globe are publishing and sharing new research preliminary results to understand and tackle the COVID-19 pandemic. Since the COVID-19 came in the knowledge of people (6-9 months), a considerable amount of work has been published. Results have been shared to promote the understanding of the disease and developing innovative methods to control it. On March 13, 2020, the Allen Institute for AI released the COVID-19 Open Research Dataset (CORD-19 [45]) in partnership with a coalition of research groups. The corpus currently contains over 100,000 scholarly articles, including over 80,000 full-text articles, about COVID-19 and corona virus-related research more broadly (e.g., SARS and MERS). A variety of sources including PubMed, a curated list of articles from the WHO, as well as preprints from bioRxiv and medRxiv, feed the collection. Exploring an ever-increasing amount of scientific papers that are produced about COVID-19 related topics is challenging for scientists. Indeed, the study presented in [20] shows that a large part of the time (about 23%) is spent by scientists in searching and reading research literature. Our work applies Natural Language Processing (NLP) and Natural Language Understanding (NLU) techniques to let the scientific community pore through articles for answers or evidence to COVID-19-related questions.

This paper introduces S_Covid an engine for exploring COVID-19 Open Research Dataset. When the user asks an initial query, S_Covid, not only returns a set of papers (like in a traditional search engine) but also returns sentences from the paper that are possible answers to the query. The user can review the sentences and decide on whether or not that paper is worth further reading. In our approach, we first discover several topics using LDA (Latent Dirichlet Allocation) and then use them to find a set of related articles for a given query. After that, we use this set to extract a collection of sentences that can contain the answer to the query. We then use sentence embedding and cosine similarity to select the most similar sentences. Finally, we select the top K articles which have highest number of related sentences.

We worked with biologists and physicians from '*Golestan University of Medical Sciences*' and developed an assessing method for data exploration algorithm. They tested our model on a collection of Kaggle tasks based on scientific questions developed with the World Health Organization and the National Academies of Sciences, Engineering, and Medicine to evaluate the accuracy and effectiveness of our documents exploration engine. To assess S_Covid we ran a systematic comparative analysis of S_Covid and various baseline models. As a result of our experiments and comparative study we conclude that through S_Covid, scientists can foster knowledge exploration and efficient gathering of evidence for scientific hypotheses.

The rest of the paper is organised as follows. Section 2 provides a detailed description of the approach behind the engine S_Covid. Section 3 describes our experimental setting including the dataset and its preprocessing steps. It also describes the assessment method that we adopted for S_Covid results. It is a scoring method used by scientists to assess the results provided S_Covid for a set of representative queries. Section 4 provides an overview of various baseline models along with a systematic comparative study of various model's performance in the experiment and presents an in-depth analysis of the results. Section 5 summarises the related work done in the field of COVID-19 information retrieval systems. Section 6 concludes the paper and discusses future work.

2 EXPLORING COVID-19 SCIENTIFIC LITERATURE WITH S_COVID

Figure 1 shows the general pipeline implemented by S_Covid to processing documents. The pipeline consists of three phases: (i) topic modelling for indexing the initial documents of the data collection for building a corpus. Then, given a query (ii) data exploration and relevant paper extraction (Searching Unit); and (iii) selecting the top-k papers that answer a query by retrieving and ranking candidate sentences from papers which possibly

answer the user query, and ranking of selected papers (Ranking Unit).

2.1 Phase-A: Building the S_Covid corpus

S_Covid first processes and indexes the papers collection for discovering a set of topics (step-1) and then assigning relevant topics to the paper according to their content (step-2).

2.1.1 ***Step-1 Generating a vocabulary.*** The COVID-19 dataset consists of a large number of scientific papers containing text in the form of words, sentences and paragraphs. To explore the documents (i.e., papers), it is necessary to generate a model of their content. In our approach, this model is a vocabulary. The vocabulary is particularly useful in the case of the COVID-19 because there is for the time being no official vocabulary about this topic.

We first pre-processed documents to generate vector representations for words i.e., word embedding. The word embedding representation is the one that best captures words contextual, lexical, and sentimental characteristics.

Once texts had been pre-possessed, we used the `Gensim Phrases` Python library [38] to automatically detect common phrases (bigrams) from each paper in the corpus. For example, sentences like 'infectious disease' or 'public health' must occur together. We applied the `skip-gram` method to predict the context of words. The principle of the method `skip-gram` is the use of a word for predicting its target context.

We trained a `Word2vec` model [33] which is a two-layer neural net that processes text by "vectorizing" the words in the document to build a vocabulary. The underlying assumption of `Word2Vec` is that two words sharing similar contexts also share a similar meaning and vector representation in the model. We set the dimensionality of the feature vectors to 300 and trained the model for 15 epochs. Finally, the resulting `Word2Vec` model was stored for future use in the document exploration pipeline.

2.1.2 ***Step-2 Topic modelling.*** The COVID-19 dataset used in our model is an unstructured text corpus. This characteristic made it difficult to extract relevant and desired information from the dataset. To tackle this challenge, we assume that a way to understand large text documents is by their topics. The statistical process of learning and extracting topics from documents collections is called topic modelling [19] [1].

In the S_Covid pipeline, we used the topic modelling method called Latent Dirichlet allocation (LDA) [4]. It is a Bayesian model to classifying discrete data having uncorrelated topics. Through topic modelling, we represented each scientific document in the corpus as a distribution of topics described as a distribution of words. Being an unsupervised machine learning method, LDA automatically analyses a text for clustering the words from a given set of documents.

The number of topics plays a very crucial role in deciding the performance of the model computed with LDA. After several iterations, we discovered that 50 is the optimum number of latent semantic topics that can be extracted from the COVID-19 literature corpus. For producing fine-grained results, the discovered topics were specified and refined with minimum overlapping. Then, having a set of topics assigned to each paper (i.e., a distribution over words), we organised papers in an LDA space, namely a simplex. The dimensionality of the space depends on the number of topics. Depending upon a topic's words distribution in a paper, each paper in the corpus is closer to the topics that represent it more strongly.

[1] Through topic modelling, one can identify the topics that best describe a set of documents. Knowing the topics representing the content of documents can be useful for search engines and customer service automation.

2.2 Phase-B: Exploring COVID-19 scientific articles

The objective of this phase is to extract, out of large COVID-19 scientific literature corpus, the relevant papers possibly containing the answer to a user query about COVID-19.

A query sentence (consisting of words/terms) is assigned with relevant topics and represented as a query vector. The semantic of a query is: *which are the papers closest to these topics?* A ranked result set is computed by determining which are the papers of the corpus containing the most probable topics that are closest to those of the query and ordering them.

For this, we first calculate the probability distributions of 50 topics over a given scientific literature corpus "P_p". The probability "P_p" represents how well each topic describes a paper content. We also calculate the probability distribution of topics over the user query "P_q".

Then, we calculate the similarity score "S" to measure the similarity between two probability distributions i.e., "P_p" and "P_q" representing the paper and query probability vectors respectively. The similarity score is given as: **S = 1 - Jensen–Shannon distance**. Jensen–Shannon distance is the square root of the Jensen-Shannon divergence, where Jensen–Shannon divergence is a method of measuring the similarity between two probability distributions.

$$S = 1 - \sqrt{\frac{D(P_p \parallel m) + D(P_q \parallel m)))}{2}} \tag{1}$$

Where "m" represents the mean of "P_p" and "P_q", D represents Kullback-Leibler divergence and "S" represent the similarity score between a user query and the research paper in the corpus. The range of "S" lies between 0 to 1. The value of the similarity score determines the similarity between the topic distribution of a paper and a query. Consequently, the probability of a paper represents its relevance concerning the query.

Now, using the similarity score we select the top relevant candidate papers to build a `related articles set`. This set contains papers that are the closest to the user query. We experimented with various values of similarity score ranging from 0.2 to 1. The domain experts from biology and medical field manually reviewed the quality of extracted papers using different similarity score. Finally, they agreed to considered articles having a similarity score more than **0.5** on a scale of 0 to 1 as a candidate for being in the related articles set.

2.3 Phase-C: Extracting candidate answers and ranking relevant papers

The objective of this phase is to extract out of a relevant papers list computed in Phase-B, the top-k papers possibly containing the answer to a user query. This phase consists of two steps described next.

2.3.1 ***Step 1: Extracting candidate answer sentences.*** This step converts each relevant paper in the `top related papers set` computed in Stage-B, into a set of "representative" sentences. We create a list of tuples *Top_relevant_paper_senten*

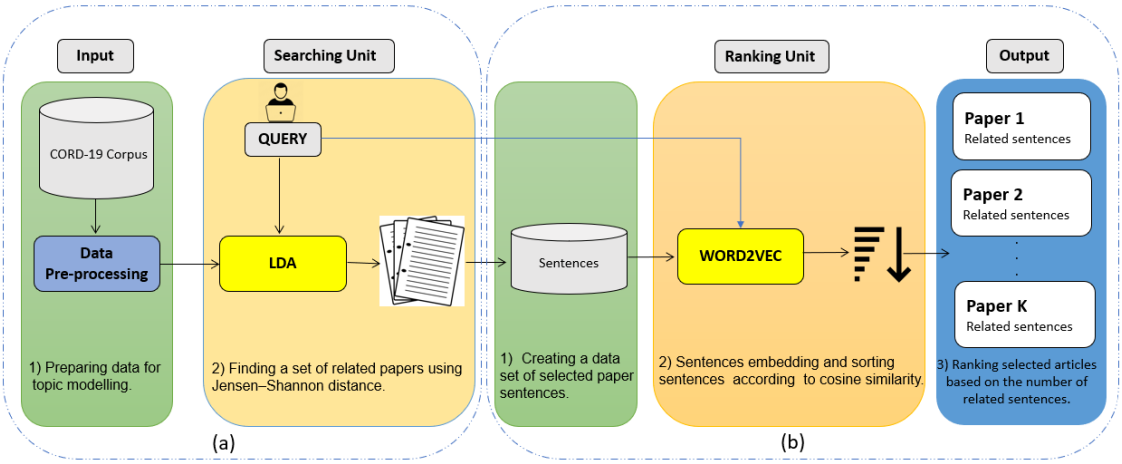


Figure 1: S_Covid exploration pipeline: a-1) Modelling topics representing the content of the corpus documents; a-2) Exploring COVID-19 articles to extract top relevant papers given a query. b) Extracting and scoring candidate answer sentences for every relevant paper according to the query-sentences similarity and ranking the relevant papers based on the sentences score.

ces $:< sentence, paperID >$. Each sentence in the list is converted into a vector representation using our previously trained `word2vec` model.

Given the query vector representing the user query, we use a Cosine similarity, to calculate a similarity score between the query vector and all the elements of the *Top_relevant_paper_sentences* list. Based on the similarity score, we choose sentences with the similarity score above a fixed threshold and then sort the sentences in a manner such that the sentences with higher similarity score are placed above in the list of candidate answers. Our experiments showed that sentences having a Cosine similarity score equal or above 0.5 on a scale of 0 to 1, can be considered candidates for an answer to the query.

2.3.2 ***Step 2: Ranking relevant papers.*** Once we have top candidate papers, we group the sentences with their respective research paper using $< paperID >$. So, now for each paper having a set of answer candidates, our model finally re-ranks the papers based on the number of answer candidates of each paper. For example: If papers 'A' and 'B' have respectively ten versus seven, candidate sentences-answers then 'A' is ranked above 'B' in the ranking of relevant papers list. The idea behind this method is that if a paper has a high number of candidate sentences, it means that its content is contextually more aligned with the user query. Consequently, the paper should be ranked higher in the relevant paper ranking.

Algorithm 1 presents the pseudo-code of the S_Covid algorithm. Given the dataset consisting of 80,000 COVID-19 papers. For a given user question input:- Query, the number of relevant papers user wants to extract:- k; S_Covid outputs:- Top-k related paper to the Query.

In **Stage-A** (refer to line 1), first creating the LDA topic model to discover 50 topics- *LDA*, then using CORD-19 [45] dataset to train the LDA model. Followed by calculating the probability distribution of topics over each paper in dataset- *doc_topic_dist* and also for the user question- *query_topic_dist*. For each paper in dataset calculating similarity score between user query and the research paper probability distribution using Jensen–Shannon

Algorithm 1: S_Covid algorithm

Data: 80000 COVID-19 articles
input :(Query , k , Data)
output:Top-k related paper to the Query

1 STAGE-A;
2 $lda \longleftarrow LatentDirichletAllocation(n\_components = 50)$;
3 lda.fit $(data\_vectorized)$;
4 $doc\_topic\_dist \longleftarrow lda.transform(data\_vectorized)$;
5 $query\_topic\_dist \longleftarrow lda.transform(query\_vectorized)$;
6 **foreach** *article in the Data* **do**
7 $similarity\_score(article) \longleftarrow 1 - Jensenshannon(query\_topic\_dist, doc\_topic\_dist[article])$;
if $similarity\_score(article) > 0.5$ **then**
8 $top\_nearest\_articles \longleftarrow article$;
9 **end**
10 **end**
11 STAGE-B;
12 $sentences \longleftarrow extract\_sentence(top\_nearest\_articles)$;
13 $sentences\_vectorized \longleftarrow word2vec(sentences)$;
14 $query\_vectorized \longleftarrow word2vec(Query)$;
15 **foreach** *sentence in the sentences_vectorized* **do**
16 $similarity \longleftarrow cosine\_similarity(query\_vectorized, sentence)$;
17 **if** $similarity > 0.5$ **then**
18 $top\_similar\_sentence \longleftarrow sentence$;
19 **end**
20 **end**
21 $sorted\_sentences \longleftarrow sorted(top\_similar\_sentence)$;
22 $candidated\_result \longleftarrow Group\_by\_paperid(sorted\_sentences)$;
23 $Top\_k\_related\_paper \longleftarrow$ Select top k articles which have most sentences in $candidated\_result$;

distance- *similarity_score (article)*. Filtering the papers with condition on similarity score- *similarity_score (article) > 0.5*. In stage-A, we created a set of most relevant papers- *top_nearest_articles*.

In **Stage-B** (refer to line 11) for papers in *top_nearest_articles* we extract a list of sentences- *sentences*, vectorize them using our trained `Word2vec` model - *sentences-vectorized*, vectorizing user question- *query_vectorized*. Then for each sentence, calculating cosine similarity between the sentence vectors and query vector- *similarity*. Using a condition on cosine similarity score- $similarity > 0.5$, selecting top sentences as candidate answer sentences- *top_similar_sentence*. Followed by sorting the candidate sentences based on the cosine similarity score- *sorted_sentences*. Then grouping the sentences with their respective research paper using paper ID- *Group_by_paperid* to have candidate relevant papers- *candidated_results*. Finally **Stage-C** ranks the papers in *candidated_results* such that the papers with greater number of answer candidate sentences are at the top i.e. more relevant and presenting *Top_k_related_paper*.

3 EXPERIMENTS

This section describes the experimental setting for the assessment of our approach and a comparative study of various baseline models along with our proposed S_Covid model.

3.1 Experiment settings

We performed our experiments on the Google Colab with 25GB NVIDIA GPU RAM, Xeon Processors @2.3Ghz, CUDA-10.0. We used the COVID-19 Open Research Dataset (CORD-19) [45] which is released by Allen Institute for AI and contains more than 100,000 scholarly articles, including over 80,000 with full text, about COVID-19 and coronavirus.

3.1.1 ***Dataset***. In our experiment we used CORD-19 [45] (COVID-19 Open Research Dataset) dataset. The corpus currently includes over 100,000 scholarly articles and updates weakly. The Allen Institute for AI published this dataset for the global research community. The objective was to let the community apply techniques in natural language processing/natural language understanding and Artificial Intelligence techniques to generate new insights about this infectious disease. CORD-19 is a collection of research papers published by various international publishing bodies.

The sources of papers come from the databases PMC, bioRxiv, medRxiv, Elsevier, Springer Nature and WHO. The metadata involves papers published by bioRxiv, medRxiv, PMC, WHO and individual publishers. The data was provided in JSON format which we converted into CSV files with the schema columns: paper_id title, authors affiliations, abstract, text, bibliography, raw_authors and raw_bibliography. We further preprocessed the data to produce a clean and structured dataset before using it in S_Covid.

3.1.2 ***Data pre-processing***. We used the version-52 of original CORD-19 [45] dataset which contained around 116,005 COVID-19 related research papers. The dataset schema consisted of the following attributes: *paper_id*, *body_text*, *methods*, *results*, *cord_uid*, *source*, *title*, *doi*, *pmcid*, *pubmed_id*, *license*, *abstract*, *publish_time*, *authors*, *journal*, *mag_id*, *who_covidence_id*, *arxiv_id*, *pdf_json_files*, *pmc_json_files*, *url*, *s2_id*, *is_covid19* and *publish_year*.

We did a first filtering process choosing 8 features from the dataset schema that represented papers content, i.e., *paper_id*: unique paper id for each article, *title*: title of the paper, *abstract*: abstract of the article, *body_text*: full length text of the article, *doi*: DOI id of the article, *url*: hyperlink to the article's publication website *publish_year*: publication year of the article and *is_covid19*: contains either true (COVID-19 only article) or false value.

We went further into an attribute engineering process to add an attribute named '*complete text*' to the schema. This new attribute concatenates three attributes of the dataset schema - Title, Abstract and Body text. In that way, we simplified the schema. We removed those research papers which were not written in English. We also removed the abstract-only papers [2]. Finally, we end up with a collection of 80,000 COVID-19 research papers.

We then performed the text preprocessing on the '*complete text*'. We used `ScispaCy` [34], a Python package containing spaCy models for processing biomedical, scientific or clinical text. We removed the stop words and performed word lemmatization on the text data. We also removed some common unnecessary words such as author, figure, copyrights, license, fig from the '*complete text*'. But we kept important information such as citation numbers in papers intact so that user would get a complete answer without any missing values. We used '*complete text*' to train `word2vec` model and to extract sentences from each paper.

3.2 S_Covid data exploration algorithm assessing method

We assessed our algorithm in every step of progress, helping the machine to evolve by continuous feedback. For the case of the COVID related topics, there is not enough standardized question answering dataset. So, we worked with three field experts of medicine and biology from National Institute of Genetic Engineering and Biotechnology, Tehran, Iran and Golestan University of Medical Sciences to assess the S_Covid algorithm and the baseline models used to compare S_Covid (see next section). To quantify our evaluation, we devoted scores related to each result according to their relevance to our queries. In this regard, we allocate three scores for our findings, as mentioned in the following formula:

$$Evaluation(article) = \begin{Bmatrix} 1 & Relevant \\ 0 & PartiallyRelevant \\ -0.5 & NotRelevant \end{Bmatrix} \quad (2)$$

In the formula, the score of -0.5 is as a penalty for completely irrelevant articles, zero for papers that may be a candidate for parts of our findings—partially relevant—and score one is concerning the documents that are completely related to our questions. We calculated the evaluation score of all top k articles extracted by the models for each given query as:

$$Evaluation\_score(Q_j^m) = \frac{\sum_{i=1}^{k} Evaluation(topkarticle_j^m(i))}{k} \quad (3)$$

where m represents a model, i.e.:
m ∈ { LDA , LDA+BM25, LDA+Whoosh,Google, kdcovid.nl,covidex.ai, S_Covid } and j is a question id which is $1 \leqslant j \leqslant 10$.

To have a standard set of diverse queries, we chose a subset consisting of 30 questions (see Table 2 and 4) out of 100 questions

[2]Research [28] has proven that using full text for information retrieval is more effective than just using the abstract section of the research paper.

Table 1: COVID-19 Tasks

Task Id	Task details
$Task1$	What is known about transmission, incubation, and environmental stability?
$Task2$	What do we know about COVID-19 risk factors?
$Task3$	What do we know about vaccines and therapeutics?
$Task4$	What do we know about virus genetics, origin, and evolution?
$Task5$	What has been published about medical care?
$Task6$	What do we know about non-pharmaceutical interventions?
$Task7$	What has been published about ethical and social science considerations?
$Task8$	What do we know about diagnostics and surveillance?
$Task9$	What has been published about information sharing and inter-sectoral collaboration?

Table 2: COVID-19 Questions

Id	Question
$Q1$	The incubation period of corona virus disease
Q_2	The effect of seasons on transmission of COVID-19
Q_3	Risk factors for severe disease and death
Q_4	Efforts to develop a SARS-CoV vaccine
Q_5	Risk-reduction strategies
Q_6	Misinformation relate to COVID-19
Q_7	The economic impact of COVID-19
Q_8	Use of diagnostics markers to detect early covid-19 disease
Q_9	Protocols for screening and testing of covid-19
Q_{10}	Outcomes data for COVID-19 after mechanical ventilation

of Kaggle CORD-19-research-challenge[3] (CORD-19). The challenge consists of 17 sub-tasks out of which 9 sub-tasks (refer table 1) were related to information retrieval. Each of the 9 sub-tasks consists of 5-10 questions. We selected 3-4 questions for each of the 9 tasks. The selection pattern of our choice was according to two criteria. First, we chose topics with enough variety to cover all aspects of the pandemic. Second, we selected questions that engaged scientific minds about the COVID-19. Since, out of 100 questions, various questions were overlapping and repetitive. To overcome this issue, our biology and medical experts identified 30 most interesting questions which were most distinctive.

To illustrate our scoring method, we show three results in response to a COVID-19 question. As given in the table 2, let us take "the effect of seasons on the transmission of COVID-19" as one of searched query item. The following papers are the three of our first five related paper extracted by the algorithm.

(1) "Climate effect on COVID-19 spread rate: an online surveillance tool [5]."
(2) "Projecting the transmission dynamics of SARS-CoV-2 through the post pandemic period [22]."

[3]https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

Table 3: Related sentences extracted by S_Covid for the questions in the table 2

Id	S_Covid Answer
Q_1	In our analysis of 44 patients with clear contact history, we found that the mean incubation period of COVID-19 was 8 [1].
Q_2	Our findings of decreased replication and spread rates of COVID-19 in warm climates may suggest that the inevitable seasonal variance will alter the dynamic of the disease spread in both hemispheres in the coming months [5].
Q_3	Evidence from China, Italy and the USA indicates that older individuals, males and those with underlying conditions, such as CVD, diabetes and CRD, are at greater risk of severe COVID-19 illness and death [11].
Q_4	Here, we discuss therapeutic and prophylactic interventions for SARS-CoV-2 with a focus on vaccine development and its challenges [2].
Q_5	In addition, we provide suggestions to aid further development of epidemic prevention and control strategies, and scientific decision-making [44].
Q_6	The World Health Organization have emphasized that misinformation - spreading rapidly through social media - poses a serious threat to the COVID-19 response [23].
Q_7	We identify a total of 35 potential determinants that describe a diverse ensemble of social and economic factors,including healthcare infrastructure, societal characteristics,economic performance, demographic structure etc [42].
Q_8	Those findings indicate that our N antigen assay is an accurate, rapid, early and simple diagnosis method of COVID-19 [13].
Q_9	We established routines for SARS-CoV-2 RNA extraction-free single-reaction RT-qPCR testing [40].
Q_{10}	All patients demonstrated stable physiology and ventilation for the duration of shared ventilation [26].

(3) "Excess cases of influenza suggest an earlier start to the corona virus epidemic in Spain than official figures tell us: an analysis of primary care electronic medical records from over 6 million people from Catalonia [12]."

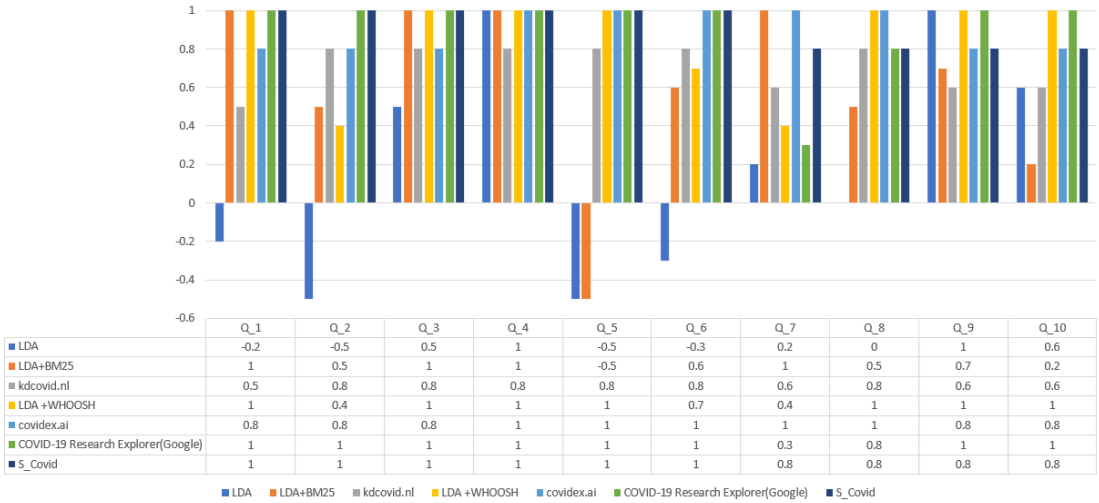
The first one gained score one because it has relevant data related to our query. The second article gained zero because it "seems" to be related to our questions, and the third one gained -0.5 score as a penalty because of its irrelevant result. Indeed, it is not about our query.

4 COMPARATIVE SURVEY OF S_COVID WITH BASELINE MODELS

We performed an in-depth quality analysis of S_Covid model by comparing its performance with the existing tools and baseline models. We used a set of COVID19 related questions shown in table 2 and table 4 for the comparative study.

4.1 Search Engines

4.1.1 ***COVIDEX****:*. It is a COVID-19 research papers' search engine [49]. It leverages the BM25 [39] algorithm and T5 [36] language model fine-tuned on medical text data for relevant paper



	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_10
LDA	-0.2	-0.5	0.5	1	-0.5	-0.3	0.2	0	1	0.6
LDA+BM25	1	0.5	1	1	-0.5	0.6	1	0.5	0.7	0.2
kdcovid.nl	0.5	0.8	0.8	0.8	0.8	0.8	0.6	0.8	0.6	0.6
LDA +WHOOSH	1	0.4	1	1	1	0.7	0.4	1	1	1
covidex.ai	0.8	0.8	0.8	1	1	1	1	1	0.8	0.8
COVID-19 Research Explorer(Google)	1	1	1	1	1	1	0.3	0.8	1	1
S_Covid	1	1	1	1	1	1	0.8	0.8	0.8	0.8

Figure 2: Question wise comparative evaluation of S_Covid against baseline models based on *Evaluation_score* (refer 3).

retrieval and ranking. It also uses a pre-trained BioBERT [25] model to extract and highlights the interesting sentences in each paper based on the user query.

4.1.2 ***KDCOVID****:.* It is a tool for exploring COVID-19 research dataset [4]. It uses the similarity between the user query and sentences in the full text of papers in CORD19 corpus using a similarity metric derived from BioSentVec [7] to find relevant papers and highlight key points. Higher the similarity between the query and a particular paper sentence, higher will be the ranking of that paper in terms of relevance.

4.1.3 ***COVID-19 Research Explorer (Google)****:* It is based on neural networks for natural language understanding [5]. This model is trained on 100,000+ scholarly articles on COVID-19. It is a supervised model powered by semantic search to help the model understand the context of the query. It encodes the query and documents as vectors and then performs retrieval by looking for the document vectors that are most similar to the query vector using k-nearest neighbour retrieval.

4.2 Baseline Models

We implemented different LDA based information retrieval models. We coupled LDA with various searching algorithms (i.e. BM25 and WHOOSH) to design COVID-19 literature exploration model such as (LDA, LDA-BM25 and LDA-WHOOSH). In our study, We compared and analysed the performances of these models against our proposed S_Covid model.

4.2.1 ***LDA****:.* For this baseline model, we used LDA topic modelling for finding relevant papers. First, it calculates the probability distribution of topics over the papers in the corpus and the given query. Then, it uses similarity score **(S = 1 - Jensen–Shannon distance)** based on Jensen–Shannon distance to select and rank the papers based on their relevance for the query. For the experiment, we took top relevant papers based on the similarity score. Papers with score equal or above 0.5 on a scale of 0 to 1 were considered as relevant papers and they were ranked based on the score.

4.2.2 ***LDA-BM25****:* For this baseline model, we used the same LDA topic modelling for finding relevant papers. Besides, in this model, we ranked again the relevant papers using the scoring function- Okapi BM25 [39]. BM25 is a bag-of-words retrieval function that ranks COVID-19 research papers based on the user query terms appearing in each document, regardless of their proximity within the papers.

4.2.3 ***LDA-WHOOSH****:.* In this baseline model, we used the LDA topic modelling and Whoosh[6] to find and rank relevant papers. First, it selects the papers using LDA and the similarity score **(S = 1 - Jensen–Shannon distance)**. Then it uses Whoosh to create an index for selected papers and, based on the query, it searches the index to find the relevant papers. Finally, the relevant papers are sorted using BM25, the Whoosh default ranking algorithm.

4.3 Experimental comparison

The experimental comparison consists of following three steps:

(1) **Step-1:** Each model outputs a set of papers ranked according to the relevance of a paper for a given query. We selected top 5 (k=5) papers from the set for each model.
(2) **Step-2:** Using our scoring method experts assign a score to each of the paper extracted by that model based on the given query. The expert thoroughly reads the paper's title and abstract to determine the quality of the paper and its relevance for a given query.
(3) **Step-3:** For each paper, we received three scores, each one corresponding to an expert. We finally assigned the most frequent score (0, 1 or -0.5) to the paper. Then, for comparing different models performance, we calculated the average performance score of each model based on the score obtained by the papers by using the *Evaluation_score* formula 3.

For evaluating the performance of the seven models (engines, baselines and S_Covid), experts manually reviewed 1050 papers (35 papers for each of the 30 questions). In some cases, experts also considered reading the extracted research paper's full text for

[4]http://kdcovid.nl/about.html
[5]https://covid19-research-explorer.appspot.com

[6]Whoosh is a python library for indexing text and then searching indexes https://pypi.org/project/Whoosh

a better understanding of why they were relevant for a query and making their final decision. A complete comparative evaluation of S_Covid against baseline models for 10 randomly selected questions (out of 30 questions), where each question belongs to one of the 9 Kaggle's challenge task (refer table 1) is shown in Figure 2. A detailed analysis of the results is given next.

4.4 Results and Discussion

This section summarises the results of the experiments and conducted a detailed analysis of various models performance. The figure 3 shows the overall performance of S_Covid and the baseline models over 30 selected COVID-19 questions. We can see that S_Covid outperformed other models in terms of average paper quality. One thing to be noticed is that the performance score of 'COVID-19 Research Explorer (Google)' and 'coveidex.ai' is very close to S_Covid. To understand this we can refer to figure 2 which shows question wise performance of each model.

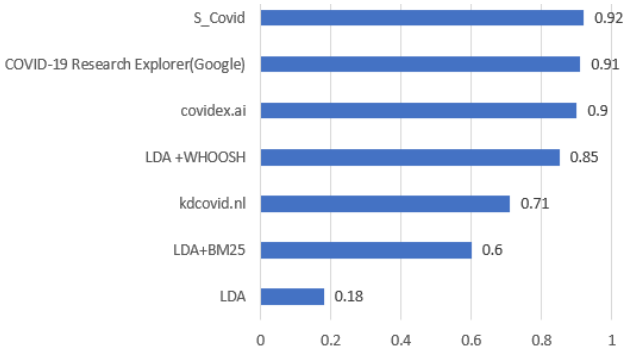


Figure 3: Overall performance score obtained by baseline models and S_Covid

In our survey, we found that almost all algorithms faced problem while extracting answer for the question Q_2 (see 2). One possible reason might be that since the prevalence of COVID-19 in all countries, there is no noticeable seasonal change, hence there not any published work which can answer the query. For question Q_7, also there is a similar reason: because the COVID-19 outbreak is a relatively new phenomenon in the world, there is not enough publication in our database related to its economic impacts on people and countries as of now. In the later update, we can evaluate these search algorithms concerning noted questions.

As shown in figure 2, S_Covid achieves much more accurate results than other search methods. Why did COVID-19 Research Explorer (Google) perform better with respect to S_Covid in questions Q_9,Q_{10}? Because they have a larger dataset (100,000+ papers). As a result, they can retrieve more papers than S_Covid. But having a larger dataset might be disadvantageous, for question Q_7. In this case, COVID-19 Research Explorer (Google) performance dropped because of the fourth retrieved article "A national prospective cohort study of SARS/COV2 pandemic outcomes in the U.S.: The CHASING COVID Cohort [39]" which is not related to the query. Therefore, by gaining a -0.5 penalty, its overall result falls to 0.3. For Q_7, coveidex.ai gave better results then S_COVID and COVID-19 Research Explorer (Google), the S_COVID scored bit low because the fourth paper in its result ("COVID-19 outbreak: Migration, effects on society, global environment and prevention [6]") is partially related to our query, so do not gain perfect score 1. On the other hand, For questionQ_2 coveidex.ai got zero scores because of the retrieved article "Modeling strict age-targeted mitigation strategies for COVID-19 [10] which is partially related to our query whereas S_Covid and COVID-19 Research Explorer (Google) performed relatively better.

Finally, we have also showcased S_Covid's capability of finding the most relevant candidate answer sentences (see Table 3 and 4) out of selected papers. This shows that the quality of related output papers is not the only advantage of S_COVID. Our application response to question Q_1 in the table 2 is shown in the figure 4. This app brings related papers and the most significant sentence of it [7]. The user can customise and refine the results by year of article publication and selecting only COVID-19 associated articles, which have terms like COVID-19, Coronavirus, SARS-CoV-2 and 2019-nCoV. The application can also be customised based on the number of relevant papers and sentences the user wants to see. It is clear that finding and reading-related papers are one of the most critical but time-consuming processes of scientific work. It is mainly about the researchers who work on systematic reviews and meta-analysis. Therefore, such a "helper" like S_Covid can facilitate research processes and drive the force of researchers toward scientific innovation rather than just wandering and being lost in an immense amount of scientific papers.

5 RELATED WORK

This section summarises research work in information retrieval and question-answering proposal devoted to searching and exploring COVID-19 scientific literature.

Information retrieval is the most researched area when it comes to using NLP techniques to build a model using CORD-19 dataset. In [49], the author proposes a neural search and ranking engine for COVID-19 literature exploration based on T5 [36] language model fine-tuned on the medical text. It leverages the BM25 algorithm for ranking the documents, followed by reranking using T5 model. The model also highlights the most relevant sentences of each research paper using the pre-trained BioBERT [25] unsupervised model.

A number of research organisations have launched online web applications to provide a platform to the scientist and research to explore and understand COVID-19 research literature. The COVIDSCHOLAR[8] is a web application that adapts the MATSCHOLAR [46] system for exploring and searching relevant COVID-19 research papers based on entity-centric queries.

The KDCOVID[9] also presents a similar document searching framework which uses BioSentVec [7] to encode the query sentences and scientific literature followed by KNN search to find the relevant papers. Papers ranking was done based on the similarity score of the query vector and the document sentences vector. The key sentences were also highlighted in the paper. This also uses DisGeNET [35] to represent the relation between genes and disease in the form of a knowledge graph.

Plenty of work has adopted question-answering approaches based on the CORD-19 dataset. In [43], the authors presented a COVID-19 specific question-answering dataset built using CORD-19 [45]. The paper also presents a performance analysis of various language models for question-answering the same dataset.

The Google AI team also developed an NLU-powered tool to explore COVID-19 research papers- COVID-19 Research Explorer

[7] In the time of writing this paper, we are working on its user-friendly-related features, so figure 4 is just an example of its functionality on Google Colab.
[8] https://covidscholar.org
[9] http://kdcovid.nl/about.html

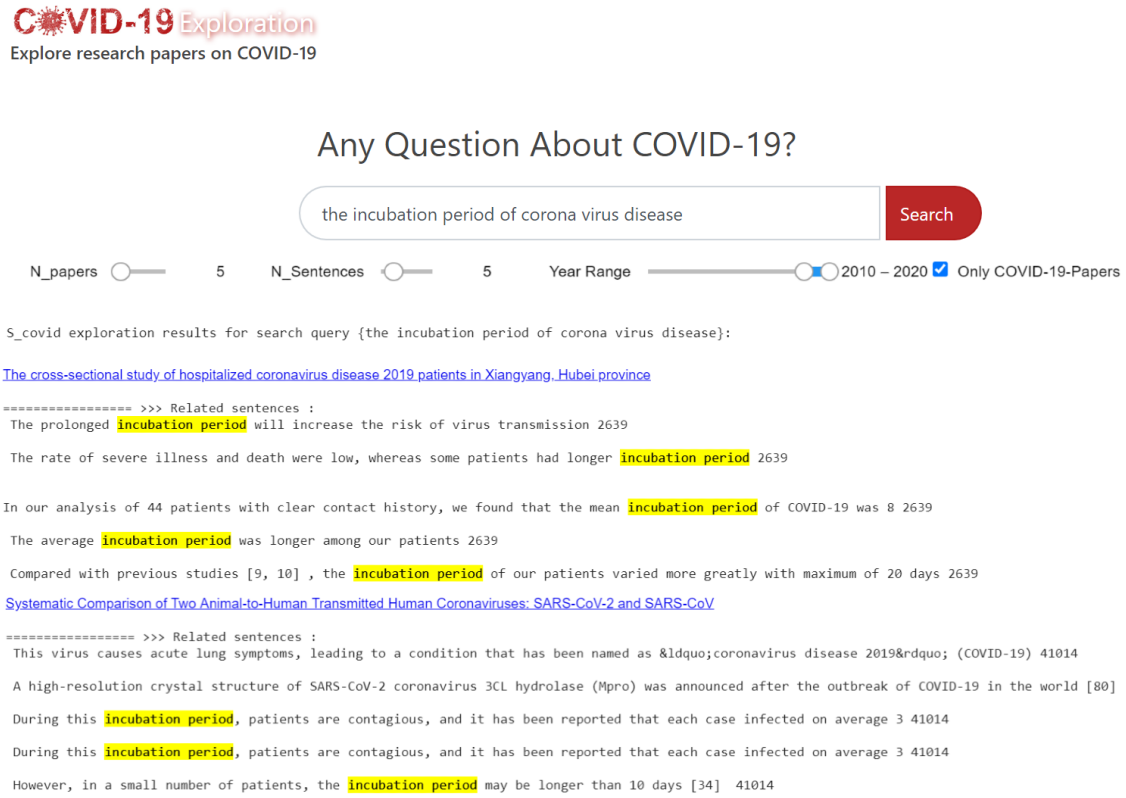


Figure 4: Screenshot of our S_Covid application, which builds on LDA and our rankings algorithm.

[10]. It is a framework for searching and ranking relevant papers. It also helps users to find the portion of each research paper related to the query. COVIDASK[11] and AUEB system[12] also present a question-answering framework that present user answer snippets for their query.

Existing work has some limitations. For example, in [43] the dataset was small in size. The performance of the language model shows that a larger question-answer dataset is necessary for training a supervised model. The paper [49] reports the use of BioBERT to generate a vector of words present in sentences. Since BioBERT's vocabulary might not have those words used within the COVID-19 papers, the contextual understanding and the word/sentence vector representation quality can decrease. In our proposal, we addressed this problem by training a COVID-19 specific `word2vec` model using CORD-19 dataset. The `word2vec` model generates a more representative contextual and semantic vector representation of all the COVID-19 terms and words present in the dataset. Also, the large transformer based information retrieval models such as [49] and Google's COVID-19 Research Explorer, tends to have a higher computational resource requirement as compared to S_covid which have comparatively simpler architecture design.

6 CONCLUSION AND FUTURE WORK

This paper introduces the S_Covid engine that assists subject-matter experts in their searching for papers given specific queries. Our work evaluates the existing COVID-19 literature search engines performance using our proposed S_Covid model. With the help of medical domain experts, the paper identifies the strengths and weaknesses of existing models and how our model addressed their limitations.

As future work, we aim to develop several data exploration techniques such as query morphing and queries as answers and query by examples mechanisms can provide data collection exploration tools for data science processes.

[10]https://covid19-research-explorer.appspot.com
[11]https://covidask.korea.ac.kr
[12]http://cslab241.cs.aueb.gr:5000

7 ACKNOWLEDGMENTS

This work was partially funded by the Iranian Ministry of Science, Research and Technology through the fellowship of Mehrdad Farokhnejad.

REFERENCES

[1] Jinwei Ai, Junwen Chen, Yong Wang, Xiaoyun Liu, Wufeng Fan, Gaojing Qu, Meiling Zhang, Shengduo Polo Pei, Bowen Tang, Shuai Yuan, et al. 2020. The cross-sectional study of hospitalized coronavirus disease 2019 patients in Xiangyang, Hubei province. *MedRxiv* (2020).
[2] Fatima Amanat and Florian Krammer. 2020. SARS-CoV-2 vaccines: status report. *Immunity* (2020).
[3] Julien Arino, Nicolas Bajeux, Stephanie Portet, and James Watmough. 2020. Assessing the risk of COVID-19 importation and the effect of quarantine. *medRxiv* (2020).
[4] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[5] Gil Caspi, Uri Shalit, Soren Lund Kristensen, Doron Aronson, Lilac Caspi, Oran Rossenberg, Avi Shina, and Oren Caspi. 2020. Climate effect on COVID-19 spread rate: an online surveillance tool. *medRxiv* (2020).
[6] Indranil Chakraborty and Prasenjit Maity. 2020. COVID-19 outbreak: Migration, effects on society, global environment and prevention. *Science of the Total Environment* (2020), 138882.
[7] Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. BioSentVec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 1–5.
[8] Zhimin Chen, Lin Tong, Yunlian Zhou, Chunzhen Hua, Wei Wang, Junfen Fu, Qiang Shu, Liang Hong, Huiqing Xu, Zhen Xu, et al. 2020. Childhood COVID-19: a multicentre retrospective study. *Clinical Microbiology and Infection* 26, 9

(2020), 1260–e1.
[9] Yichun Cheng, Ran Luo, Kun Wang, Meng Zhang, Zhixiang Wang, Lei Dong, Junhua Li, Ying Yao, Shuwang Ge, and Gang Xu. 2020. Kidney disease is associated with in-hospital death of patients with COVID-19. *Kidney international* (2020).
[10] Maria Chikina and Wesley Pegden. 2020. Modeling strict age-targeted mitigation strategies for COVID-19. *arXiv preprint arXiv:2004.04144* (2020).
[11] Andrew Clark, Mark Jit, Charlotte Warren-Gash, Bruce Guthrie, Harry HX Wang, Stewart W Mercer, Colin Sanderson, Martin McKee, Christopher Troeger, Kanyin I Ong, et al. 2020. How many are at increased risk of severe COVID-19 disease? Rapid global, regional and national estimates for 2020. *medRxiv* (2020).
[12] Ermengol Coma, Nuria Mora, Albert Prats-Uribe, Francesc Fina, Daniel Prieto-Alhambra, and Manuel Medina-Peralta. 2020. Excess cases of influenza suggest an earlier start to the coronavirus epidemic in Spain than official figures tell us: an analysis of primary care electronic medical records from over 6 million people from Catalonia. *medRxiv* (2020).
[13] Bo Diao, Kun Wen, Jian Chen, Yueping Liu, Zilin Yuan, Chao Han, Jiahui Chen, Yuxian Pan, Li Chen, Yunjie Dan, et al. 2020. Diagnosis of Acute Respiratory Syndrome Coronavirus 2 Infection by Detection of Nucleocapsid Protein. *medRxiv* (2020).
[14] Hatem A Elshabrawy. 2020. SARS-CoV-2: An Update on Potential Antivirals in Light of SARS-CoV Antiviral Drug Discoveries. *Vaccines* 8, 2 (2020), 335.
[15] Caroline X Gao, Yuguo Li, Jianjian Wei, Sue Cotton, Matthew Hamilton, Lei Wang, and Benjamin J Cowling. 2020. Multi-route respiratory infection: when a transmission route may dominate. *medRxiv* (2020).
[16] Ya Gao, Ming Liu, Shuzhen Shi, Yamin Chen, Yue Sun, Ji Chen, and Jinhui Tian. 2020. Cancer is associated with the severity and mortality of patients with COVID-19: a systematic review and meta-analysis. *medRxiv* (2020).
[17] Hui Poh Goh, Wafiah Ilyani Mahari, Norhadyrah Izazie Ahad, Liling Chaw, Nurolaini Kifli, Bey Hing Goh, Siang Fei Yeoh, and Long Chiau Ming. 2020. Risk factors affecting COVID-19 case fatality rate: A quantitative analysis of top 50 affected countries. *medRxiv* (2020).
[18] Emanuel Goldman. 2020. Exaggerated risk of transmission of COVID-19 by fomites. *The Lancet Infectious Diseases* 20, 8 (2020), 892–893.
[19] Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 50–57.
[20] Katharine E Hubbard and Sonja D Dunbar. 2017. Perceptions of scientific research literature and strategies for reading papers depend on academic career stage. *PloS one* 12, 12 (2017).
[21] Lindsay Kim, Shikha Garg, Alissa O'Halloran, Michael Whitaker, Huong Pham, Evan J Anderson, Isaac Armistead, Nancy M Bennett, Laurie Billing, Kathryn Como-Sabetti, et al. 2020. Risk factors for intensive care unit admission and in-hospital mortality among hospitalized adults identified through the US coronavirus disease 2019 (COVID-19)-associated hospitalization surveillance network (COVID-NET). *Clinical Infectious Diseases* (2020).
[22] Stephen M Kissler, Christine Tedijanto, Edward Goldstein, Yonatan H Grad, and Marc Lipsitch. 2020. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* 368, 6493 (2020), 860–868.
[23] Samuli Laato, AKM Islam, Muhammad Nazrul Islam, and Eoin Whelan. 2020. Why do people share misinformation during the Covid-19 pandemic? *arXiv preprint arXiv:2004.09600* (2020).
[24] Paul-Henri Lambert, Donna M Ambrosino, Svein R Andersen, Ralph S Baric, Steven B Black, Robert T Chen, Cornelia L Dekker, Arnaud M Didierlaurent, Barney S Graham, Samantha D Martin, et al. 2020. Consensus Summary Report for CEPI/BC March 12-13, 2020 Meeting: Assessment of Risk of Disease Enhancement with COVID-19 Vaccines. *Vaccine* (2020).
[25] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.
[26] Matthew Levin, Martin D Chen, Anjan Shah, Ronak Shah, George Zhou, Erica Kane, Garrett Burnett, Shams Ranginwala, Jonathan Madek, Christopher Gidiscin, et al. 2020. Differential ventilation using flow control valves as a potential bridge to full ventilatory support during the COVID-19 crisis. *medRxiv* (2020).
[27] Xue Li, Wei Xu, Marshall Dozier, Yazhou He, Amir Kirolos, Evropi Theodoratou, et al. 2020. The role of children in transmission of SARS-CoV-2: A rapid review. *Journal of global health* 10, 1 (2020).
[28] Jimmy Lin. 2009. Is searching full text more effective than searching abstracts? *BMC bioinformatics* 10, 1 (2009), 46.
[29] Vanessa Aparecida Marcolino, Tatiana Colombo Pimentel, and Carlos Eduardo Barão. 2020. What to expect from different drugs used in the treatment of COVID-19: A study on applications and in vivo and in vitro results. *European Journal of Pharmacology* 887 (2020), 173467.
[30] Alexandra Martin, Alexandre Storto, Barbara Andre, Allison Mallory, Remi Dangla, Benoit Visseaux, and Olivier Gossner. 2020. High-sensitivity COVID-19 group testing by digital PCR. *arXiv preprint arXiv:2006.02908* (2020).
[31] Eduardo Massad, Marcos Amaku, Annelies Wilder-Smith, Paulo Cesar Costa dos Santos, Claudio Jose Struchiner, and Francisco Antonio Bezerra Coutinho. 2020. Two complementary model-based methods for calculating the risk of international spreading of anovel virus from the outbreak epicentre. The case of COVID-19. *Epidemiology & Infection* (2020), 1–19.
[32] Eugene Merzon, Dmitry Tworowski, Alessandro Gorohovski, Shlomo Vinker, Avivit Golan Cohen, Ilan Green, and Milana Frenkel-Morgenstern. 2020. Low plasma 25 (OH) vitamin D level is associated with increased risk of COVID-19 infection: an Israeli population-based study. *The FEBS journal* 287, 17 (2020), 3693–3702.
[33] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
[34] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669* (2019).
[35] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I Furlong. 2016. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research* (2016), gkw943.
[36] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).
[37] Jennifer M Reckrey. 2020. COVID-19 Confirms It: Paid Caregivers are Essential Members of the Healthcare Team. *Journal of the American Geriatrics Society* (2020).
[38] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.
[39] Stephen E Robertson, Steve Walker, MM Beaulieu, Mike Gatford, and Alison Payne. 1996. Okapi at TREC-4. *Nist Special Publication Sp* (1996), 73–96.
[40] Ioanna Smyrlaki, Martin Ekman, Martin Vondracek, Natali Papanicoloau, Antonio Lentini, Johan Aarum, Shaman Muradrasoli, Jan Albert, Björn Högberg, and Björn Reinius. 2020. Massive and rapid COVID-19 testing is feasible by extraction-free SARS-CoV-2 RT-qPCR. *medRxiv* (2020).
[41] Leonardo Stella, Alejandro Pinel Martínez, Dario Bauso, and Patrizio Colaneri. 2020. The Role of Asymptomatic Individuals in the COVID-19 Pandemic via Complex Networks. *arXiv preprint arXiv:2009.03649* (2020).
[42] Viktor Stojkoski, Zoran Utkovski, Petar Jolakoski, Dragan Tevdovski, and Ljupco Kocarev. 2020. The socio-economic determinants of the coronavirus disease (COVID-19) pandemic. *arXiv preprint arXiv:2004.07947* (2020).
[43] Raphael Tang, Rodrigo Nogueira, Edwin Zhang, Nikhil Gupta, Phuong Cam, Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly Bootstrapping a Question Answering Dataset for COVID-19. *arXiv preprint arXiv:2004.11339* (2020).
[44] Jia Wang and Zhifeng Wang. 2020. Strengths, Weaknesses, Opportunities and Threats (SWOT) Analysis of China's Prevention and Control Strategy for the COVID-19 Epidemic. *International Journal of Environmental Research and Public Health* 17, 7 (2020), 2235.
[45] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020. CORD-19: The Covid-19 Open Research Dataset. *arXiv preprint arXiv:2004.10706* (2020).
[46] Leigh Weston, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Amalie Trewartha, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. *Journal of chemical information and modeling* 59, 9 (2019), 3692–3702.
[47] Zi-Wei Ye, Shuofeng Yuan, Kit-San Yuen, Sin-Yee Fung, Chi-Ping Chan, and Dong-Yan Jin. 2020. Zoonotic origins of human coronaviruses. *International journal of biological sciences* 16, 10 (2020), 1686.
[48] Shu Yuan, S Jiang, Zi-Lin Li, et al. 2020. Do Humidity and Temperature Impact the Spread of the Novel Coronavirus? *Frontiers in Public Health* 8 (2020), 240.
[49] Edwin Zhang, Nikhil Gupta, Rodrigo Nogueira, Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly Deploying a Neural Search Engine for the COVID-19 Open Research Dataset: Preliminary Thoughts and Lessons Learned. *arXiv preprint arXiv:2004.05125* (2020).
[50] Patrick Zimmerman, Stephanie Stroever, Timothy Burton, Karri Hester, Minha Kim, Ryan Fahy, Kimberly Corbitt, Joann Petrini, and Jeffrey Nicastro. 2020. Mortality Associated With Intubation and Mechanical Ventilation in Patients with COVID-19. *medRxiv* (2020).

Table 4: S_Covid answers for COVID-19 related questions

Question	S_Covid Answer
11. Persistence of virus on surfaces of different materials	A report by van Doremalen and colleagues found survival of both SARS-CoV and SARS-CoV-2 of up to 2 days (on surfaces) and 3 days (in aerosols generated in the laboratory), but again with a large inoculum [18].
12. Role of the environment in transmission	The Respiratory diseases are often simply assumed to be transmitted via "close contact"; however, the complex transmission mechanisms often involve with more than one transmission route including direct or indirect contact, large droplet, and airborne routes [15].
13. Approaches to evaluate risk for enhanced disease after vaccination	NHPs could be utilized to evaluate COVID-19 vaccine candidates without adjuvants and guide in the selection of vaccines that elicit desired attributes that could reduce the risk of vaccine-mediated enhanced disease [24].
14. Effectiveness of drugs being developed and tried to treat COVID-19 patients	Remdesivir (GS-5734™) is an antiviral drug developed by Gilead Sciences initially developed to treat Ebola, but experimental tests are also being carried out to treat diseases such as MERS and COVID-19 [29].
15. Outcomes data for COVID-19 after mechanical ventilation adjusted for age	As age increased, so did the proportion of patients who required ICU admission, invasive mechanical ventilation, and vasopressors. Median age was 76 years (IQR, 66-85); 58% (n=244) were male; 71% (n=299) were admitted to the ICU; and 59% (n=246) received invasive mechanical ventilation [21].
16. Guidance on the simple things people can do at home to take care of sick people and manage disease.	The best health advice for older, frail patients was to stay home and health care providers offered televisits and telephonic symptom management to avoid unnecessary emergency department visits[37].
17. Policies and protocols for screening and testing.	In this study we propose a novel group testing protocol using a commercially available RT-dPCR assay and compare empirically the sensitivity of individual identification through RT-PCR with group testing by RT-dPCR for three groups sizes of 8, 16 and 32 samples [30].
18. Efforts to track the evolution of the virus	Mutation and adaptation have driven the co-evolution of coronaviruses (CoVs) and their hosts, including human beings, for thousands of years [47].
19. Effectiveness of case quarantine of exposed individuals	If 90% of cases are asymptomatic or undetected, as could happen in a location making no effort to follow people during their isolation, the efficacy of quarantine would be about 70% [3] .
20. Effectiveness of inter/inner travel restriction	During the peak of the COVID-19 outbreak in Europe, about 3-6% of air passengers were SARS-CoV-2 positive on repatriation flights [31].
21. Effectiveness of school distancing	We fit our model with the data until August 29th, and then simulate what would happen in the event that schools open in mid-September [41].
22. How does temperature and humidity affect the transmission of 2019-nCoV?	Analyzed meteorological data of 30 cities in China and suggested that low temperature, mild diurnal temperatures, and low humidity likely aid the transmission of novel coronavirus disease 2019 (COVID-19) [48].
23. What is the efficacy of novel therapeutics being tested currently?	Remdesivir and favipiravir are the most promising antiviral drugs that have been tested in clinical trials so far [14].
24. Risk factor studies related to impact of diabetes	Conclusion: Older people above 65 years old and diabetic patients are significant risk factors for COVID-19 [17].
25. Risk factor studies related to impact of male gender	The multivariate analyses, age over 50 years, male gender and low-medium socioeconomic status were also positively associated with the risk of COVID-19 infection [32].
26. Risk factor studies related to impact of kidney disease	Kaplan-Meier analysis demonstrated that patients with kidney disease had a significantly higher risk for in-hospital death [9].
27. Risk factor studies related to impact of cancer	Combined with previously published results , we can conclude that patients with cancer have an increased risk of COVID-19[16].
28. Risk factor studies related to impact of overweight	Our findings are consistent with other reports from New York that did not find obesity to be an independent risk factor for mortality, though reports from reviews suggest obesity does play a role in mortality [50].
29. What do we know about viral shedding in stool?	In addition, we identified six studies presenting indirect evidence on the potential for SARS-CoV-2 transmission by children, three of which found prolonged virus shedding in stools [27].
30. What is the longest duration of viral shedding?	This happens to coincide with the fact that the viral RNA shedding of children is much longer than that of adults(13, 14) [8].