Title
COVIDScholar: An automated COVID-19 research aggregation and analysis platform

Authors
Trewartha, Amalie
Dagdelen, John
Huo, Haoyan
et al.

Publication Date
2020-12-07

Peer reviewed

COVIDScholar: An automated COVID-19 research aggregation and analysis platform

Amalie Trewartha[1], John Dagdelen[1,2], Haoyan Huo[1,2], Kevin Cruse[1,2], Zheren Wang[1,2], Tanjin He[1,2], Akshay Subramanian[3], Yuxing Fei[4], Benjamin Justus[2], Kristin Persson[1,2], and Gerbrand Ceder[1,2]

[1] Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[2] Department of Materials Science & Engineering, University of California, Berkeley, Berkeley, CA 94720, USA

[3] Indian Institute of Technology Roorkee, Roorkee, Uttarakhand 247667, India

[4] Wuhan University, Wuhan, Hubei 430072, China

Abstract. The ongoing COVID-19 pandemic has had far-reaching effects throughout society, and science is no exception. The scale, speed, and breadth of the scientific community's COVID-19 response has lead to the emergence of new research literature on a remarkable scale — as of October 2020, over 81,000 COVID-19 related scientific papers have been released, at a rate of over 250 per day. This has created a challenge to traditional methods of engagement with the research literature; the volume of new research is far beyond the ability of any human to read, and the urgency of response has lead to an increasingly prominent role for pre-print servers and a diffusion of relevant research across sources. These factors have created a need for new tools to change the way scientific literature is disseminated.

COVIDScholar is a knowledge portal designed with the unique needs of the COVID-19 research community in mind, utilizing NLP to aid researchers in synthesizing the information spread across thousands of emergent research articles, patents, and clinical trials into actionable insights and new knowledge. The search interface for this corpus, https://covidscholar.org, now serves over 2000 unique users weekly.

We present also an analysis of trends in COVID-19 research over the course of 2020.

1 Introduction

The scientific community has responded to the COVID-19 pandemic with unprecedented speed, and as a result an enormous amount of research literature is rapidly emerging, at a rate of over 250 papers a day [1]. The urgency and volume of emerging research has caused pre-prints to take a prominent role in lieu of traditional journals, leading to widespread usage of pre-print servers for the first time in many fields, most prominently biomedical sciences[2][3]. While this allows new research to be disseminated to the community sooner, this also circumvents the role of journals in filtering poor or flawed papers and highlighting relevant research [4]. Additionally, the uniquely multi-disciplinary nature of the scientific community's response to the pandemic has lead to pertinent research being dispersed across many open access and pre-print services - no single one of which captures the entirety of the COVID-19 literature.

These challenges have created a need and opportunity for new tools and methods to rethink the way in which researchers engage the wealth of available COVID-19 scientific literature.

COVIDScholar is an effort to address these issues by using natural language processing (NLP) techniques to aggregate, analyze, and search the COVID-19 research literature. We have developed an automated, scalable infrastructure for scraping and integrating new research as it appears, and used it to construct a targeted corpus of over 81,000 scientific papers and documents pertinent to

COVID-19 from a broad range of disciplines. The search interface for this corpus, https://covidscholar.org, now serves over 2000 unique users weekly.

While a variety of other COVID-19 literature aggregation efforts exist [5, 6, 7], COVIDScholar differs in the breadth of literature collected. In addition to the biological and medical research collected by other large-scale aggregation efforts such as CORD-19 [6] and LitCOVID [7], COVIDScholar's collection includes the full breadth of COVID-19 research, including public health, behavioural science, physical sciences, economics, psychology, and humanities.

In this paper, we present a description of the COVIDScholar data intake pipeline and back-end infrastructure, and the NLP models used to power directed searches on the front-end search portal. We also present an analysis of the COVIDScholar corpus, and discuss trends in the dynamics of research output during the pandemic.

2 Data Pipeline & Infrastructure

At the heart of COVIDScholar is the automated data intake and processing pipeline, depicted in Fig. 1. Data sources are continually checked for new or updated papers, patents, and clinical trials, which are then parsed, cleaned, analyzed with NLP models, and made searchable on https://covidscholar.org. [5].

The COVIDScholar research corpus consists of research literature from 14 different open-access and pre-print services, listed in Table. 1. For each of these, a web scraper regularly checks for new documents and updates to existing ones. Missing metadata is then collected from Crossref, and citation data is collected from OpenCitations [8].

[5] The complete codebase for the data pipeline is available at https://github.com/COVID-19-Text-Mining

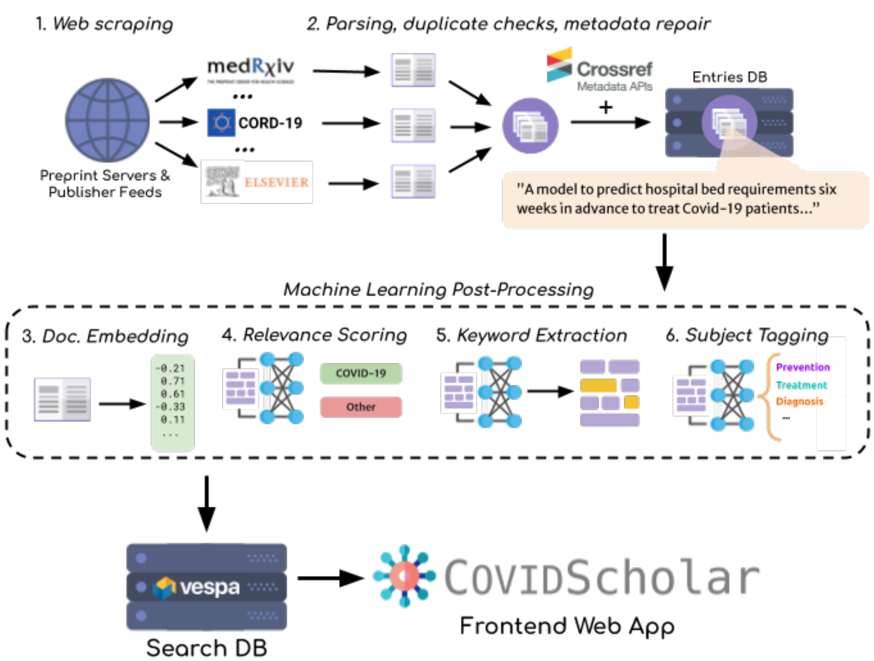

Fig. 1: The data pipeline used to construct the COVIDScholar research corpus.

After collection, these publications are then parsed into a unified format, cleaned, and resolved to remove duplicates. Publications are identified as duplicates when they share any of doi (up to version number), pubmed id, or uncased title. For clinical trials without valid document identifiers, a shared title is used to identify duplicates. In cases where there are multiple versions of a single paper (most commonly, a pre-print and a published version), a combined single document is produced, whose contents are selected on a field-by-field basis using a priority system. Published versions and higher version numbers (based on doi) are given higher priority, and sources are otherwise prioritized based on the quality of their text.

Source	**COVID-19 Publications Count**
preprints.org [9]	923
osf.io [10]	337
lens.org [11]	98
SSRN [12]	3491
Psyarxiv [13]	691
CORD-19 [6]	1135
Dimensions.ai [14]	6489
Elsevier [15]	6735
Chemrxiv [16]	292
LitCovid [17]	51807
Biorxiv [18]/Medrxiv [19]	8832
NBER.org [20]	261
COVIDScholar User Submission	25

Table 1: The source of papers, patents, and clinical trials in the COVIDScholar collection, with the count of COVID-19 related publications from each source.

In cases where full-text PDFs are available text is parsed from the document using pdfminer (for PDFs with embedded text [21]) or OCR. However, it is our experience that, in general text extracted in this manner is not of sufficient quality for to be used by the classification and relevance NLP models, and at this time is used solely for text searches.

Abstracts are classified based on their relevance to COVID-19, topic, discipline, and field. Publications are classified into 5 disciplines - Biological & Chemical Sciences, Medical Sciences, Public Health, Physical Sciences and Humanities & Social Sciences. A paper may belong to any number of disciplines. Each discipline is composed of 12-15 fields. The breakdown of fields by discipline is shown in the supplementary material (S.1). Publications for which an abstract cannot be found are not classified.

Keywords are also extracted from titles and abstracts using an unsupervised approach, as described in Sec. 3.

Our web portal, COVIDScholar.org, provides an accessible user interface to a variety of literature search tools and information retrieval algorithms tuned specifically for the needs of COVID-19 researchers. Because there still remains a great deal that we do not know about the disease, we have directed our efforts towards developing tools that can extend beyond information retrieval and aid researchers at the knowledge discovery phase as well. To do this, we have utilized new machine learning and natural language processing techniques together with proven information retrieval approaches to create the search algorithms behind COVIDScholar, which we describe in the remainder of this section.

Machine learning algorithms can be used to identify emerging trends in the literature and correlate them with similar patterns from pre-existing research. For this reason, we chose to base our search back end on the Vespa engine [22], which provides a high level of performance, wide scalability, and easy integration with custom machine learning models. For example, the default search result ranking profile on COVIDScholar.org combines the BM25 relevance[**BM25**] with a "COVID-19 relevance" score calculated by a classification model trained to predict whether a paper discusses the SARS-CoV-2 virus or COVID-19 using this approach. We observe that papers from before the COVID-19 pandemic that are related to certain viruses/diseases tend to receive high relevance scores, especially papers on the original SARS and other respiratory diseases. SARS-CoV-2 shares 79% of its genome sequence identity with the SARS-CoV virus[23], and there are many similarities between how the two viruses enter cells, replicate, and transmit between hosts.[24] Because the relevance classification model gives a higher score to studies on these similar diseases, search results are more likely to contain relevant information, even if it is not directly focused on COVID-19. For example, the transmembrane protease TMPRSS2 plays an important role in viral entry and spread for both SARS-CoV and SARS-CoV-2, and its inhibition is a promising avenue for treating COVID-19[25]. A wealth of information on strategies to inhibit TMPRSS2 activity and their efficacy in blocking SARS-CoV from entering host cells was available in the early days of the COVID-19 pandemic. These studies were boosted in search results because of their higher relevance scores, thereby bringing potentially useful information to the attention of researchers more directly. In comparison, results of a Google Scholar search for "TMPRSS2" (with results containing "COVID-19" and "SARS-CoV-2" filtered out) are dominated by studies on the protease's role in various cancers.

COVIDScholar also provides tools that utilizes unsupervised document embeddings so that searches can be performed within "related documents" to automatically link research papers together by topics, methods, drugs, and other key pieces of information. Documents are sorted by similarity via the cosine distances between unsupervised document embeddings[26], which is then combined with the more overall result-ranking score mentioned above. This allows users to focus their results into a more specific domain without having to repeatedly pick and choose new search terms to add to their queries. Users can also filter all of the documents in the database by broader subjects relevant to COVID-19 (treatment, transmission, case reports, etc), which are all determined though the application of machine learning models trained on a smaller number of hand-labeled examples. All combined, these tools have allowed us to create much more targeted tools for literature search and knowledge discovery that would not be possible otherwise.

3 Text Analysis NLP Models

Classification of abstracts is performed using a fine-tuned SciBERT [27] model. While other BERT models pre-trained on scientific text exist (e.g. BioBERT [28], MedBERT [29], and ClinicalBERT [30]), we select SciBERT due to its broad, multidisciplinary training corpus, which we expect to more closely resemble the COVIDScholar corpus than those pre-trained on a single discipline. SciBERT has state-of-the-art performance on the task of paper domain classification [31], as well as a number of biomedical domain benchmarks [32, 33, 34] - the most common discipline in the COVIDScholar corpus. A single fully-connected layer with sigmoid activation is used as a classification head, and the model is fine-tuned for 4 epochs using 2600 human-annotated abstracts [6]

ROC curves for the classifier's performance for each top-level discipline using 20-fold cross-validation are shown in Fig. 2. The classifier performs extremely well, with F1 scores above 0.73 for all disciplines. Performance metrics of the discipline classifier are displayed in Table. 2, compared to a baseline random forest model using TF-IDF features.

On three disciplines (Medical Sciences, Physical Sciences, and Humanities & Social Sciences) the SciBERT-based discipline classifier offers a significant performance advantage over the baseline random forest/TF-IDF model, with F1

[6] Abstracts were annotated by members of the Rapid Reviews: COVID-19 [35] editorial team.

		Biological & Chemical Sciences	Medical Sciences	Public Health	Physical Sciences	Humanities & Social Sciences
SciBERT	F1	0.92	0.85	0.73	0.78	0.92
	Precision	0.92	0.80	0.74	0.78	0.88
	Recall	0.92	0.80	0.75	0.81	0.92
	Accuracy	0.92	0.85	0.73	0.79	0.92
Random Forest	F1	0.90	0.63	0.73	0.68	0.78
	Precision	0.93	0.77	0.83	0.81	0.89
	Recall	0.89	0.55	0.67	0.59	0.73
	Accuracy	0.92	0.84	0.81	0.83	0.90

Table 2: Scoring metrics of SciBERT [27] and baseline random forest discipline classification models. Models were evaluated using 10-fold cross-validation on 2600 labeled abstracts. Input features to the random forest model generated using TF-IDF.

scores which are between 0.1 and 0.14 higher. These are the broadest disciplines, encompassing multiple disparate fields. The large variability of subjects within these domains may account for the inability of TF-IDF-based models to classify them well.

For the remaining two disciplines, Biological & Chemical Sciences and Public Health, the F1 scores are similar between SciBERT and the baseline model. In the case of Biological & Chemical Sciences, this may be explained by relatively distinctive vocabulary and narrow subjects within the discipline. Public Health was observed to have the largest inter-annotator disagreement, leading to a lower performance by the classifier.

It is also of note in each case that while precision is broadly similar between the two models, the baseline model exhibits significantly lower recall. This may be due to unbalanced training data - no single discipline accounts for more than 33% of the total corpus. For search applications, often a relatively small number of documents is relevant to each query. In this case, a high recall is more desirable than a high precision - in practice, the performance gap between the two models is larger than indicated by relative F1 scores.

On the task of binary classification as related to COVID-19, our current models perform similarly well, achieving an F1 score of 0.98. While the binary classification task is significantly simpler from an NLP perspective - the majority of related papers contain "COVID-19" or some synonym - this still represents a significant performance improvement over the baseline model, which achieves an F1-score of 0.90. Given the relative simplicity of this task, in cases where an abstract is absent we classify it as related to COVID-19 based on the title.

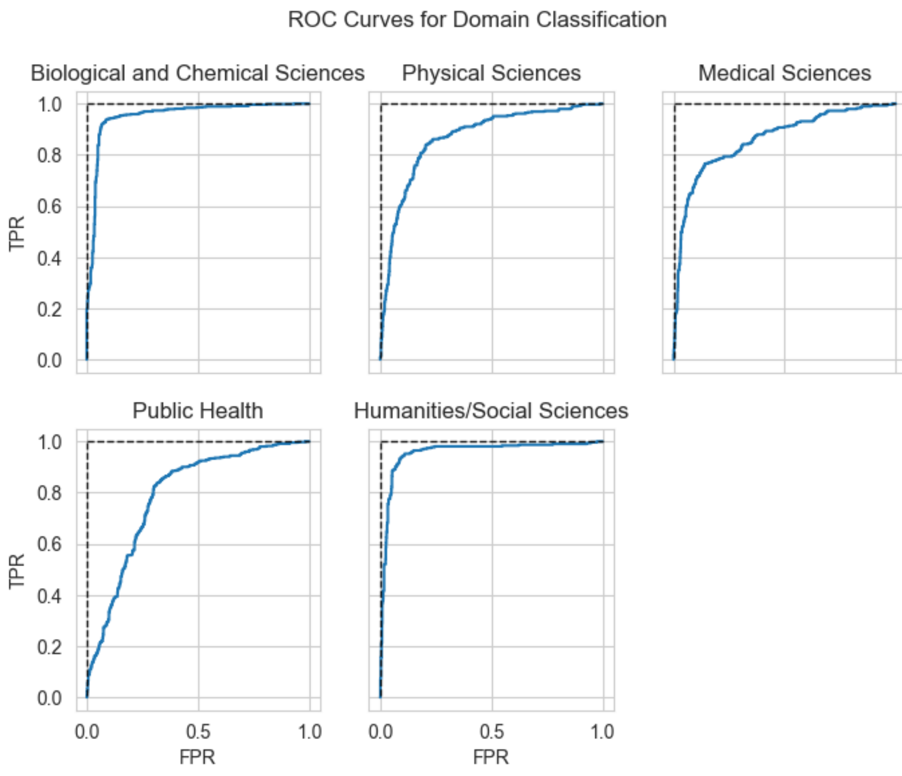

Fig. 2: ROC curves for discipline classification models of paper abstracts using a fine-tuned SciBERT [27] model adapted for classification. Training is performed using a set of 2500 human-annotated abstracts, and results shown are generated with 20-fold cross validation.

For the task of unsupervised keyword extraction, 63 abstracts were annotated by humans, and two statistical methods, TextRank [36]and TF-IDF [37], and two graph-based models, RaKUn [38] and Yake [39], were tested. Models were evaluated for overlap between human-annotated keywords and extracted keywords, and results are shown in Table. 3. Note that due to the inherent subjectivity of the keyword extraction task that scores are relatively low - the best performing model, RaKUn has an F1 score of only 0.2. However, the quality of extracted keywords from this model was deemed reasonable for display on the search portal after manual inspection.

Model	Precision	Recall	F1
RaKUn	0.17	0.33	0.2
Yake	0.11	0.45	0.15
TextRank	0.06	0.36	0.09
TF-IDF	0.10	0.09	0.08

Table 3: Precision, recall, and F1 scores for 4 unsupervised keywords extractors, RaKUn[38], Yake[39], TextRank[36], and TF-IDF[37]. Output from keyword extractors was compared to 63 abstracts with human-annotated keywords.

To better visualize the embedding of COVID-19-related phrases and find latent relationship between biomedical terms, we designed a tool based on Embedding Projector[40]. A screenshot of the tool is shown in Fig. 3

We utilize FastText[41] embeddings for the embedding projector, with an embedding dimension of 100. Embeddings are trained on the abstracts of all papers which have been classified as relevant to COVID-19.

For the purpose of visualization, embeddings must be projected to a lower dimensional space (2D or 3D). The dimensionality reduction technique used here includes principal component analysis (PCA), uniform manifold approximation and projection (UMAP) and t-distributed stochastic neighbor embedding (t-SNE). Users can set various parameters and do the dimension reduction via an

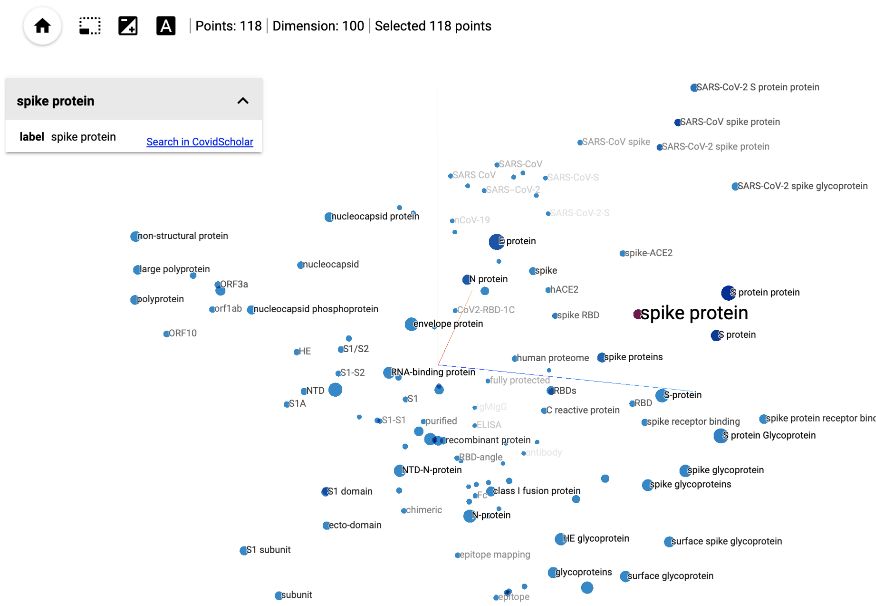

Fig. 3: A screenshot of the embedding projector visualizing tokens similar to "spike protein", using FastText[41] embeddings trained on the COVIDScholar corpus.

interactive page. They can also load and visualize the cached result on the server with default parameters.

Cosine distance is used to measure the similarity between phrases. If the cosine distance between two phrases is quite small, they are likely to have similar meaning.

$$\text{Cosine Distance}(p_1, p_2) = \frac{\boldsymbol{Emb}(p_1) \cdot \boldsymbol{Emb}(p_2)}{\|\boldsymbol{Emb}(p_1)\| \|\boldsymbol{Emb}(p_2)\|}$$

p_1, p_2 represent two phrases, $\boldsymbol{Emb}$ maps phrases to their embedded representation in the learned semantic space.

4 COVIDScholar Corpus Analysis

4.1 Corpus Breakdown

As of October 2020, the COVIDScholar corpus consists of 150,113 total documents, of which 143,887 are papers. The remainder is composed of 3306 patents, 1712 clinical trials, 1025 book chapters, and 183 datasets. Of the papers, 81,106 are classified as related to COVID-19,[7] and are approximately equally split between preprints and published papers - 44% pre-prints, 56% published. A breakdown by discipline of the COVID-19 relevant papers is shown in Table. 4. As may be expected, Public Health and Biological & Chemical Sciences are the most represented disciplines, with respectively 56% and 42% of the corpus tagged as members of these disciplines. Overlap between these two disciplines is relatively small —only 3295 papers are classified as belonging to both Public Health and Biological & Chemical Sciences—, and so the vast majority of the corpus, 50,787 papers, belongs to one of the two.

Discipline	Paper Count	Fraction of Total
Biological & Chemical Sciences	23227	0.42
Humanities & Social Sciences	17464	0.31
Medical Sciences	21023	0.38
Physical Sciences	17214	0.31
Public Health	30855	0.56

Table 4: The number of papers and fraction of total COVID-19 related papers in the COVIDScholar corpus for each discipline. Only papers with abstracts are classified and included in final count. Note that a given paper may have any number of discipline labels.

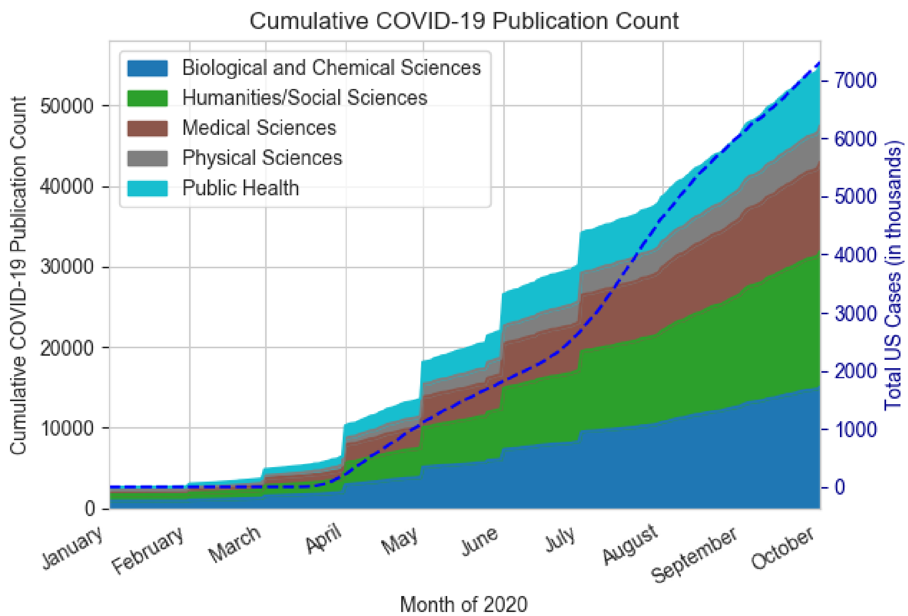

Fig. 4: Cumulative count by primary discipline of COVID-19 papers in the COVIDScholar database, and total number of reported US COVID-19 cases during the first 10 months of 2020. Papers are categorized by the classification model described in Sec. 3, and assigned to the discipline with highest predicted likelihood. Case data from The New York Times, based on reports from state and local health agencies. Note that only those papers with abstracts available are classified, and so the publication is somewhat lower than the total from Sec. 4.1.

4.2 Research Trends

The cumulative count of COVID-19 papers in the COVIDScholar collection over the first 10 months of 2020 is shown in Fig. 4. Papers are categorized by the discipline with highest predicted likelihood using the fine-tuned SciBERT model described in Sec. 3. Note that papers for which the day of publication is unknown are assigned to the first of the month, causing the step-like features visible at the beginning of each month. The total number of reported US COVID-19 cases is also plotted. Data on cases is from The New York Times, based on reports from state and local health agencies (https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html).

The rate at which publications emerged in all disciplines shows a steep increase through the early months of 2020. Between the declaration of a Public Health Emergency of International Concern[42] by the World Health Organization in January 2020 and April 2020, the rate of new publications approximately tripled each month, from just 91 papers in January to 7135 in April. From May onwards, the rate stabilized at approximately 8000 papers a month.

Given the lag between research and publication, it therefore seems that by April 2020 the COVID-19 research effort had already reached full capacity, before the US case count began to dramatically rise in the Summer. The US government passed two stimulus bills, each with over $1 billion in funding allocated for coronavirus research on March 5th [43] and March 27th [44]. The data suggests that any increase in rate of research associated with these had already fully manifested itself within 2 months of their passing, demonstrating the rapidity of the scientific community's COVID-19 response. Other notable events within this timeframe include the declaration of global pandemic by the WHO on March 11 [45].

[7] Papers marked not relevant to COVID-19 are a combination of papers on related diseases, such as SARS and MERS, and with no relation to COVID-19.

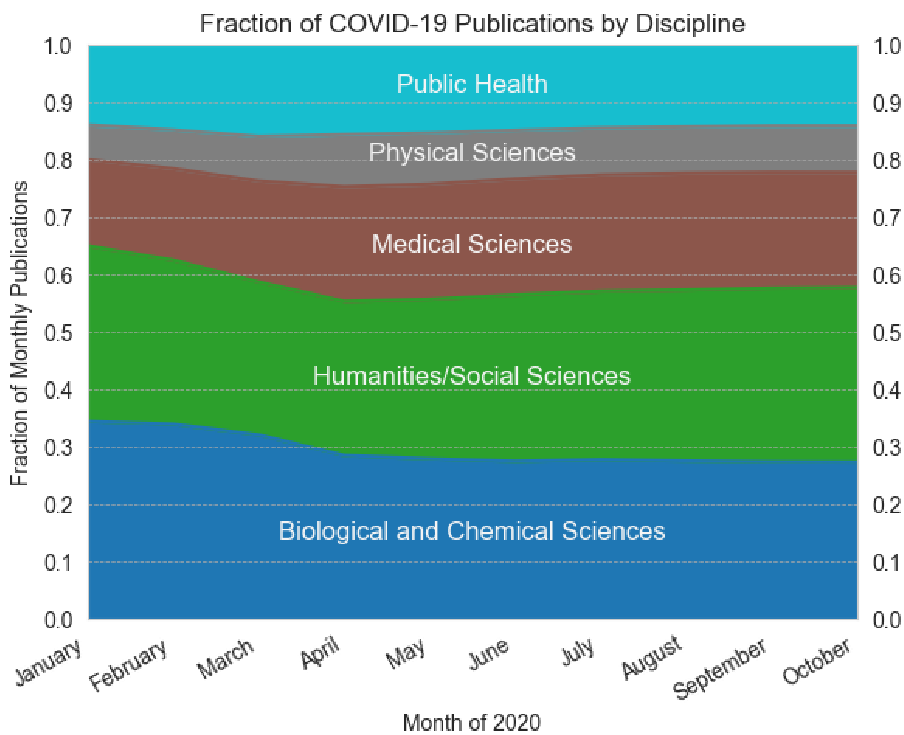

Fig. 5: Fraction of total COVID-19 papers by primary discipline. Fractions are calculated based on total over previous calendar month. Papers are categorized by the classification model described in Sec. 3, and assigned to the discipline with highest predicted likelihood.

A breakdown of research by discipline over the course of 2020 is shown in Fig. 5, which depicts the fraction of monthly COVID-19 publications primarily associated with each discipline. From January - April, the relative popularity of discipline showed some shifts. While Biological and Chemical Sciences comprised 45% of the total corpus in January, by April that had decreased to 28%. This is largely accounted for by an increase in papers from Physical and Medical Sciences - over the same period the fraction of papers from Medical Sciences increased from 15% to 20% of the total, and Physical Sciences from 5% to 8%. By April, the fraction of the corpus from each discipline seems to have stabilized, with fluctuations of relative fractions of under 1%. This is further support for the evidence in Fig. 4 that research output had already reached its maximum rate by April/May - this seems to hold true on a discipline-by-discipline basis also.

We investigate this increase in Fig. 6, where we have plotted the fraction of total monthly papers on selected mental health- and lockdown- related topics. Over the April-June period, there is a clear increase in research related to psychological impacts of lockdown and social distancing, accounting for 6-8% of total monthly papers. Between March and April, many countries and territories instituted lockdown orders, and by April, over half of the world's population was under either compulsory or recommended shelter-in-place orders [46]. The corresponding emergence of a robust literature on psychological impacts associated with this is the major driving force behind the increase in COVID-19 literature from Humanities & Social Sciences.

5 Summary and Future Work

We have developed and implemented a scalable research aggregation, analysis, and dissemination infrastructure, and created a targeted corpus of over 81,000

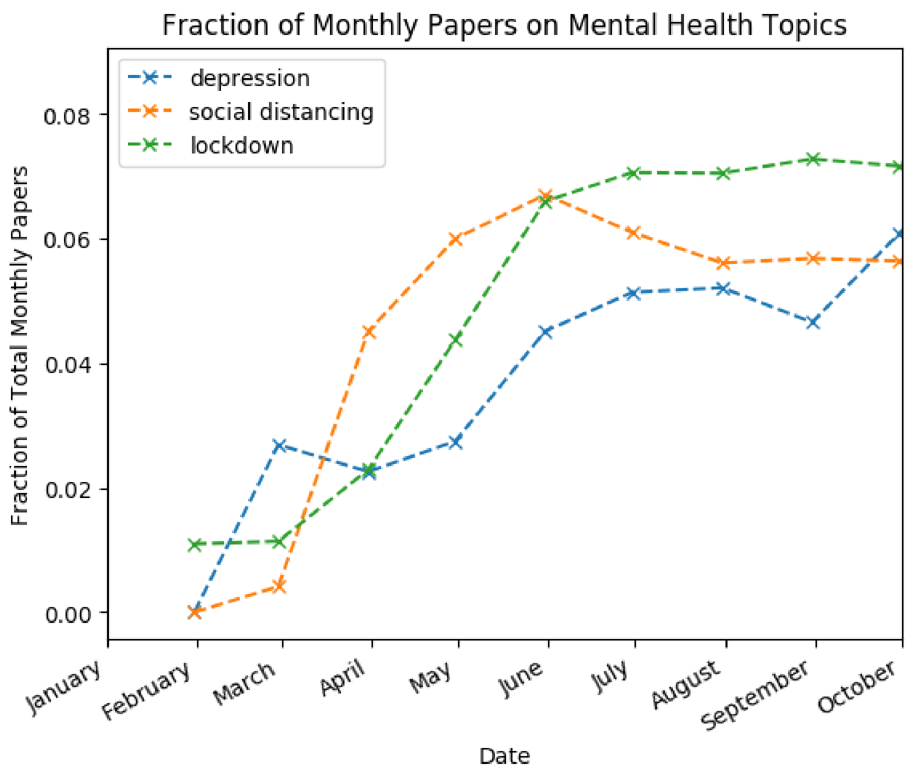

Fig. 6: Fraction of COVID-19 literature on mental health- and lockdown- related topics on a monthly basis.

COVID-19 relevant research documents. The associated search portal, https://covidscholar.org, serves over 2000 weekly scientific users.

While the large amount of open data and enormous scientific interest in COVID-19 have made it an ideal use-case, the infrastructure is domain-agnostic, and presents a blueprint for future large-scale scientific literature aggregation efforts.

While to-date the COVIDScholar research corpus has primarily been used for front-end user search, it provides a rich opportunity for NLP analysis. Recent work [47] has highlighted the ability of NLP to discover latent knowledge from unstructured scientific text, utilizing information from thousands of research papers. We are now moving to employ similar techniques here, applied to such problems as drug re-purposing and predicting protein-protein interactions.

6 Acknowledgements

Portions of this work were supported by the C3.ai Digital Transformation Institute and the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under U.S. Department of Energy Contract No. DE-AC02-05CH11231.

The text corpus analysis and development of machine learning algorithms were supported by the DOE Office of Science through the National Virtual Biotechnology Laboratory, a consortium of DOE national laboratories focused on response to COVID-19, with funding provided by the Coronavirus CARES Act.

This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

We are thankful to the editorial team of Rapid Reviews: COVID-19 for their assistance in annotating text.

References

[1] URL: https://covidscholar.org/stats.

[2] Michael A. Johansson et al. "Preprints: An underutilized mechanism to accelerate outbreak science". In: *PLOS Medicine* 15.4 (Apr. 2018), pp. 1–5. DOI: 10.1371/journal.pmed.1002549. URL: https://doi.org/10.1371/journal.pmed.1002549.

[3] Nicholas Fraser et al. "Preprinting the COVID-19 pandemic". In: *bioRxiv* (2020). DOI: 10.1101/2020.05.22.111294. eprint: https://www.biorxiv.org/content/early/2020/09/18/2020.05.22.111294.full.pdf. URL: https://www.biorxiv.org/content/early/2020/09/18/2020.05.22.111294.

[4] Areeb Mian and Shujhat Khan. "Coronavirus: The spread of misinformation". English. In: *BMC Medicine* 18.1 (Jan. 2020). DOI: 10.1186/s12916-020-01556-3.

[5] *WHO COVID-19 Database*. URL: https://search.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/.

[6] Lucy Lu Wang et al. *CORD-19: The COVID-19 Open Research Dataset*. 2020. arXiv: 2004.10706 [cs.DL].

[7] Qingyu Chen, Alexis Allot, and Zhiyong lu. "Keep up with the latest coronavirus research". In: *Nature* 579 (Mar. 2020), pp. 193–193. DOI: 10.1038/d41586-020-00694-1.

[8] S. Peroni and D. Shotton. "OpenCitations, an infrastructure organization for open scholarship". In: *Quantitative Science Studies* 1 (2019), pp. 428–444.

[9] *The Multidisciplinary Preprint Platform*. URL: https://www.preprints.org/.

[10] URL: https://osf.io/.

[11] *The Lens COVID-19 Data Initiative*. URL: https://about.lens.org/covid-19/.

[12] *Social Science Research Network*. URL: https://www.ssrn.com/index.cfm/en/.

[13] Sean Rife. *Introducing PsyArXiv: a preprint service for psychological science*. Oct. 2016. URL: http://blog.psyarxiv.com/2016/09/19/introducing-psyarxiv/.

[14] *Dimensions COVID-19 Dataset*. URL: https://www.dimensions.ai/covid19/.

[15] *Elsevier Novel Coronavirus Information Center*. Nov. 2020. URL: https://www.elsevier.com/connect/coronavirus-information-center.

[16] *Chemrxiv*. URL: https://chemrxiv.org/.

[17] Qingyu Chen, Alexis Allot, and Zhiyong Lu. "Keep up with the latest coronavirus research". In: *Nature* 579.7798 (2020), pp. 193–193. DOI: 10.1038/d41586-020-00694-1.

[18] 2013 Jocelyn KaiserNov. 12 et al. *New Preprint Server Aims to Be Biologists' Answer to Physicists' arXiv*. Dec. 2017. URL: https://www.sciencemag.org/news/2013/11/new-preprint-server-aims-be-biologists-answer-physicists-arxiv.

[19] Claire Rawlinson and Theodora Bloom. *New preprint server for medical research*. 2019.

[20] *NBER Working Papers*. URL: https://www.nber.org/papers.

[21] URL: https://github.com/pdfminer/pdfminer.six.

[22] URL: https://vespa.ai/.

[23] Roujian Lu et al. "Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding". In: *The Lancet* 395.10224 (2020), pp. 565–574. ISSN: 1474547X. DOI: 10.1016/S0140-6736(20)30251-8. URL: http://dx.doi.org/10.1016/S0140-6736(20)30251-8.

[24] Ali A. Rabaan et al. "SARS-CoV-2, SARS-CoV, and MERS-CoV: A comparative overview". In: *Infezioni in Medicina* 28.2 (2020), pp. 174–184. ISSN: 11249390.

[25] Konrad H Stopsack et al. "TMPRSS2 and COVID-19: Serendipity or Opportunity for Intervention?" In: *Cancer discovery* 10.6 (2020), pp. 779–782.

[26] Quoc Le and Tomas Mikolov. "Distributed Representations of Sentences and Documents". In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. ICML'14. Beijing, China: JMLR.org, 2014, II–1188–II–1196.

[27] Iz Beltagy, Kyle Lo, and Arman Cohan. "SciBERT: Pretrained Language Model for Scientific Text". In: *EMNLP*. 2019. eprint: arXiv:1903.10676.

[28] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* (Sept. 2019). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz682. URL: https://doi.org/10.1093/bioinformatics/btz682.

[29] Laila Rasmy et al. *Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction.* 2020. arXiv: 2005.12833 [cs.CL].

[30] Emily Alsentzer et al. "Publicly Available Clinical BERT Embeddings". In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop.* Minneapolis, Minnesota, USA: Association for Computational Linguistics,

June 2019, pp. 72–78. DOI: 10.18653/v1/W19-1909. URL: https://www.aclweb.org/anthology/W19-1909.

[31] Arnab Sinha et al. "An Overview of Microsoft Academic Service (MAS) and Applications". In: *WWW - World Wide Web Consortium (W3C)*. May 2015. URL: https://www.microsoft.com/en-us/research/publication/an-overview-of-microsoft-academic-service-mas-and-applications-2/.

[32] Wonjin Yoon et al. "CollaboNet: collaboration of deep neural networks for biomedical named entity recognition". In: *BMC Bioinformatics* 20.S10 (May 2019). ISSN: 1471-2105. DOI: 10.1186/s12859-019-2813-6. URL: http://dx.doi.org/10.1186/s12859-019-2813-6.

[33] Benjamin Nye et al. "A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 197–207. DOI: 10.18653/v1/P18-1019. URL: https://www.aclweb.org/anthology/P18-1019.

[34] Sangrak Lim and Jaewoo Kang. "Chemical–gene relation extraction using recursive neural network". In: *Database* 2018 (June 2018). bay060. ISSN: 1758-0463. DOI: 10.1093/database/bay060. eprint: https://academic.oup.com/database/article-pdf/doi/10.1093/database/bay060/27438554/bay060.pdf. URL: https://doi.org/10.1093/database/bay060.

[35] "Rapid Reviews: COVID-19, publishes reviews of COVID-19 preprints". In: *Rapid Reviews COVID-19* (Aug. 11, 2020). https://rapidreviewscovid19.mitpress.mit.edu/pub/wfavs1 URL: https://rapidreviewscovid19.mitpress.mit.edu/pub/wfavs1oc.

[36] Rada Mihalcea and Paul Tarau. "TextRank: Bringing Order into Text". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Lan-*

guage Processing. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 404–411. URL: https://www.aclweb.org/anthology/W04-3252.

[37] Gerard Salton and Christopher Buckley. "Term-weighting approaches in automatic text retrieval". In: *Information Processing & Management* 24.5 (1988), pp. 513–523. ISSN: 0306-4573. DOI: https://doi.org/10.1016/0306-4573(88)90021-0. URL: http://www.sciencedirect.com/science/article/pii/0306457388900210.

[38] Blaz Skrlj, Andraz Repar, and S. Pollak. "RaKUn: Rank-based Keyword extraction via Unsupervised learning and Meta vertex aggregation". In: *ArXiv* abs/1907.06458 (2019).

[39] Ricardo Campos et al. "YAKE! Collection-Independent Automatic Keyword Extractor". In: Feb. 2018. DOI: 10.1007/978-3-319-76941-7_80.

[40] Daniel Smilkov et al. "Embedding projector: Interactive visualization and interpretation of embeddings". In: *arXiv preprint arXiv:1611.05469* (2016).

[41] Piotr Bojanowski et al. "Enriching Word Vectors with Subword Information". In: *arXiv preprint arXiv:1607.04606* (2016).

[42] URL: https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov).

[43] URL: https://www.congress.gov/bill/116th-congress/house-bill/6074/text.

[44] URL: https://www.congress.gov/116/bills/hr748/BILLS-116hr748eas.pdf.

[45] URL: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

[46] URL: https://www.euronews.com/2020/04/02/coronavirus-in-europe-spain-s-death-toll-hits-10-000-after-record-950-new-deaths-in-24-hou.

[47] Vahe Tshitoyan et al. "Unsupervised word embeddings capture latent knowledge from materials science literature". In: *Nature* 571 (July 2019), pp. 95–98. DOI: 10.1038/s41586-019-1335-8.