Exploring the evolution of research topics during the COVID-19 pandemic

Francesco Invernici[a], Anna Bernasconi[a,*] and Stefano Ceri[a]

[a]Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milan, Italy

ARTICLE INFO

Keywords:
Research Data
Scientific Literature
Natural Language Processing
Topic Modeling
COVID-19
Time Series

ABSTRACT

The COVID-19 pandemic has changed the research agendas of most scientific communities, resulting in an overwhelming production of research articles in a variety of domains, including medicine, virology, epidemiology, economy, psychology, and so on. Several open-access corpora and literature hubs were established; among them, the COVID-19 Open Research Dataset (CORD-19) has systematically gathered scientific contributions for 2.5 years, by collecting and indexing over one million articles—this corpus, however, does not provide an easy-to-access overview of its content. Here, we present the CORD-19 Topic Visualizer (CORToViz), a method and associated visualization tool for inspecting the CORD-19 textual corpus of scientific abstracts. Our method is based upon a careful selection of up-to-date technologies (including large language models), resulting in an architecture for clustering articles along orthogonal dimensions and extraction techniques for temporal topic mining. Topic inspection is supported by an interactive dashboard, providing fast, one-click visualization of topic contents as word clouds and topic trends as time series, equipped with easy-to-drive statistical testing for analyzing the significance of topic emergence along arbitrarily selected time windows. Overall, our pipeline is very fast and its results match our expectations on topic identification (F1-score 0.854). The processes of data preparation and results visualization are completely general and virtually applicable to any corpus of textual documents–thus suited for effective adaptation to other contexts.

1. Introduction

COVID-19 was the first pandemic event in the Internet age. In addition to well-known social and economic implications, the worldwide community had to deal with an information overload (Valika et al., 2020), due to huge knowledge production about the SARS-CoV-2 virus and the associated COVID-19 disease. To support research, several open-access datasets, corpora, and literature hubs have been collected; among them, we mention the datasets from the Novel Coronavirus Information Center (Elsevier, 2023), the iSearch COVID-19 Portfolio (National Institutes of Health, 2023), the Human Coronaviruses Data Initiative (The Lens, 2021), LitCovid (Chen et al., 2021), COVIDScholar (Dagdelen et al., 2023), the COVID-19 Research Database (World Health Organization, 2023b), and the COVID-19 Open Research Dataset (CORD-19) (Wang et al., 2020).

CORD-19, curated by the Allen Institute for AI in collaboration with the White House Office of Science and Technology, Microsoft Research, and Kaggle, had the overall largest impact. Thanks to weekly updates throughout the pandemic, targeted to cover new preprints and publications, it provided a multidisciplinary, accurate, and timely view of the pandemic evolution. The ready-to-use dataset includes curated metadata, abstracts, full-text papers, as well as vectorial representations generated by the SPECTER transformer-based language model (Cohan et al., 2020). At its final release, in June 2022, CORD-19 indexed more than 1 million papers (out of which 370 thousand with full text), extracted from more than 50 thousand journals and authored by more than 2 million researchers.

CORD-19 has enabled many text mining approaches (Wang and Lo, 2021), leading to remarkable results (Wang et al., 2021), building for instance knowledge graphs for research acceleration (Logette et al., 2021; Wise et al., 2020) and drug repurposing (Wang et al., 2021), resource annotation services (Huang et al., 2020; Serna García et al., 2023), claim verification systems (Wadden et al., 2020), and purpose-specific language models (Korn et al., 2021). Since May 5th, 2023, the pandemic is no longer considered a public health emergency by the Worlds Health Organization (United Nations News, 2023); then, we may finally consider it as a concluded phenomenon and therefore analyze its history

*Corresponding author

francesco.invernici@polimi.it (F. Invernici); anna.bernasconi@polimi.it (A. Bernasconi); stefano.ceri@polimi.it (S. Ceri)

ORCID(s): 0009-0002-5423-6978 (F. Invernici); 0000-0001-8016-5750 (A. Bernasconi); 0000-0003-0671-2415 (S. Ceri)

as a whole. In this direction, this paper aims to show how the big literature corpus CORD-19 can be successfully exploited to gather a comprehensive overview of the pandemic, tracing the trends that have characterized its scientific literature narrative.

To this end, we follow an unsupervised statistical approach based on natural language processing, specifically focused on topic modeling (Krause et al., 2006). In the post-Large Language Models era, instead of resorting to classic topic modeling techniques like Latent Dirichlet Analysis (LDA) or Non-negative Matrix Factorization, we have chosen to exploit pre-trained language models (PLMs), providing representations that effectively embed both syntactic and semantic meaning (Shao et al., 2018). PLMs can be used *as is* (i.e., without any retraining) for several tasks such as summarization (Radford et al., 2019), information retrieval (Thakur et al., 2021), and clustering (Reimers and Gurevych, 2019).

Hereby, topic modeling is interpreted as a clustering task (Jayabharathy et al., 2011) over the latent space generated by the PLMs, as opposed to other approaches that build and train end-to-end models for topic modeling, both based on classical methods (Moody, 2016) and on language models (Meng et al., 2022). Along with the suggestions from a known survey on topic modeling (Egger and Yu, 2022), we selected BERTopic (Grootendorst, 2022) to implement our analyses based on topic modeling from document clustering. BERTopic has already been proven a valid topic modeling framework for social sciences (Falkenberg et al., 2022; Ebeling et al., 2022; Šćepanović et al., 2023) since it is very flexible, can be scaled for big data corpora, and can be embedded in an end-to-end data pipeline.

In addition, we consider the textual representations extracted from TF-IDF-based models to be particularly useful and powerful when compared to both other classical methods (LDA) (Chen et al., 2019) and transformer-based methods, such as Top2Vec (Angelov, 2020), since it also has the advantage of becoming a knowledge retrieval proxy to discover topics by their textual representations.

Other works have previously focused on topic analysis for COVID-19-related matters; some analyzed the early stages of the pandemic (Zhang et al., 2021; Tran et al., 2020), others analyzed the broader field of coronaviruses (Pourhatami et al., 2021), focused on topic distribution by country (Berchialla et al., 2021) or on the delineation and impact in scientometric terms of the early CORD-19 (Colavizza et al., 2021).The approach conducted in this study, hereon called CORToViz (CORD-19 Topic Visualizer) is broader, as it applies to the entire pandemic history without choosing a specific field of investigation *a priori*.

As our input, we consider all the English-language abstracts of CORD-19 with high-quality metadata that were published after December 2019. First, we provide a pipeline for ingesting huge data corpora, built upon state-of-the-art technologies (Grootendorst, 2022), and extracting from them highly relevant topics, clustered along orthogonal dimensions. Then, our system enables the discovery, from a given literary corpus, of topics of interest through a keyword-based search interface. For the discovered topics, a word cloud representation is rendered to the user, who can select, based on the insights of the content, a set of topics whose trends should be visualized on the timeline of the pandemic. When any topic presents evident trends for distinct time periods, a statistical test can be run to determine if it is a significant behavior or just a stochastic event. Our approach leverages the existing technology of BERTopic exclusively for topic analysis, as results are further elaborated by binning topic-clustered documents within temporal ranges, then obtaining relative bin-representation frequency, and finally producing interactive statistical testing.

Remarkably, CORToViz enables a fast exploratory analysis of a big data corpus along the time dimension, in a way that was not well supported previously. Additionally, the proposed technology and method are completely general and agnostic to the specific domain; our full-stack process applies to any corpus of medium-sized textual documents, using any topic model of choice, and a time-series visualizer. We foresee that a methodology similar to the one presented here, once deployed on the Web, can support the lightweight analytics of arbitrary domains.

2. Materials and Methods

CORD-19 anatomy

The content of CORD-19 was explored with an in-depth analysis of its metadata and embeddings (Cohan et al., 2020). CORD-19 presents 1,056,660 articles with associated metadata. Fig. 1*A* presents their distribution along their month of publication, showing a rapid increase in scientific production from the pandemic outbreak until the summer of 2020, and then a stable production. Metadata is typically incomplete, due to missing entries in several fields; this can be observed in Fig. 1*B*, where we show eight metadata fields for a representative 20% of the dataset. Several papers come with duplicates – as analyzed in Fig. 1*C*; in such cases, we retained only the most representative paper of duplicate clusters (see details in the *Materials and Methods*). After data filtering, we retained 357,170 distinct abstracts

written in English that presented adequate metadata for our research purposes. We only selected abstracts equipped with publish_time information (see Fig. 1*B*).

Preliminary and fine-grain clustering

Then, we performed a preliminary feasibility assessment of the topic modeling analysis. We aimed to verify if the latent topics' structure of CORD-19 can be modeled as an unsupervised clustering task. First, we selected the optimal value for executing k-Means clustering (see Fig. 1*D*). As a result of this exploration, we obtained a comprehensive view of the corpus, split into five clusters, reported as a scatter plot in Fig. 2*A*. Thanks to the word clouds generated from the most frequent words for each cluster, we were able to identify five macro-topics:

1. *Biology of coronavirus*, associated with words 'protein', 'cell', and 'viral', with 68,278 articles.
2. *Therapy and treatment*, associated with words 'group', 'treatment', and 'patient', with 54,391 articles.
3. *Epidemiology*, associated with words 'cases', 'data', and 'risk', with 99,554 articles.
4. *Psychology*, associated with words 'social', 'mental', 'care', and 'students', with 72,980 articles.
5. *Society*, associated with 'model'/'social' accompanied by 'analysis', 'research', 'public', and 'lockdown', denoting a broad cluster on the impact of the pandemic on society, with 61,967 articles.

In order to best identify and track specific phenomena that characterized the pandemic, however, we needed a much finer grain of the topics' structure. Thus, we ran a technology-rich, state-of-the-art pipeline, detailed next, obtaining a considerably richer topics' structure. In particular, Given the hierarchical nature of the adopted algorithm, we obtained a hierarchy of 354 clusters, each of which defines one topic; finely-grained topics are aggregated as a list of 29 high-level topics, shown on the right of Fig. 2*B*; these topics, while not perfectly overlapping, can be related to the macro-topics of the exploratory analysis. For instance: (1) the high-level cluster on n-glycans at a finer-grain level also includes clusters on the nucleocapsid, lysosomes, and methyladenosine—it relates to the 'biology' macro-topic; (2) the high-level cluster covering ACE2 and angiotensin is expanded by clusters on prothrombic coagulopathies, COVID-19-associated conditions, severe Coronavirus Disease, sepsis, and ventilators—it relates to the 'therapy and treatement' aspects; (3) fine-grain clusters connected to Italy's COVID-19 pandemic, outbreak, government restrictions, and immunization—they relate to the 'epidemiological' macro-topic; (4) clusters on e-learning, caretakers, loneliness, and sleep-related issues relate to the 'psychology' macro-topic; and (5) clusters on traffic, transportation, pollution, and economy are related to the 'society and environment' macro-topic.

Architecture for topic extraction and dynamic modeling

To process the original dataset, extract topics, and prepare their time-series data, we adopted a full-fledged, technology-rich complete pipeline, illustrated in Fig. 3. The pipeline assembles up-to-date technologies and is fully portable, after adaptation, to any organized repository such as CORD-19.

The top subfigure provides a bird-eye-view of the pipeline. The 'Data Preparation' stage is in charge of removing records with null values or incomplete dates, retaining only records referring to articles written in English and published between December 2019 and June 2022. Record deduplication is also performed since – especially during the pandemic – several scientific contributions were exposed through different portals and registered with multiple digital object identifiers. The 'Hyper-parameter Optimization' stage selects the specific large language model instance and optimizes parameters for dimensionality reduction and topic learning. Finally, the 'Fit BERTopic model and transform data' stage includes a) embedding operations (executing a dimensionality reduction and density-based clustering – with optimized parameters); b) textual representation operations; and c) time series mining. The processes in (a) and (b) enable the keyword-based topic search, whereas (c) enables plotting time series of topic data in the visualization tool. The next paragraphs describe the pipeline steps more in detail.

Metadata selection

The pipeline ingests the CORD-19's metadata table, and applies several preprocessing operations, as shown in the 'Data Preparation' stage depicted in Fig. 3; each row of the metadata table (1,056,660 records) corresponds to a distinct document of the CORD-19 collection (e.g., an abstract, an article, and so on). The pipeline filters rows with *missing mandatory information*, such as the title, abstract DOI, and publishing time; in particular, rows with incomplete time information (e.g., with just the publishing year) were also filtered. Next, records were *deduplicated*, thanks to the `cord_uid`, linked with groups of CORD-19 entries describing the same paper. Most CORD-19 deduplication refers to differences in external identifiers, while duplicates generally agree on textual metadata, the focus of this work. Finally,

only documents written in English within the time window between December 2019 and June 2022 were selected, yielding a final collection of 327K entries.

Terminology

In the following, several methods and technologies are derived from different domains (i.e., language models, clustering, vectorization, topic modeling, and search), each using specific terminology, which should be interpreted as follows.

- **Token, word, term.** Scientific abstracts are composed of *words*, whose meaning is dictated by common sense; language models transform words into *tokens*; vectorization methods associate frequencies to each token, and these are used for building word clouds. Finally, search systems denote as *terms* the words used in keyword-based methods (e.g. TF-IDF) and search.

- **Cluster, class, topic**. Once abstracts are represented as points in the space of the embeddings, they form *clusters*, or *classes* of similar elements; they are then called *topics* after being associated with a textual representation understandable by humans (e.g., a word cloud).

Unsupervised topic modeling

Unsupervised Topic Modeling is used to discover and analyze latent topics within a document, without pre-existing labels or supervision. The methods work at best under the assumption that each document represents a single topic, or at least that one topic is preponderant, so as to exclude featuring multiple topics at the same time. We applied unsupervised topic modeling to CORD-19 abstracts; we based this work on BERTopic (Grootendorst, 2022), a topic modeling tool that leverages transformers and clustering models for latent topic identification and a class-based term frequency–inverse document frequency model for textual representation learning. We produced as output the topics' identification, their representations as word clouds, and the temporal distribution of topics over time. In the lack of ground truth, we employed quantitative methods, wherever possible, to assess the quality of our topic model through its sub-models, as shown in the following two sections.

Preliminary feasibility assessment

At first, we tested the feasibility of learning the latent topics' structure of CORD-19 as an unsupervised clustering task, by applying the K-means clustering algorithm to embeddings of the dataset texts. Embeddings were generated by SPECTER (Cohan et al., 2020), a generic multi-task document-level transformer model. To select the optimal number k of clusters, we maximized (in the 2-60 range) the silhouette score (Shahapure and Nicholas, 2020) over a random sample that covers 25% of the dataset, as shown in Fig. 1*D*. In this way, we identified five macro-clusters. To assess the content of these macro-topics, we computed the 100 most frequent words present in the abstracts (after stopword removal), by using `gensim` (Řehůřek and Sojka, 2010), a Python package for topic modeling. Then, we generated a word cloud for each cluster; words are displayed in different sizes based on word frequency. In this way, we determined that clusters cover different areas of research, as reported in Fig. 2A, generated by projecting the embeddings on the first two principal components (James et al., 2021).

Learning the best latent representation

In order to learn the latent topic structure of the dataset, we mapped each abstract from CORD-19 to a point in an embedding representation, consisting of a dense, 768-dimensional, vector space. Vectorial representations were generated by a transformer model, selected among the sentence transformer models (SBERT) (Reimers and Gurevych, 2019) compatible with BERTopic. Specifically, we used from the `huggingface` repository the model `pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb` (Deka et al., 2022), which provides robust sentence embeddings for clustering and information retrieval tasks for scientific and medical literature.

Thus, we modeled the latent topic learning problem as a clustering task. We adopted HDBSCAN (McInnes et al., 2017) because it is a density-based clustering algorithm. This characteristic helps to learn clusters that are not perfectly shaped as hyperspheres. HDBSCAN is also fairly tunable, to avoid a cluster structure with degenerate characteristics, such as a single massive cluster surrounded by multiple single-item outlier clusters. In order to improve the quality of clustering we compressed the embeddings with UMAP (McInnes et al., 2018), which is a stochastic dimensionality reduction algorithm, before feeding them to HDBSCAN. By tuning UMAP, we found a sweet spot in the trade-off between keeping local structures and prioritizing the representation of the global structure.

Table 1
Summary of parameters' values selected during the 'Hyper-parameter Optimization' stage

Pipeline step	Parameter name	Parameter value
UMAP optimization	n_neighbors	50
	n_components	50
	min_dist	0.0
	metric	`"cosine"`
HDBSCAN optimization	min_cluster_size	100
	min_samples	10
	metric	`"euclidean"`
	cluster_selection_method	`"leaf"`

We evaluated the quality of our topic modeling framework through a quantitative assessment of the learned topics' structure, which is embodied in the clustering results. We assessed the clusters' one-to-one relative density connection using the Density-Based Clustering Validation (DBCV) (Moulavi et al., 2014) index. This index spans from -1 (lowest quality) to 1 (highest quality); we targeted it during the optimization stage, where we selected the highest-scoring hyperparameters for the joined UMAP and HDBSCAN models in the grid search process. To select the hyperparameters of the UMAP and HDBSCAN models, we performed an optimization step, by grid searching the values on the two jointed models targeting the highest DBCV value, which is a score for the goodness of a density-based clustering model spanning [-1, 1]. At every iteration of the grid search, we randomly sampled 25% of the abstracts from the data preparation. In the best run, we obtained a DBCV score of 0.36, using the parameters' values reported in Table 1.

Extraction of textual and visual representations for topics

After we found clusters in the latent space of embeddings, we searched for a synthetic representation of each cluster, to understand the content of the abstracts and finally define topics. Again, we adopted the stack proposed in (Grootendorst, 2022), which is available in BERTopic. As it can be appreciated in Fig. 3, the task consists of two steps: 1) abstract vectorization, and 2) fitting of per-class TF-IDF (Ceri et al., 2013) models. We used the `scikit-learn` (Pedregosa et al., 2011) CountVectorizer, which converts a collection of text documents to a matrix of token counts; we set `stop_words` to "English" and `token_pattern` as a regular expression to keep together hyphenated words, such as COVID-19 and SARS-CoV-2, which are common in biomedical writings. Similarly, we fitted the c-TF-IDF model with the `reduce_frequent_words` parameter set, which considers the square root of the normalized frequency of the terms (i.e., words). With this model, we obtained -for each class- the most relevant terms (i.e., topics) and their frequency. In this way, we computed a textual, human-understandable representation for each cluster, and then retrieved the most important topics using the TF-IDF representations. Finally, we adopted the `word cloud` (Mueller, 2023) package to generate word clouds with the most frequent terms of each topic, thereby providing a visual representation to inspect the content of the topic.

Dynamic topic modeling

To understand the trends of the research topics during the pandemic, we used the dynamic modeling tool of BERTopic. In this tool, the classical definition of topic modeling is extended by including the temporal dimension; to do so, (Grootendorst, 2022) employs the concept of absolute counts for each topic in equally-sized temporal bins, whose size in days has to be defined before the extraction of the time series with such data. In this work, we build on this idea, but we used relative frequencies, by normalizing each absolute count of topic observations with the amount of abstracts published in that period. In this way, we can interpret these values as pointwise measures of the intensities of the topic, as other previous works on dynamic topic modeling (Krause et al., 2006). In practice, we extract the absolute frequency of each topic within time bins of equal size, specifically 1, 2, 3, and 4 weeks; we paired the abstracts with their publication date (`publish_time`). We then pivoted the resulting data frames in order to obtain actual time-series data for each topic. We also normalized each row of the pivoted dataframe, representing a time bin, to obtain the relative frequency of each topic for that time frame. Taking advantage of these time-series, we generated line plots and stacked histograms for the counts of abstracts per bin. To put these visualizations in chronological context, we added as background the plot of the global number of active COVID-19 cases, retrieved from Our World In Data (Mathieu et al., 2020), and a timeline of significant events that marked the evolution of the pandemic, such

as lockdowns, vaccination campaigns, and variants' outbreak, collected from various resources (American Society for Microbiology, 2023; MacMillan Learning, 2022; Wikipedia, 2023).

Statistical evidence for topic dynamics

To check if a topic's trend is statistically significant, we used a procedure based on the non-parametric Kruskal-Wallis test (Kruskal and Wallis, 1952), used for comparing sample medians, checking if two groups are sampled from the same population. The test can be conveniently parametrized by choosing on which topic and time frames (T1 and T2) it should be run; starting and ending times of time frames T1/T2 can be set. The test produces a p-value and H statistics, enabling the acceptance or rejection of the simple null hypothesis, which corresponds to "there is no significant difference in the topic representation in periods T1 versus T2". Specifically, we adopted the implementation of the test available in the Python library `SciPy.stats` (Virtanen et al., 2020), which implements the formulation of the H-test statistic with correction for ties (i.e., two observations among all the groups are equal).

At first, each observation is assigned a rank, starting from 1 for the lowest value. If there are ties, each observation is given the mean of the ranks for which it is tied. Then, the H statistic is computed as

$$H = \frac{\frac{12}{N(N+1)}\left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n^2}\right) - 3(N+1)}{1 - \frac{\sum T}{N^3 - N}} \tag{1}$$

where n_1 and n_2 are the numbers of observations in the two groups, R_1 and R_2 are the sums of the ranks of the observations of the two groups, N is the number of total of observations of the two groups combined and the summation of T, where $T = (t-1)t(t+1)$, is over all the groups of ties. The H test statistic takes positive values and the critical value for the 5% p-value, which we use as the threshold for significance, for two groups, is 3.85. Higher values imply lower p-values and, hence, the rejection of the null hypothesis (Kruskal and Wallis, 1952).

An interactive, discovery-enabling dashboard

All the components described before, such as the c-TF-IDF model, the time-series plots, the word clouds, and the statistical box, are embedded in a single-page, interactive, and responsive dashboard. Users can interactively set the topic and time frames to be tested, therefore impacting the selection of considered abstracts; they can visualize the p-value and H statistics, easily testing their own hypotheses and drawing conclusions.

For implementing the interface, we used `Streamlit` (Streamlit, 2023), a Python package for building single-script web applications. In this way, we enabled multiple users to explore the topics of CORD-19 at the same time. The application, with the data and the model, has been dockerized to facilitate distribution and deployment.

Results

Topic visualizer

To support the full exploration of the fine-grain topic structure of CORD-19, we developed the Topic Visualizer, available as an interactive dashboard (`http://gmql.eu/cortoviz/`). The tool includes several interfaces for topics exploration, tracking their evolution in time, and associating behavior with their statistical significance. Users will experience a free keyword-based topic search, producing – as an outcome – results in the format displayed in Fig. 4 within a few seconds.

Fig. 4 portrays the four main areas of the dashboard. Panel (A) shows the search feature, where users can enter an arbitrary search keyword to start their search session. To demonstrate the approach, Fig. 4*A* shows the search of topics related to 'ventilator'.

Panel (B) shows six top-ranked topics through their word clouds, ordered by similarity (spanning from 0.91 for topic ID 320 to 0.45 for topic 151). The user can filter topics for further visualization and analysis (in the example, topics 'ventilator' and 'prone-positioning' are selected).

Panel (C) shows the plots of the time series of articles associated with the selected topics, respectively through line plots (above) and stacked histograms (below). The y-axis of the line plot shows the number of scientific abstracts represented in the shown topic using their relative frequency w.r.t. to the total number of abstracts for each bin, whereas the y-axis of the histogram holds absolute counts of scientific papers for each bin. The relative frequencies

of appearance for determined periods represent pointwise measures of the topic intensity (Krause et al., 2006). The x-axis is used to scan the temporal evolution, binned according to time resolution, ranging between one and four weeks, that can be manually changed. Plots are augmented with contextualizing statistics: a plot of global active cases of COVID-19 (expressed in million units and highlighted using a gray area underneath the plot) and vertical red lines marking peculiar events of the pandemic (first cases in China, the world's emergency declaration, first lockdowns, start of vaccine testing, start of vaccination campaign, onset of Delta and Omicron variants, worldwide peak of 400 million cases, final release of CORD-19).

Panel (D) shows the box for testing the statistical significance of changes in topic intensity. Users specify the topic to be tested among those shown in panel (B). The user also uses sliders to define two time intervals. All articles whose publishing date falls within the two intervals are included within two groups, and the Kruskal-Wallis test – a non-parametric test on the difference of the medians of intensity values – is performed; the null hypothesis states that there are no significant differences among the groups, while the alternative hypothesis suggests that the two groups differ. The null hypothesis is rejected when the test's H statistic is above a certain threshold (correspondingly, the test's p-value is below a certain threshold, i.e., 5%). Fig. 4*D* shows that the topic with ID 320 (focused on different formulations of ventilator/ventilation) was significantly different (indeed, much more intense) in the March-September 2020 interval when compared with the June-December 2021 interval, having a p-value below 0.03%; when the difference is not significant, a red cross appears (see *Materials and Methods* for details on the H statistic of the test).

Search sessions

A broad spectrum of search sessions can be performed quickly and flexibly, as the system's average response time is about three seconds regardless of the number of bins. Fig. 5 reports a collection of use cases regarding topics that characterized the COVID-19 pandemic and discusses insights provided by the tool; plots use the default 2-week resolution.

Panel (A) shows two interesting topics extracted by searching the word 'variant'.

- Topic-107 (focused on infectivity and pathogenicity of specific (sub)variants, among which omicron) appeared in December 2021 and had a dramatic rise peaking at the beginning of 2022, when related abstracts covered more than 8/1000 abstracts in CORD-19, with 55 in absolute numbers. Interestingly, we observe that topic-107's profile anticipates a peak of COVID-19 cases in the background, corresponding to the fourth wave of the pandemic.

- Topic-262 (focused on the delta variant, its sub-kinds, and infectivity effect) started between July and November 2021, with the rise of SARS-CoV-2 variants, and remained present throughout the pandemic.

Panel (B) shows two topics extracted by searching the word 'vaccine'.

- Topic-3 (focused on vaccines and vaccinations) appears during the first months of the first lockdown period and starts increasing intensity at the beginning of large-scale vaccination testing; it further increases intensity until August-September 2021, then remains stable at 5/1000 abstracts for every bin (2 weeks).

- Topic-180 (on immunization and immunization-related effects) is already present at the beginning of 2020, but at a very low intensity until January 2021, when vaccines become available in many countries (World Health Organization, 2023a); after that date, the topic increases its intensity until the end of the observation in June 2022.

Panel (C) shows topics related to the 'outbreak' keyword with very similar decreasing trends.

- Topic-202 (on epidemic outbreaks) peaks at the beginning of the pandemic, when the World Health Organization declares COVID-19 as a global health emergency; then, it rapidly decreases.

- Topic-80 (on the influenza virus) peaks at the beginning of the pandemic, with more than 1/100 of abstracts per bi-weekly bin, an indication that before the pandemic outbreak, a relevant number of articles on viral species were focused on influenza. After the first six months, the topic relatively decreases and stabilizes at a lower intensity, reflecting the typical interest in influenza, rather independent of the course of the pandemic.

Panel (D) shows topics related to 'olfactory' and 'long covid'.

- Topic-34 (olfactory) peaks at the beginning of the pandemic, when loss of odor sensing was first associated with COVID; it declines when this symptom becomes more known.
- Topic-44 (long covid) is present after the first pandemic wave in May 2020 and grows, at an intermittent rate, until the end of observation, where it presents the maximum relative frequency of 4/1000 abstracts in CORD-19, i.e., 46 articles.

Panel (E) shows topics related to 'pneumonia'.

- Topic-286 (highlighting central terms in pneumonia seen as a pulmonary disease) peaks at the beginning of the pandemic, when severe covid cases were associated with interstitial pneumonia. It then decreases, while the relevance of other clinical factors rises.
- Topic-14 (highlighting other aspects such as myocarditis and dyspnoea) specularly grows in interest, while the clinical models of severe COVID-19 link it to other co-morbidities.

Panel (F) shows topics related to 'telemedicine' and 'contact tracing'.

- Topic-1 (telemedicine) is quite relevant at all times of the pandemic, as the use of remote, domotic controls for monitoring a person's health has been a central theme throughout.
- Topic-104 (contact tracing) is less relevant, and it is possible to spot a decreasing trend, as the practical limitations of contact tracing became more evident.

Topic trend comparison

Table 2 describes 55 manually curated topics, grouped into 10 classes, which present interesting facets of the evolution of the pandemic.

Pandemic outbreak. The first group includes topics that were very popular (high frequency) during the pandemic outbreak, then lost interest. Among them, the outbreak in Italy – the first European country hit by the virus (Capobianchi et al., 2020; Cerqua and Di Stefano, 2022) – and the lack of preparedness in terms of organization or medical, protective, and testing equipment. Still, in the early period, the burden of COVID-19 caused several postponements and cancellations of surgeries and operations, with deleterious effects on other pathologies.

Understanding the causes of severe disease. At the beginning of the pandemic, clinicians were struggling to understand the causes of the pandemic, originally considered a pulmonary disease, and later on better classified as a vascular-inflammatory disease. Red areas illustrate well the progression of the understanding process, throughout the first year of the pandemic (2020). We note that the neuroinflammation and neuropathies, related to the long-term effects of COVID-19, are slightly delayed, compared to the other topics in this group.

Coronavirus severity and general traits. Effects of general traits (pregnancies, alcoholism, obesity) were studied throughout the pandemic. At a given point, the term *long Covid* started to denote symptoms that persist after the end of the acute COVID-19 disease.

Coronavirus and co-morbidities. Important studies relate COVID-19 severity to cancer, breast cancer, diabetes and hyperglycemia, and other more specific comorbidities such as myocarditis, Kawasaki disease, and fungal infections.

Innovative treatments. Throughout the pandemic, a number of innovative treatments were tried, with quite controversial effects. Among them, are chloroquine, angiotensin, heterocyclic compounds, flavonoids, and antimycobacterial treatments. Other classical treatments, such as corticosteroids and opioids, were evaluated throughout the pandemic.

SARS-CoV-2. During the pandemic, virologists have studied the SARS-CoV-2 virus and its evolution; some terms relate to immunological aspects (reception bindings, serology, epitopes), other to the evolution of the genome, including variants, and specifically some important variants such as Delta and Omicron; of course, these terms are created at the time when variants first appear.

Vaccination. The interest in these topics grew after the end of 2020 when the first vaccines became available to the public for mass immunization. Discussions on vaccination constantly increased until the end of the study.

Social aspects. Aspects related to the mental health of caregivers rise after the first year of the pandemic, then remain at high intensity. Similarly, changes in lifestyle imposed by the lockdowns, such as the work-from-home policies, show an increasing trend. Tourism and traveling exhibit a seasonal trend, with several equally spaced peaks. The effects of distance learning were also studied, including the impact on children and adolescents, such as lack of socialization and physical activity.

Technology. Telemedicine and telehealth, made necessary by the criticality of health systems, show an increase in intensity, as many studies were devoted to the benefits induced by telemedicine in COVID-19. Instead, technologies for contact tracing, which seemed promising at the beginning of the pandemic, had an initial peak and then a decreasing trend.

Pandemic models. Several communities studied the pandemic from a variety of viewpoints, including epidemiological models, virological models, zoonotic models that focus on interactions with animal species, and environmental models (principally on pollution); these exhibit different peaks of intensity all over the pandemic.

Pipeline assessment

Execution and time performances

End-to-end execution of the pipeline on the CORD-19 dataset (1,056,660 initial records, reduced to 327K after filtering), employing our one-node server powered by an Intel Xeon CPU E5-2660 with 56 cores and 378GB of RAM (without any GPU), took 19 hours and 47 minutes. Our logs reported the time required for specific steps: Data preparation 8 minutes; Optimization, including model selection (5 hours and 9 minutes) and UMAP + HDBSCAN parameter selection (8 hours and 16 minutes); fitting of BERTopic model (2 hours and 36 minutes); and dynamic modeling (3 hours and 38 minutes).

Evaluation

The goal of the proposed pipeline is to discover a possible representation of the latent structure of the dataset (unsupervised learning problem). Then, our pipeline applies to datasets that do not provide any classification into pre-defined topics' taxonomies. To evaluate the effectiveness of our pipeline we propose a small experiment based on manually curated classification; we limit this assessment to a randomly selected sample of 50 abstracts. Specifically, by using the 354 identified topics as predefined classes, we manually classified 50 papers not included in the CORD-19 but relevant to the COVID-19-related literature. Then, we compared this ground truth with the classification obtained automatically by our pipeline (see Supplementary Material). The results indicate a weighted 0.867 precision, 0.86 recall, and 0.854 F1-score. These were obtained by leveraging the `precision_recall_fscore_support` function of the `sklearn.metrics` Python module, using the '`weighted`' option of the average parameter, which calculates metrics for each label and finds their average weighted by the number of true instances for each label.

Applying the paradigm to a novel domain

The full-stack process described in this research is virtually applicable to any corpus of medium-to-large-sized textual documents, using any topic model of choice, and a time-series visualizer. The main challenge stands in the data collection stage, in charge of gathering data on other domains of interest.

To demonstrate the applicability of the pipeline, we conducted a data collection activity targeting climate change-related scientific literature. To this end, we implemented a method to build another corpus of abstracts from scientific literature exploiting the public endpoint of the APIs provided by the Springer Nature Group. Specifically, we requested from the Springer Meta API (`http://api.springernature.com/meta/v2/json`) all the articles listed under the subject "Climate Change" (specified in the query of the endpoint (Springer Nature, 2023)). We obtained a dataset of 33,723 scientific abstracts, upon which we performed an exploratory analysis of records and their metadata—see Fig. 6, whose panel (A) shows the significant increase in the volume of scientific publications about climate change in the last 40 years. We then cleaned the dataset through the 'Data Preparation' stage of the CORToViz pipeline, obtaining the final dataset of abstracts (29,886, after language selection). Then, the self-tuning data pipeline developed for CORToViz (Fig. 3) was reapplied on the abstracts to obtain a new topic model, that includes 166 topics; these could be then explored with dashboards similar to CORToViz; note that our work required minimal adaptation and was performed in about two days by the first author.

End-to-end execution of the pipeline on this smaller dataset, using the same setup as for CORD-19, took 3 hours overall.

3. Discussion

By critically analyzing the scientific abstracts of the CORD-19 dataset, this research has shed light on key factors of the evolution of research questions on COVID-19, SARS-CoV-2, and the whole pandemic phenomenon. The

proposed technological pipeline and associated visualizer combine state-of-the-art methods, targeting both data extraction efficiency and user-friendly topic extraction and visualization with integrated statistical testing. Through the lens of scientific research, CORToViz enables the understanding of individual aspects (topics) that characterized the COVID-19 crisis worldwide, with their interactions and timing. Additionally, we also showcased the benefit of having a statistical approach for dynamic topic modeling built on the results of deep learning-based language models. For the purpose of this work, we have chosen the most comprehensive dataset (CORD-19) available to date, without the need for integrating different datasets. Overall, our pipeline is very fast and its results match our expectations on topic identification (F1-score 0.854).

Our proposed pipeline exploits BERTopic to build a fully-fledged domain-independent automatic topic exploration architecture; with this, it presents some limitations: (1) Single modules of the pipeline are used *as is*, while employing alternative technologies may improve the overall results of the topic modeling unsupervised learning. (2) The user experience provided in the CORToViz visualizer is currently restricted to a simple interface, aimed at proving our concept. (3) The pipeline has so far been applied to self-contained limited-in-time text corpora. (4) The current architecture has been demonstrated only for topics in research/scientific literature.

Future work will include mitigation strategies addressing these aspects. To address (1), we will explore other language models and large language models, topic modeling algorithms with neural components instead of clustering, as well as enhanced methods to extract high-quality textual representations for each topic, taking into account the coherence and diversity of the extracted topics. To address limitation (2), we plan to enhance the flexibility of allowed queries and visualization of word clouds; these changes will be evaluated with extensive user studies. To address limitation (3), progress of this research could consider extending the COVID-19 corpus with other corpora targeting the last year of the pandemic. In particular, LitCOVID is still actively curated at the time of writing. Unfortunately, it is focused on PubMed, thus it covers a smaller fraction of research articles w.r.t. CORD-19; it excludes non-medical articles, which are instead an important portion of CORD-19. Along this direction, we will consider the implementation of an evolving, self-updating pipeline, including the use of time-series analysis for the potential identification of trending topics.

Since the methods embedded in the modeling pipeline and in the dashboard are not specific to COVID-19 and SARS-CoV-2, the CORD-19 topic extraction pipeline and visualizer are easily adapted to literature repositories with a similar organization. To demonstrate this aspect, we employed the public API of the Springer Nature Group to retrieve publications about climate change, and then we quickly applied the full pipeline to build a dataset, organized by topic, fully compatible with our topic visualizer.

To address limitation (4), we aim to investigate the pipeline's applicability to non-scientific text.

In perspective, an approach like the one of CORToViz can be applied to both highly technical texts (e.g., scientific research abstracts) and general texts (e.g., book reviews). The flexibility of the approach makes it applicable to very diverse domains and markets; it enables improved access to summarized and digested content both on domain-specific Web content (e.g., reviews on water-scoping machines) and more domain-general ones (e.g., climate change), possibly contributing to e-learning objectives (Badawy et al., 2021) or Web Page topics summarization (Mahmoud et al., 2016).

Overall, our contribution enables addressing the needs of stakeholders requiring a one-click stack, providing immediate high-level analytics to quickly grasp trends, for instance in e-commerce reviews, events feedback tweets, public engagement threads, or any other business in which observing temporal trends is crucial.

Data and code availability statement The original CORD-19 dataset is available at the GitHub repository of the project, at the URL `https://github.com/allenai/cord19`. Both data processing pipeline and application are available as Docker images on `https://hub.docker.com/r/frinve/cortoviz/`. The CORToViz application is freely available on `http://gmql.eu/cortoviz`.

Authorship contribution statement. F.I. and A.B. conceived the work; F.I., A.B., and S.C. jointly conceptualized and designed the framework; S.C. insisted on making the framework reusable for other domains; F.I. selected a coherent set of up-to-date technologies and developed the core pipeline; F.I. and A.B. curated the user experience; F.I. drafted the manuscript, A.B. and S.C. improved it; all authors revised the final version of the manuscript; S.C. supervised the project.

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

Funding. This research is supported by the TETYS project, a beneficiary of the NGI Search 2nd Open Call. Funded

by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them. Funded within the framework of the NGI Search project under grant agreement No 101069364.

References

American Society for Microbiology, 2023. COVID-19 (SARS-CoV-2 Coronavirus) Resources. https://asm.org:443/Resource-Pages/COVID-19-Resources. Last accessed: March 22nd, 2024.

Angelov, D., 2020. Top2Vec: Distributed Representations of Topics. arXiv:2008.09470 https://doi.org/10.48550/arXiv.2008.09470.

Badawy, A., Fisteus, J.A., Mahmoud, T.M., Abd El-Hafeez, T., 2021. Topic extraction and interactive knowledge graphs for learning resources. Sustainability 14, 226.

Berchialla, P., Urru, S., Sciannameo, V., 2021. The effect of COVID-19 on scientific publishing in Italy. Epidemiologia & Prevenzione 45, 449–451.

Capobianchi, M.R., Rueca, M., Messina, F., Giombini, E., Carletti, F., Colavita, F., Castilletti, C., Lalle, E., Bordi, L., Vairo, F., et al., 2020. Molecular characterization of SARS-CoV-2 from the first case of COVID-19 in Italy. Clinical Microbiology and Infection 26, 954–956.

Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P., Quarteroni, S., 2013. Information Retrieval Models, in: Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P., Quarteroni, S. (Eds.), Web Information Retrieval, Springer, Berlin, Heidelberg. pp. 27–37.

Cerqua, A., Di Stefano, R., 2022. When did coronavirus arrive in europe? Statistical Methods & Applications 31, 181–195.

Chen, Q., Allot, A., Lu, Z., 2021. LitCovid: An Open Database of COVID-19 Literature. Nucleic Acids Research 49, D1534–D1540.

Chen, Y., Zhang, H., Liu, R., Ye, Z., Lin, J., 2019. Experimental explorations on short text topic mining between LDA and NMF based Schemes. Knowledge-Based Systems 163, 1–13.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D., 2020. SPECTER: Document-level representation learning using citation-informed transformers, in: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 2270–2282.

Colavizza, G., Costas, R., Traag, V.A., van Eck, N.J., van Leeuwen, T., Waltman, L., 2021. A Scientometric Overview of CORD-19. PLOS ONE 16, e0244839.

Dagdelen, J., Trewartha, A., Huo, H., Fei, Y., He, T., Cruse, K., Wang, Z., Subramanian, A., Justus, B., Ceder, G., et al., 2023. Covidscholar: An automated covid-19 research aggregation and analysis platform. PloS one 18, e0281147.

Deka, P., Jurek-Loughrey, A., P, D., 2022. Evidence Extraction to Validate Medical Claims in Fake News Detection, in: Traina, A., Wang, H., Zhang, Y., Siuly, S., Zhou, R., Chen, L. (Eds.), Health Information Science, Springer Nature Switzerland, Cham. pp. 3–15.

Ebeling, R., Sáenz, C.A.C., Nobre, J.C., Becker, K., 2022. Analysis of the influence of political polarization in the vaccination stance: the brazilian covid-19 scenario, in: Proceedings of the International AAAI Conference on Web and Social Media, pp. 159–170.

Egger, R., Yu, J., 2022. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Frontiers in Sociology 7, 886498.

Elsevier, 2023. Novel Coronavirus Information Center. https://www.elsevier.com/connect/coronavirus-information-center. Last accessed: March 22nd, 2024.

Falkenberg, M., Galeazzi, A., Torricelli, M., Di Marco, N., Larosa, F., Sas, M., Mekacher, A., Pearce, W., Zollo, F., Quattrociocchi, W., et al., 2022. Growing polarization around climate change on social media. Nature Climate Change 12, 1114–1121.

Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794 https://doi.org/10.48550/arXiv.2203.05794.

Huang, T.H.K., Huang, C.Y., Ding, C.K.C., Hsu, Y.C., Giles, C.L., 2020. CODA-19: Using a non-expert crowd to annotate research aspects on 10,000+ abstracts in the COVID-19 open research dataset, in: Verspoor, K., Cohen, K.B., Dredze, M., Ferrara, E., May, J., Munro, R., Paris, C., Wallace, B. (Eds.), Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Association for Computational Linguistics, Online.

James, G., Witten, D., Hastie, T., Tibshirani, 2021. An introduction to statistical learning: with applications in R. Springer New York, NY. eBook ISBN 978-1-0716-1418-1, https://doi.org/10.1007/978-1-0716-1418-1.

Jayabharathy, J., Kanmani, S., Parveen, A.A., 2011. Document clustering and topic discovery based on semantic similarity in scientific literature, in: 2011 IEEE 3rd International Conference on Communication Software and Networks, pp. 425–429.

Korn, D., Bobrowski, T., Li, M., Kebede, Y., Wang, P., Owen, P., Vaidya, G., Muratov, E., Chirkova, R., Bizon, C., Tropsha, A., 2021. COVID-KOP: Integrating emerging COVID-19 data with the ROBOKOP database. Bioinformatics (Oxford, England) 37, 586–587.

Krause, A., Leskovec, J., Guestrin, C., 2006. Data association for topic intensity tracking, in: Proceedings of the 23rd international conference on Machine learning, pp. 497–504.

Kruskal, W.H., Wallis, W.A., 1952. Use of Ranks in One-Criterion Variance Analysis. Journal of the American Statistical Association 47, 583–621.

Logette, E., Lorin, C., Favreau, C., Oshurko, E., Coggan, J.S., Casalegno, F., Sy, M.F., Monney, C., Bertschy, M., Delattre, E., et al., 2021. A machine-generated view of the role of blood glucose levels in the severity of COVID-19. Frontiers in Public Health 9, 695139.

MacMillan Learning, 2022. Information about COVID-19. https://covid19.macmillanlearning.com/. Last accessed: March 22nd, 2024.

Mahmoud, T., Abd-El-Hafeez, T., El-Deen, D., 2016. A design of an automatic web page classification system. British Journal of Applied Science & Technology 18, 1–14.

Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D., Ortiz-Ospina, E., Roser, M., 2020. Coronavirus Pandemic (COVID-19). https://ourworldindata.org/coronavirus. Last accessed: March 22nd, 2024.

McInnes, L., Healy, J., Astels, S., 2017. hdbscan: Hierarchical Density Based Clustering. Journal of Open Source Software 2, 205.

McInnes, L., Healy, J., Melville, J., 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 https://doi.org/10.48550/arXiv.1802.03426 .

Meng, Y., Zhang, Y., Huang, J., Zhang, Y., Han, J., 2022. Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations, in: Proceedings of the ACM Web Conference 2022, pp. 3143–3152.

Moody, C.E., 2016. Mixing Dirichlet Topic Models and Word Embeddings to Make Lda2vec. arXiv:1605.02019 https://doi.org/10.48550/arXiv.1605.02019 .

Moulavi, D., Jaskowiak, P.A., Campello, R.J., Zimek, A., Sander, J., 2014. Density-based clustering validation, in: Proceedings of the 2014 SIAM international conference on data mining, SIAM. pp. 839–847.

Mueller, A.C., 2023. Wordcloud. https://github.com/amueller/wordcloud. Last accessed: March 22nd, 2024.

National Institutes of Health, 2023. NIH OPA iSearch COVID-19 Portfolio. https://icite.od.nih.gov/covid19/search/. Last accessed: March 22nd, 2024.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Édouard Duchesnay, 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Pourhatami, A., Kaviyani-Charati, M., Kargar, B., Baziyad, H., Kargar, M., Olmeda-Gómez, C., 2021. Mapping the intellectual structure of the coronavirus field (2000–2020): A co-word analysis. Scientometrics 126, 6625–6657.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. OpenAI blog 1, 9.

Řehůřek, R., Sojka, P., 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. pp. 45–50.

Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Inui, K., Jiang, J., Ng, V., Wan, X. (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China. pp. 3982–3992.

Serna García, G., Al Khalaf, R., Invernici, F., Ceri, S., Bernasconi, A., 2023. CoVEffect: interactive system for mining the effects of SARS-CoV-2 mutations and variants based on deep learning. GigaScience 12, giad036.

Shahapure, K.R., Nicholas, C., 2020. Cluster Quality Analysis Using Silhouette Score, in: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), pp. 747–748.

Shao, Y., Taylor, S., Marshall, N., Morioka, C., Zeng-Treitler, Q., 2018. Clinical Text Classification with Word Embedding Features vs. Bag-of-Words Features, in: 2018 IEEE International Conference on Big Data (Big Data), pp. 2874–2878.

Springer Nature, 2023. Querystring parameter - api portal. https://dev.springernature.com/querystring-parameters. Last accessed: March 22nd, 2024.

Streamlit, 2023. A Faster Way to Build and Share Data Apps. https://streamlit.io/. Last accessed: March 22nd, 2024.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I., 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). Https://openreview.net/forum?id=wCu6T5xFjeJ.

The Lens, 2021. Human coronaviruses data initiative. https://about.lens.org/covid-19/. Last accessed: March 22nd, 2024.

Tran, B.X., Ha, G.H., Nguyen, L.H., Vu, G.T., Hoang, M.T., Le, H.T., Latkin, C.A., Ho, C.S., Ho, R.C., 2020. Studies of Novel Coronavirus Disease 19 (COVID-19) Pandemic: A Global Analysis of Literature. International Journal of Environmental Research and Public Health 17, 4095.

United Nations News, 2023. WHO Chief Declares End to COVID-19 as a Global Health Emergency. https://news.un.org/en/story/2023/05/1136367. Last accessed: March 22nd, 2024.

Valika, T.S., Maurrasse, S.E., Reichert, L., 2020. A Second Pandemic? Perspective on Information Overload in the COVID-19 Era. Otolaryngology-Head and Neck Surgery 163, 931–933.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 17, 261–272.

Wadden, D., Lin, S., Lo, K., Wang, L.L., van Zuylen, M., Cohan, A., Hajishirzi, H., 2020. Fact or Fiction: Verifying Scientific Claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online. pp. 7534–7550.

Wang, L.L., Lo, K., 2021. Text mining approaches for dealing with the rapidly expanding literature on COVID-19. Briefings in Bioinformatics 22, 781–799.

Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R.M., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D.A., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A.D., Wang, K., Wang, N.X.R., Wilhelm, C., Xie, B., Raymond, D.M., Weld, D.S., Etzioni, O., Kohlmeier, S., 2020. CORD-19: The COVID-19 open research dataset, in: Verspoor, K., Cohen, K.B., Dredze, M., Ferrara, E., May, J., Munro, R., Paris, C., Wallace, B. (Eds.), Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Association for Computational Linguistics, Online.

Wang, Q., Li, M., Wang, X., Parulian, N., Han, G., Ma, J., Tu, J., Lin, Y., Zhang, R.H., Liu, W., Chauhan, A., Guan, Y., Li, B., Li, R., Song, X., Fung, Y., Ji, H., Han, J., Chang, S.F., Pustejovsky, J., Rah, J., Liem, D., ELsayed, A., Palmer, M., Voss, C., Schneider, C., Onyshkevych, B., 2021. COVID-19 literature knowledge graph construction and drug repurposing report generation, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, Association for Computational Linguistics, Online. pp. 66–77.

Wikipedia, 2023. Timeline of the COVID-19 pandemic. https://en.wikipedia.org/wiki/Timeline_of_the_COVID-19_pandemic. Last accessed: March 22nd, 2024.

Wise, C., Calvo, M.R., Bhatia, P., Ioannidis, V., Karypus, G., Price, G., Song, X., Brand, R., Kulkani, N., 2020. COVID-19 knowledge graph: Ac-

celerating information retrieval and discovery for scientific literature, in: Proceedings of Knowledgeable NLP: the First Workshop on Integrating Structured Knowledge and Neural Networks for NLP.

World Health Organization, 2023a. A brief history of vaccination. https://www.who.int/news-room/spotlight/history-of-vaccination/a-brief-history-of-vaccination. Last accessed: March 22nd, 2024.

World Health Organization, 2023b. Global research on coronavirus disease (COVID-19). https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov. Last accessed: March 22nd, 2024.

Zhang, Y., Cai, X., Fry, C.V., Wu, M., Wagner, C.S., 2021. Topic evolution, disruption and resilience in early COVID-19 research. Scientometrics 126, 4225–4253.

Šćepanović, S., Constantinides, M., Quercia, D., Kim, S., 2023. Quantifying the Impact of Positive Stress on Companies from Online Employee Reviews. Scientific Reports 13, 1603.

Figures and Tables

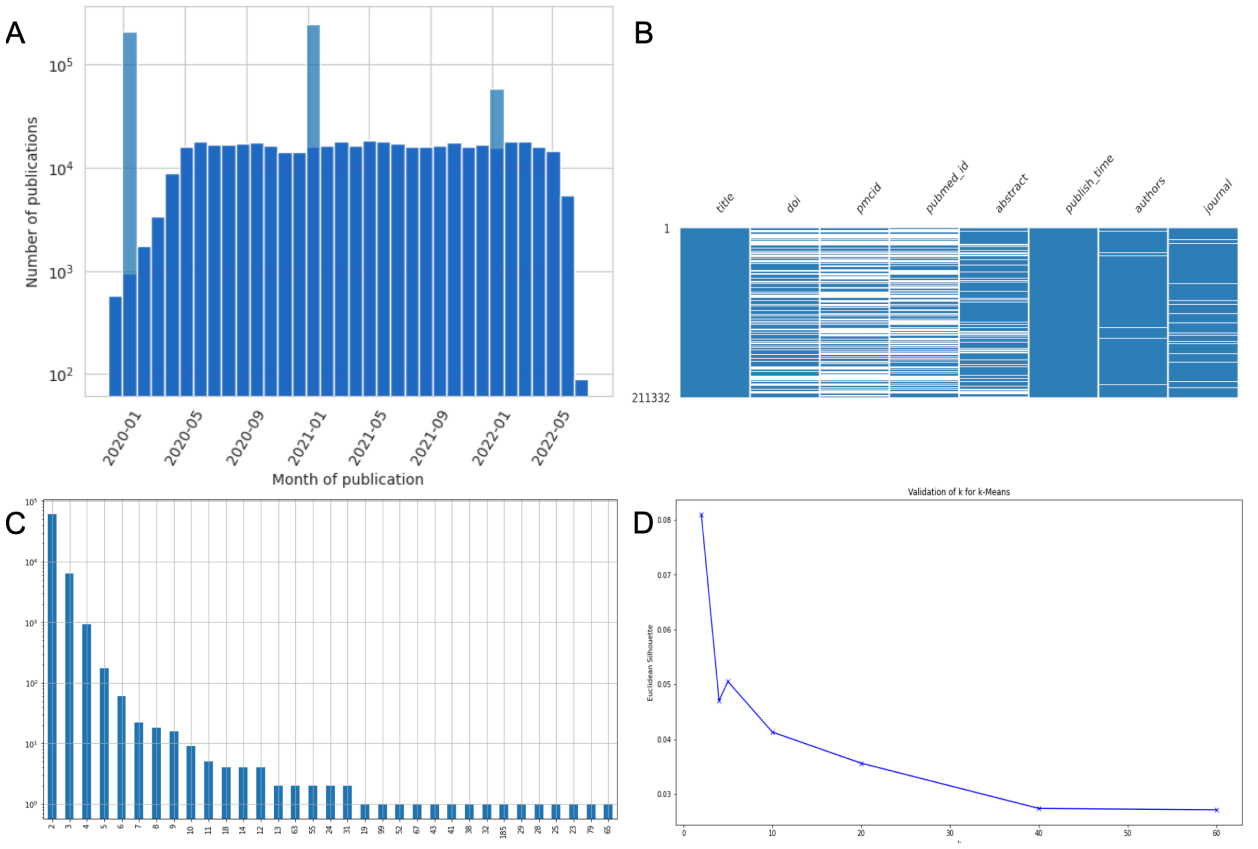

Figure 1: Visualizations of the exploratory analyis of CORD-19 data and metadata. (A) Monthly number of publications in CORD-19. The number increases in the first months of 2020, then is rather stable, until April 2022, when the trend starts decreasing; CORD-19 was updated until June 2022. In light color, the spikes of publications with just the year in their metadata were converted to the first of January; these entries were removed. (B) Data-density display of eight metadata fields for a sample of 20% of the dataset. We retain articles with abstract and publish_time metadata. (C) Distribution of the number of duplicates. The majority of articles, on the left of the distribution, have a single duplicate (typically without the doi), representing a preprint non-peer-reviewed version uploaded on public archives before publication; only a few documents are present in the dataset with a high number of replicated entries. (D) Silhouette score for k-Means for different values of k, which indicates the number of clusters. A spike in the line plot means that that value is a good candidate for the number of clusters; the figure clearly indicates that five is a good candidate, and then selected for the exploratory clustering analysis.

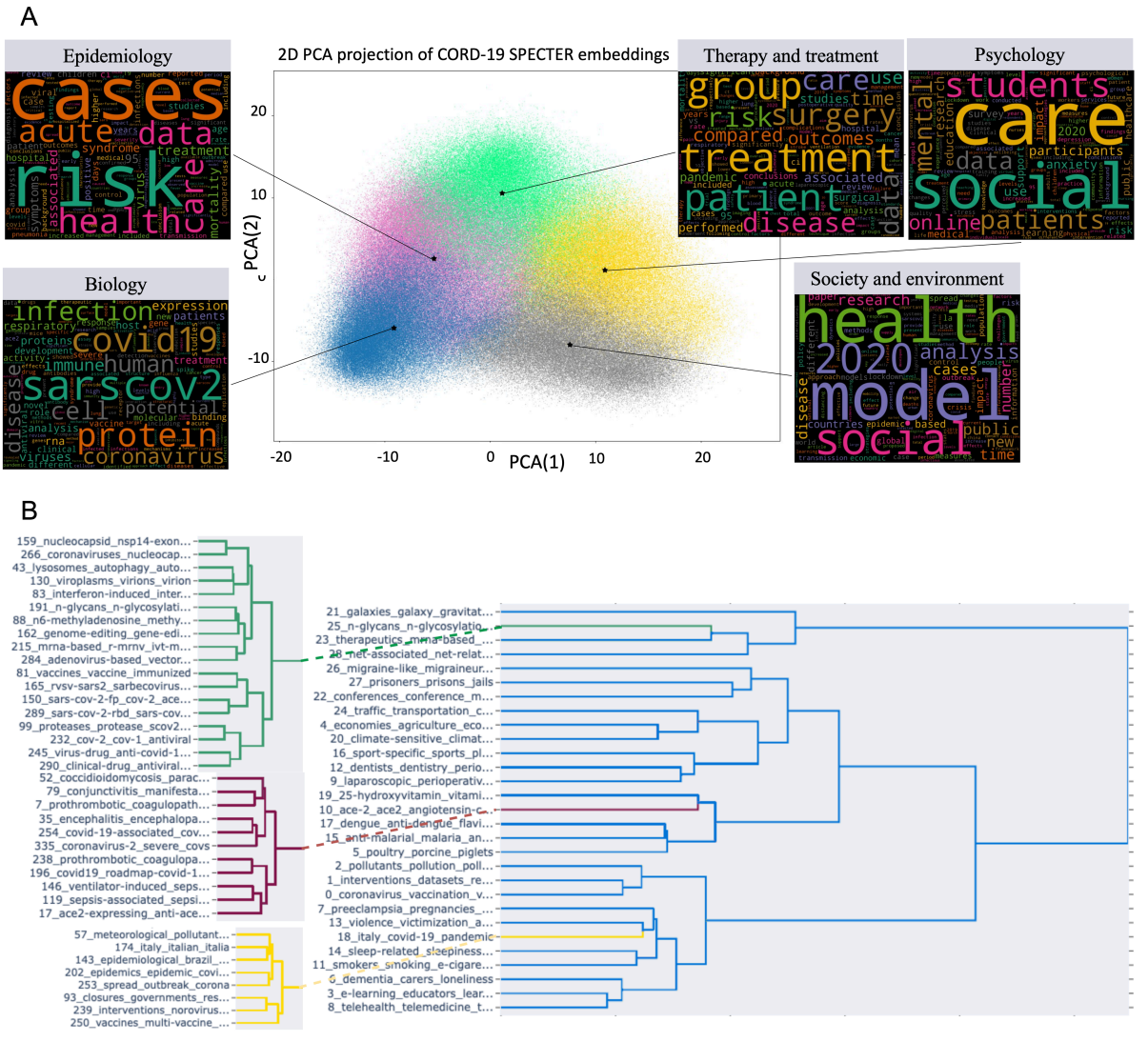

Figure 2: Topic clustering produced by the preliminary and fine-grain clustering methods. We show and compare the clusters introduced in the Results section. (A) Scatter plot of the exploratory clustering analysis. The analysis has been performed with k-Means, a classic clustering algorithm. We found five macro-topics and we assessed their content with word clouds. As shown in the figure, the five clusters identify distinct classes of topics, well described by word clouds, which nicely partition the set of articles of CORD-19. (B) Dendrograms of the hierarchical density-based clustering. We then explored topics using a technology-rich pipeline, resulting in a fine-grain topic clustering. The high-level cluster hierarchy, with only 29 clusters, resembles the five macro-topics structure of the preliminary clustering. The full hierarchy includes 354 fine-grained clusters, each related to a specific high-level cluster. We show the hierarchy of the *n-glycans*-related topics, of the *ACE2*-related topics, and of an epidemiology-related topic.

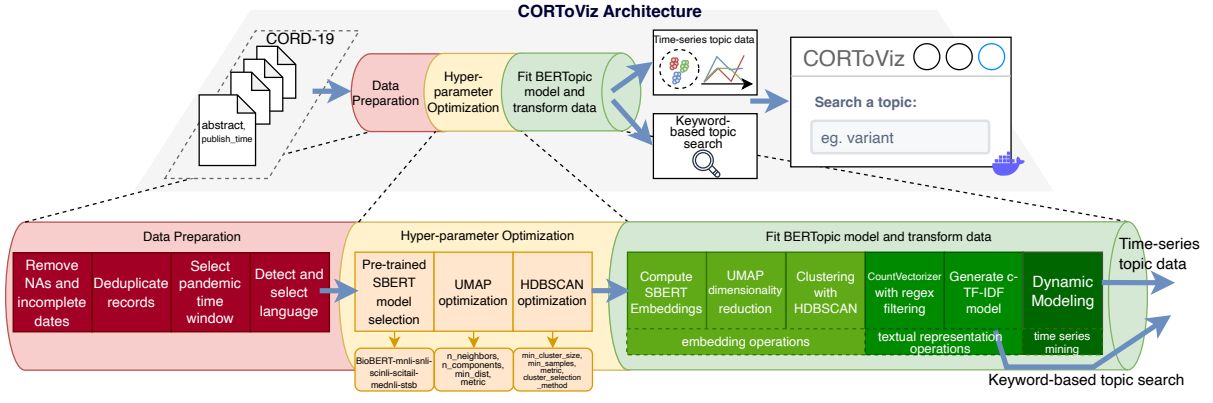

Figure 3: General architecture of CORToViz. The data pipeline consists of three stages: data preparation (red), hyperparameter optimization (yellow), and topic extraction using the BERTopic model (green); the pipeline produces as output the ingredients for the dashboard application, a user-friendly interface for topic selection and display. In the data pipeline, The data preparation stage selects the abstracts with the appropriate metadata from CORD-19; the hyper-parameter optimization finds the values that maximize the performance of the models operating on embeddings; finally, the data transformation generates the artifacts used by the CORToViz dashboard application. The dashboard supports keyword-based topic search and then visualizes the time series information for each topic; each topic is associated with a word cloud, providing insight into the topic's content.

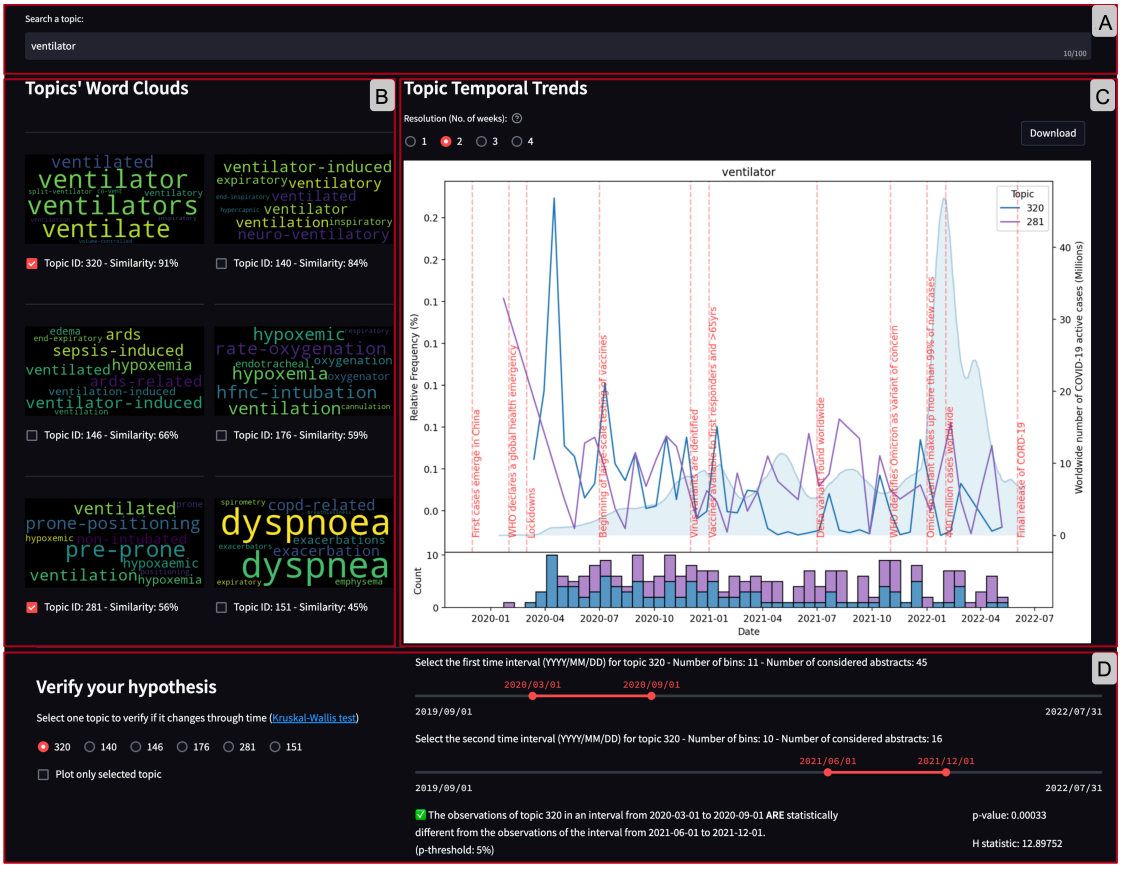

Figure 4: User interface of the CORToViz dashboard. (A) Keyword-based search bar - the example query "*ventilator*" is entered by the user. (B) Six top-ranked topics, explained through their word clouds. The user selects two topics "*ventilator*" and "*prone-positioning*". (C) Line plot of the intensities (i.e., the relative frequencies of appearance) of the two selected topics. The user sets (above) the bin resolution to 2 weeks (options are 1-4 weeks). Histograms (below) show the count of articles associated with the selected topics, with the given bin size of two weeks. (D) Panel showing statistical testing. The user selects the "*ventilator*" topic and sets two time windows, a six-month window at the beginning of the pandemic and a 6-month window at the end of the second year of the pandemic. The tool, at the bottom, reports the result of the Kruskal-Wallis test for the difference between groups. Specifically, it shows the H statistic of 12.89752, which is the statistic of the aforementioned non-parametric test, and that determines the p-value (0.00033). Therefore, since the p-value is below the 5% threshold, the null hypothesis (i.e., no difference in groups) can be rejected, and a green check indicates a statistically significant difference between observations in the two intervals.

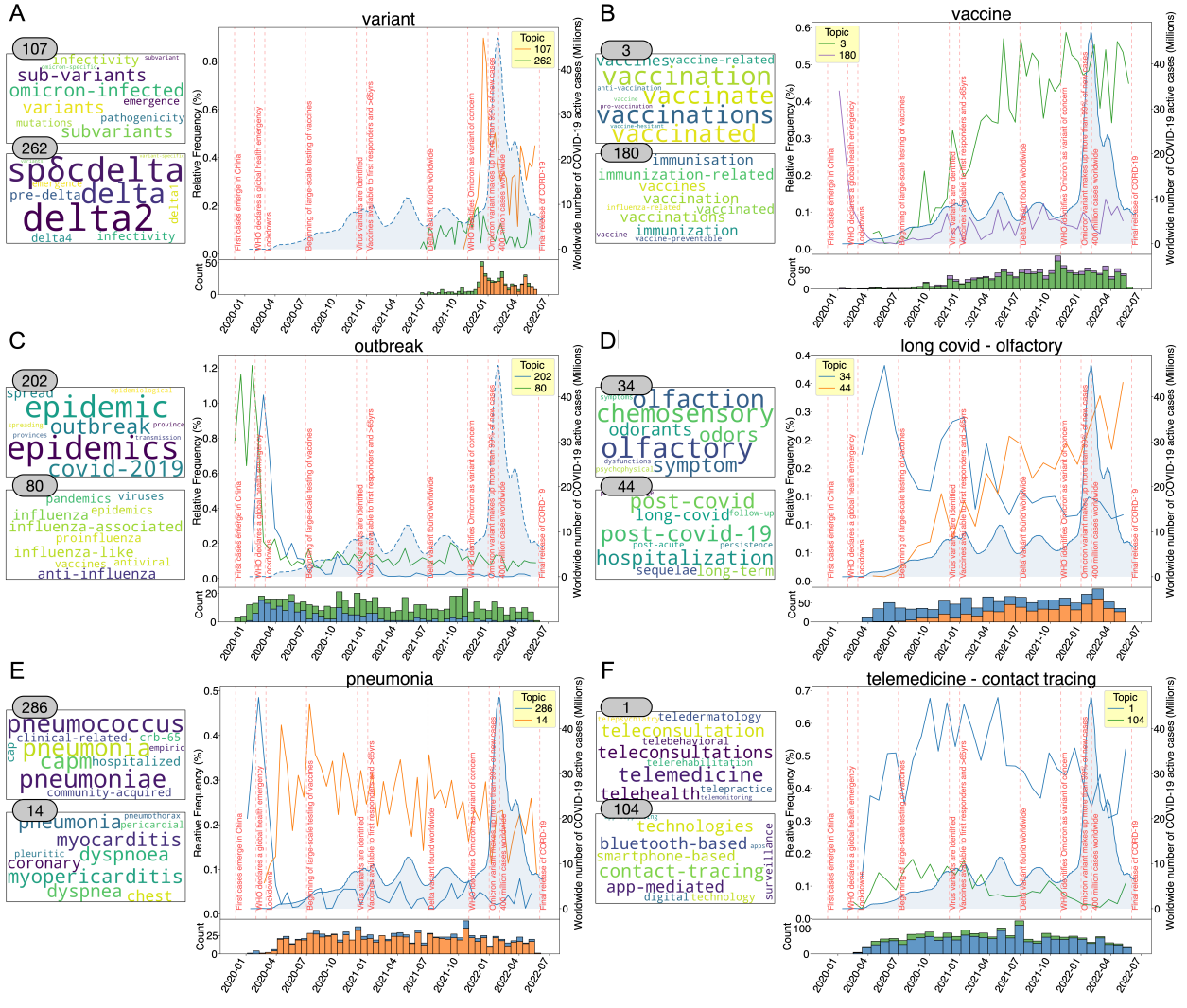

Figure 5: Visualizations of relevant example cases. Each panel corresponds to the search of a keyword on CORToViz (see the title of plots). For each one, we show the word clouds generated for two topics and the line plots of the topics' time series. (A) Variant: topic on (sub)variants, among which omicron, whose spike anticipates a peak in active COVID cases shown in the background; topic on delta that increases when the variant spreads worldwide; (B) Vaccine: generic topic showing an increase in interest over time; immunization topic, more specific, with a similar trend. (C) Outbreak: epidemic-related topic interesting at the beginning, but not very interesting after the first months; influenza, a topic with an early peak representing the large fraction of articles written on influenza prior to COVID, then less relevant and almost unrelated to COVID cases. (D) Olfactory and long covid: the former peaking at the beginning of the pandemic and then decreasing; the latter showing a steadily increasing trend. (E) Pneumonia: the first topic is decreasing in mid-2020 while the second topic, highlighting other co-morbidities, grows in interest. (F) Telemedicine and contact tracing: the first is steadily interesting; instead contact tracing is most interesting in the first months, but then loses interest (as it revealed hard to deploy in reality).

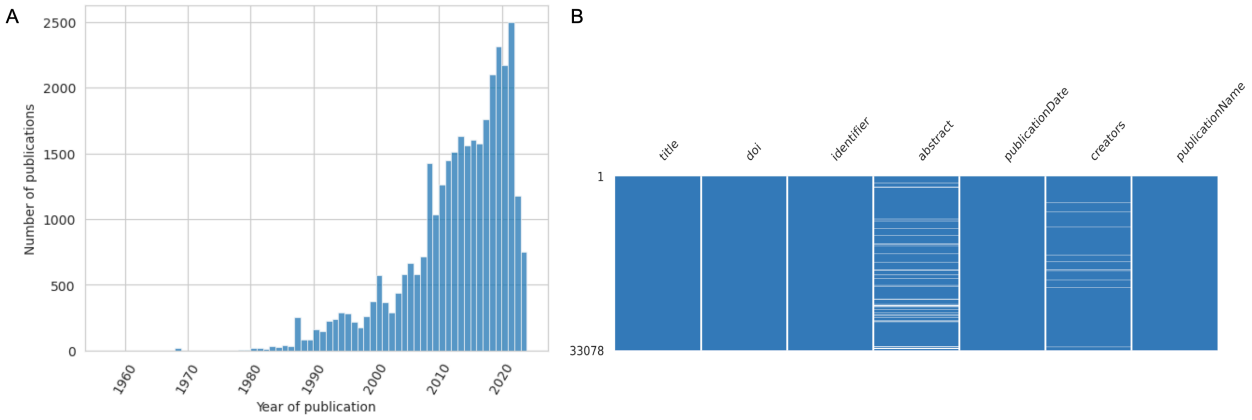

Figure 6: Visualizations of the exploratory analysis of the dataset on climate change from Springer Nature Group. (A) Yearly number of publications. The number is below 500 articles per year until 2008. Then, it ramps to 1,500 articles per year for about one decade. In the last 5 years, the number exceeds 2,000 articles. (B) Data-density display of seven metadata fields for the whole dataset, which shows, in general, the good quality of the dataset. We note some unavailable data in the abstract and creators fields.

Table 2
Heatmap of trends of a manually-curated sample of 55 topics. For each topic, the median intensity (× 1000) has been computed for two-month intervals throughout the pandemic, as observed in CORD-19 (March 2020-June 2022). These values are represented by color nuances from white (lowest) to bright red (highest). To ease reading, we grouped topics with close semantics. Interest trends are either: increasing (e.g., vaccinations, variants, teleworking, long COVID), decreasing (e.g., outbreak topics, symptoms-related, testing and epidemiological modeling), or seasonal (e.g., tourism).

		2020					2021						2022		
Topic ID	**Description**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
	Pandemic outbreak														
174	Italy	4,35	2,07	1,61	1,06	0,88	1,29	0,25	0,45	0,22	0,50	0,21	0,32	0,34	0,28
144	Facemasks and Personal Protective Equipment	1,45	1,03	0,96	1,11	1,17	1,41	0,42	0,71	0,54	0,36	0,59	0,63	0,45	0,33
320	Ventilators	0,73	0,89	0,32	0,43	0,56	0,63	0,45	0,24	0,27	0,28	0,35	0,23	0,33	0,26
279	Management of waste and disinfection	1,27	0,53	0,82	0,66	0,30	0,31	0,58	0,49	0,57	0,52	0,53	0,65	0,34	0,37
136	Self testing	1,12	1,34	1,26	0,66	0,57	0,72	0,91	0,79	0,40	0,56	0,60	0,64	0,70	0,56
131	Postponement and cancellation of surgeries and operations	0,78	1,48	1,36	0,66	1,14	1,12	0,99	0,80	0,48	0,28	0,67	0,50	0,80	0,33
11	Surgeries, laparoscopies and endoscopies	10,25	3,04	3,73	3,32	2,77	1,83	2,48	2,02	2,39	1,70	2,40	2,47	2,29	1,57
	Understanding of the causes of severe disease														
37	Pneumonia and chest scans	7,25	3,55	2,92	2,58	2,62	2,15	1,82	1,29	1,00	0,56	0,60	1,01	0,90	0,56
173	Hyperinflammation, cytokine storm, interleukin and tocilizumab	0	0,95	1,15	0,89	0,88	1,06	0,52	0,68	0,50	0,28	0,67	0,32	0,45	0,44
196	Cardiomyopathy, myocarditis, and cytokines	0,68	1,55	0,96	0,71	0,88	0,90	0,29	0,25	0,40	0,47	0,39	0,55	0,42	0,29
238	Coagulopathies, thrombosis and thromboembolism	0	0,76	0,65	0,73	0,59	0,48	0,28	0,49	0,41	0,24	0,67	0,36	0,47	0,28
34	Symptomps: olfactory and chemosensory disfunctions	1,81	3,08	3,41	1,59	2,22	2,51	1,74	1,82	1,44	1,11	1,77	1,06	1,14	1,00
35	Manifestations of neuroinflammation, neuropathies and encephalitis	0,78	1,90	2,30	2,21	1,76	2,52	2,46	1,47	1,86	1,20	1,51	1,27	1,47	1,54
	Coronavirus severity and general traits														
200	COVID-19 infection in pregnancies: studies on preeclampsia	2,35	0,96	0,97	0,51	0,59	0,50	0,28	0,49	0,57	0,40	0,49	0,53	0,65	0,44
49	Relations with alcoholism and alcohol-associated conditions	1,27	0,53	1,06	1,42	1,18	1,26	1,38	1,58	1,62	1,57	1,57	2,30	2,52	1,63
58	Comorbidity of obesity and COVID-19	1,27	2,00	1,89	1,85	1,77	1,30	1,77	1,97	1,44	1,57	0,94	1,07	0,68	0,89
44	Long COVID and long-term post-COVID effects	0	0,30	0,30	0,66	0,88	0,94	1,24	1,85	1,99	1,84	2,16	1,99	3,16	3,09
	Coronavirus and co-morbidities														
71	Comorbidity and mortality of cancer and malignancies	2,31	3,01	1,57	1,59	1,71	1,01	1,35	0,74	1,24	0,84	0,90	0,36	0,47	0,76
230	Interactions of COVID-19 with HER2 breast cancer	0	0,46	0,32	0,43	0,44	0,54	0,51	0,49	0,45	0,56	0,60	0,58	0,91	0,86
4	Interactions of COVID-19 and diabetes and hyperglycemia	1,82	0,79	1,59	2,32	2,35	2,39	2,98	2,25	3,38	3,41	3,43	4,27	4,74	4,78
220	Vaccine-related and associated myocarditis	0	0	0	0	0	0,26	0,29	0,38	0,49	0,71	0,90	0,76	1,05	1,96
248	Correlation of COVID-19 and Kawasaki disease	0	0,52	0,54	0,76	0,55	0,57	0,68	0,73	0,41	0,47	0,48	0,32	0,25	0,61
52	Interactions of Coccidioidomycosis, fungal infections and COVID-19	0	0,29	0,65	0,40	0,78	0,50	1,12	1,18	2,17	2,51	2,92	2,54	2,80	2,22
	Innovative treatments														
23	Hydroxychloroquine and cardiotoxic side effects	4,23	6,42	4,09	3,32	3,66	2,52	1,59	1,25	1,69	0,80	0,81	0,93	0,63	0,56
179	Clinical studies on antihypertensives targeting angiotensin	0,78	1,64	0,96	0,90	0,88	1,04	0,55	0,51	0,46	0,24	0,25	0,25	0,45	0,33
13	Treatments based on heterocyclic compounds	2,13	1,40	1,93	4,15	2,09	2,44	1,82	2,21	2,89	2,41	2,56	2,70	2,35	1,98
22	Therapies based on flavonoids and other phytochemicals	2,90	0,59	1,09	2,21	1,57	1,57	1,65	1,56	2,27	2,78	2,36	3,28	3,16	2,83
8	Efficacy of antimycobacterial and anti-TBC treatments	2,78	1,24	1,15	2,49	1,73	2,26	2,68	2,12	2,75	3,06	3,36	3,49	4,51	4,55
334	Corticosteroids in in-hospital treatment	0	0,30	0,48	0,43	0,44	0,45	0,38	0,31	0,41	0,28	0,37	0,25	0,23	0,29
24	Opioids, medical and non-medical uses in the pandemic	2,21	0,59	1,28	1,42	1,66	1,93	2,27	2,46	1,99	2,01	2,16	2,26	2,26	3,08
	SARS-CoV-2														
150	SARS-CoV-2 receptor bindings and ACE2	2,16	0,36	0,96	0,71	1,24	0,97	0,63	1,17	0,81	0,49	0,81	0,45	0,50	0,56
15	Serology and immunoassays	1,14	2,34	3,46	2,84	3,05	2,63	3,13	3,25	2,42	1,67	1,77	2,06	1,68	1,36
89	Epitopes of antigens for SARS-CoV-2	0	0,64	0,66	0,95	1,03	1,26	1,34	1,45	0,81	1,20	1,12	1,33	0,84	1,05
299	Phenotype, genome and polymorphisms of SARS-CoV-2	0	0,56	0,32	0,33	0,57	0,64	0,55	0,54	0,28	0,72	0,27	0,32	0,47	0,56
124	Variants and substitutions	0	0,31	0,32	0,28	0,29	1,01	1,36	1,13	1,34	1,20	1,12	0,82	0,46	0,66
262	Delta variant	0	0	0	0	0	0	0	0	0,37	0,83	1,33	0,87	0,69	0,44
107	Omicron variant, subvariants and infectivity	0	0	0	0	0	0	0	0	0	0	0,20	4,51	3,15	3,86
	Vaccination														
59	Infection-prevention campaigns and adherence	0,73	1,34	1,15	1,62	2,36	2,58	1,74	1,25	1,35	0,97	1,30	1,13	0,91	0,84
81	Available vaccines, immunization and immunogenicity	1,86	0,52	0,76	0,86	0,59	1,67	0,99	1,10	1,21	1,57	1,35	1,32	1,13	0,98
215	mRNA-based drugs and vaccines	0,73	0	0,42	0,48	0,29	0,67	0,58	0,50	0,55	0,52	0,90	0,77	0,80	0,70
3	Vaccination: for, against and hesitant	0	0,38	0,32	0,81	1,16	2,09	2,98	3,68	4,10	4,72	5,07	4,93	4,51	4,87
	Social aspects														
27	Coping with loneliness and disconnection for elderly people	0	0,95	1,09	1,00	1,14	1,26	1,36	2,03	1,53	1,95	1,50	1,50	1,47	1,52
39	Stress and burnout of healthcare workers	1,49	1,10	1,28	1,07	1,73	1,43	2,03	1,97	1,69	1,90	2,13	1,96	1,60	2,41
137	Caregivers	1,13	0,84	0,65	1,07	0,87	1,51	0,96	0,73	0,76	0,56	0,39	0,84	0,23	0,57
149	Work from home	0,78	0,30	0,31	0,28	0,54	0,35	0,68	0,80	0,80	0,71	0,81	1,11	1,13	1,14
267	Effects of restrictions on travelling, traffic, commuting	0	0	0,39	0,27	0,29	0,66	0,37	0,26	0,65	0,66	0,45	0,67	0,47	0,47
316	Travelling and tourism	0	0,56	0,32	0,43	0,57	0,35	0,45	0,41	0,25	0,56	0,27	0,25	0,56	0,44
5	e-Learning: teaching, education and university	0,68	1,07	1,57	1,93	2,30	2,35	3,50	3,41	3,36	3,06	3,43	2,56	4,00	3,37
	Technology														
1	Telemedicine, teleconsultations, and telehealth	1,86	4,65	4,90	5,03	5,42	6,00	4,96	6,08	3,98	4,18	5,16	4,06	3,83	3,79
104	Smartphone-based contact tracing technologies	1,73	1,09	1,40	1,32	1,18	1,43	1,49	0,50	0,92	0,88	0,90	0,54	0,68	0,93
	Pandemic models														
20	Estimations of mortality and fatality	2,17	4,20	2,30	2,65	2,77	2,90	2,53	2,04	1,53	2,23	1,50	1,48	1,51	1,80
224	Prediction and prognosis of COVID-19	0,78	0,56	0,32	0,64	0,86	0,96	0,78	0,45	0,35	0,56	0,43	0,39	0,57	0,29
50	Epidemiological modelling and simulations	2,54	3,42	3,16	2,13	1,43	1,91	1,74	1,06	0,97	1,00	1,07	0,64	1,05	1,10
319	Zoonotic and human-to-animal interactions	0	0,33	0,31	0,26	0,29	0,31	0,29	0,24	0,54	0,47	0,60	0,47	0,45	0,56
72	Air-quality, anthropogenic pollution and aerosol	1,36	0,55	0,63	0,99	1,14	1,56	0,76	1,06	1,01	1,39	1,20	1,42	2,10	1,52