



Dynamic topic modelling for exploring the scientific literature on coronavirus: an unsupervised labelling technique

Ibai Guillén-Pacho¹ · Carlos Badenes-Olmedo^{1,2} · Oscar Corcho¹

Received: 16 May 2023 / Accepted: 12 July 2024
© The Author(s) 2024

Abstract

The work presented in this article focusses on improving the interpretability of probabilistic topic models created from a large collection of scientific documents that evolve over time. Several time-dependent approaches based on topic models were compared to analyse the annual evolution of latent concepts in the CORD-19 corpus: Dynamic Topic Model, Dynamic Embedded Topic Model, and BERTopic. Then COVID-19 period (December 2019–present) has been analysed in greater depth, month by month, to explore the evolution of what is written about the disease. The evaluations suggest that the Dynamic Topic Model is the best choice to analyse the CORD-19 corpus. A novel topic labelling strategy is proposed for dynamic topic models to analyse the evolution of latent concepts. It incorporates content changes in both the annual evolution of the corpus and the monthly evolution of the COVID-19 disease. The generated labels are manually validated using two approaches: through the most relevant documents on the topic and through the documents that share the most semantically similar label topics. The labelling enables the interpretation of topics. The novel method for dynamic topic labelling fits the content of each topic and supports the semantics of the topics.

Keywords Dynamic topic model · Topic labelling · Topic interpretability · CORD-19

1 Introduction

The health crisis caused by SARS-CoV-2 has marked a milestone in human history. The management of this crisis brings with it a multitude of challenges and difficulties. One of these is the management of the large volume of information (e.g. discarded drugs, new treatments, effective preventive actions, etc.) generated due to the vast amount of research that was conducted in such a short period of time and that is still ongoing. Scientific publications were growing on a daily basis, and the information varied over time. Many of these publications have been openly collected in different data sets for further exploitation, for example, COVID-19 Open Research Dataset (CORD-19), the literature hub for tracking up-to-date scientific information about the 2019 novel Coronavirus (Lit-Covid), etc. But for health practitioners and researchers, it is

very difficult to extract knowledge from such corpus due to the high volume and large size of textual data.

One of the main techniques that can be used to analyse the content of a collection of texts in an unsupervised way is using probabilistic topic models [1]. These techniques enable the extraction of the inherent topics in a corpus. A topic is defined as a distribution of words that are related and represent a specific concept. To identify these words and discover topics, topic modelling techniques organise words that frequently appear together in the document collection. Furthermore, the words of the topic have different weights depending on the importance of the term within the topic. An example of what the topics look like is given in Table 1.

Some domains of application of probabilistic topic models are social network studies [2], news [3], political and policy texts [4], software development [5], and biomedical informatics [6], among others. For this type of specialised domain, knowledge is highly technical and requires professionals to interpret the results obtained. The main problem with these techniques is the complexity of interpreting the generated topics. Since topics are sets of words, the relationship between them may not be obvious or have no direct relationship. Because of this, many studies interpret the meaning of

✉ Ibai Guillén-Pacho
ibai.guillen@upm.es

¹ Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain

² Computer Science Department, Universidad Politécnica de Madrid, Madrid, Spain

Table 1 Topic examples

Topic	Label	Top words
T_1	Eating disorders	Food, obesity, weight, body, diet, insecurity, dietary, ..
T_2	Mental disorders	Symptom, anxiety, depression, mental, psychological, depressive, stress, ..

topics manually, looking at the most relevant words and documents, and assigning a label to the captured concept (e.g. [2, 5]). This label tends to be one or several words that represent the knowledge captured by words and documents on that topic. Other studies simply show the topic as a bag of the most important N top words (e.g. [4, 6]). An example of labelling is given in Table 1.

One of the main characteristics of working with collections of scientific articles is that each one corresponds to work done on a specific period and published on a specific date. This feature makes the corpus temporal. Therefore, the time factor acquires relevance and may become essential in the development of the topics. In a collection of scientific articles on coronavirus, this temporal aspect may be the way in which the corpus expresses the development of treatments, drugs, vaccines, etc. or the identification of symptoms, mutations, side effects, etc. Techniques for dynamic probabilistic topic modelling allow the extraction of the inherent topics in a corpus considering their evolution over time.

Labelling improves the interpretability of probabilistic topics by providing concepts that describe them [7]. However, capturing the changes that these concepts can undergo over time is a very important challenge. Because topic models are difficult to interpret, labelling is a crucial tool to express the evolution of a large corpus as the one described above [8]. Labels would represent the development of the disease, treatment, or symptoms in a more human interpretable way. In this paper, we study how to capture the evolution of concepts over time from the terms that appear in texts, in a particular domain and for a collection of documents.

In our approach, we use dynamic topic modelling techniques to analyse and understand the content of research articles associated with the SARS-CoV-2 pandemic. We want to extract topics according to their temporal evolution and facilitate the interpretation of the generated topics to understand the evolution of key concepts in the pandemic. For this purpose, we want to accomplish two things: make it easier to interpret the concept that captures each topic; and facilitate the interpretation of the temporal evolution of the concepts inherent in a collection of scientific articles. To achieve this, we develop a novel method to faithfully construct labels that express the evolution of each topic. With this method, we build dynamic labels that change over time (as does the topic

itself). In addition, we have validated the labels through the most relevant documents on their topic.

In summary, the main contributions of our work are as follows:

- A set of dynamic topic models generated from the CORD-19 corpus. To carry out this study, we created different dynamic topic models for research articles associated with the SARS-CoV-2 pandemic. The models are available at Zenodo [9].
- A novel unsupervised technique to create labels for dynamic topics. We present a method to generate labels based on the most relevant words that appear in each topic. The resulting labels of the models analysed are available at Zenodo [10].
- A methodology to validate the dynamic topic labels. We reviewed the most important documents for each topic and analysed the relationship with the label. Additionally, we represent dynamic topics vectorially based on their semantics, and we analyse the shared documents between nearest topics. This semantic representation can be used by different models with different configurations to group topics by semantic affinity or to find the most semantically similar topics.

2 Background

2.1 Probabilistic topics

A probabilistic topic is a concept or idea extracted from a corpus of documents. The same document can talk about several topics. In practice, a topic consists of two probability distributions that are generated during training, one on the documents (topic-document) and one on the vocabulary (term-topic). On the one hand, the distribution over the documents specifies the probability with which this concept is mentioned in each document of the corpus. However, the distribution over the vocabulary provides the semantic definition of the topic. These topics are considered static topics.

A dynamic topic is the same as a static topic, but assumes that the concept or idea evolves over time. This evolution is manifested in the probability distribution over the vocabulary (not over the documents). Each word may have a different probability depending on the date when the topic is seen. This represents its semantic evolution over time, but it is still the same topic. In both dynamic and static topics, the probability distributions define the importance of the elements for the topic. The higher the probability of a document in the topic-document matrix, or of a word in the term-topic matrix, the higher its relevance. Moreover, in the case of dynamic topics, the probability of words varies over time. Consequently, there is not one most important word for a topic, but one most

important word for a topic at a specific date (called time slice in this domain).

An example of such probability distributions would be the following. On the one hand, for a static topic on “prison sentences”, its topic-document distribution could be: document “Judgment X” 45%, document “Judgment Y” 45%, and document “Prison list” 10%. On the other hand, its term-topic distribution could be: “witness” 45%, “guilty” 45%, and “guard” 10%. If the topic was dynamic, the topic-document distribution would remain the same, and changes would be seen in the term-topic distribution. In this case, the term-topic distribution would be divided by time slices. In the 2022 time slice, the distribution could be: “innocent” 45%, “acquitted” 30%, and “witness” 25%, while for the year 2023 it could be: “witness” 65%, “guilty” 30%, and “convicted” 5%.

2.2 Embeddings

One of the most important ways to represent words in a computable form is through embeddings. Some topic modelling techniques employ them. These are vectors that distribute the vocabulary of a collection of texts in a multidimensional space (known in the literature as a semantic space). In this way, the distances between words will represent their semantic affinity; the closer, the more similar and the farther the more different. There are two main approaches to generate these embeddings: static and dynamic. On the one hand, the static approach generates only one vector for each word, which causes problems with polysemous words. On the other hand, the dynamic approach generates a vector for the context of each word, resulting in several vectors per word depending on the sentence in which it occurs.

One of the most widely used techniques for generating static embeddings is the continuous bag-of-words model (CBOW) and Continuous Skip-gram Model architectures [11], both implemented in the word2vec software tool (usually the two architectures are referred to by word2vec). These architectures are inverse, but both approaches use shallow neural networks to learn word embeddings. On the one hand, CBOW predicts the word corresponding to a context of neighbouring words, both preceding and following the word to be found. On the other hand, Skip-gram predicts the context of neighbouring words for a particular word. Furthermore, these models have been upgraded [12] to improve efficiency, troubleshoot problems, and eliminate frequent words from the training phase of training work.

Among the most commonly used techniques to generate dynamic word embeddings is the Bidirectional Encoder Representations from Transformers (BERT) language representation model [13], based on the transformer architecture [14]. This technique generates embeddings depending on the word and the context in which it appears, so that the same word can have many different embeddings depending on the

sentence in which it appears. But, as pointed out by [15], BERT suffers from under-training problems and presented several modifications to improve the pretraining process, calling the improved model the Robustly Optimized BERT Approach (RoBERTa). However, these models have been outperformed by several architectures, such as [16]. This work improves the BERT and RoBERTa models by modifying the way the words were represented and by adding absolute positions in the decoding layer. The authors call this architecture decoding-enhanced BERT with disentangled attention (DeBERTa). One of the main drawbacks of these BERT-based architectures is the high computational overhead when working with complete sentences, which made it impossible to tackle the task in a reasonable amount of time. To solve this, [17] suggested Sentence-BERT (SBERT). This modification of BERT includes Siamese and triplet networks capable of constructing a semantic space of sentences where similar sentences are close.

2.3 Topic models

Topic models allow for the creation of representations of large collections of texts. These techniques abstract the content of documents and extract the main points covered. In this section, we examine three of these techniques: Dynamic Topic Model (DTM) [18], Dynamic Embedded Topic Model (DETM) [19], and BERTopic [20].

One of the most important techniques is latent Dirichlet allocation (LDA) [21], the Bayesian version of probabilistic latent semantic analysis [22]. This technique follows a probabilistic generative approach and was the first topic modelling technique to use the Bayesian statistical paradigm. In LDA, the topic-document probability distribution is defined as a Dirichlet distribution and the term-topic distribution as a multinomial one. When modelling the topic-document probability distribution, each document is defined as a Dirichlet distribution with as many dimensions as topics are specified a priori. The value of the distribution in each dimension (topic) is the probability that this topic is present in that document.

Due to the power and performance of the LDA algorithm, it suffered several modifications to address different weaknesses. One of the modifications is the Dynamic Topic Model (DTM) [18]. This model captures the evolution of topics over time. DTM assumes that the data is divided by discrete time slices and that the topics of a time slice are an evolution of the previous one. This means that DTM applies LDA to the documents of each time slice, having as many LDA models as time slices. However, to represent the evolution of the topics, DTM models the distributions of topics for a date using a Gaussian distribution of the previous date one. This allows transferring the topics generated by each LDA model to the model of the next time slice. So, this can be summarised as DTM is a set of LDA models (as many as there are discrete

time spaces) connected by a temporal dynamic. However, one of the main drawbacks of this approach is that it requires a lot of computational resources to generate the models. This algorithm has been used many times on large text collections [23–25].

The main drawback of LDA, and therefore of DTM as well, is dealing with large vocabularies and taking into account the semantics of words. Moreover, they do not allow for working with words outside the vocabulary. To overcome these shortcomings, the Embedded Topic Model (ETM) [26] is a non-probabilistic generative technique based on neural networks. The main goal of this technique is to link LDA to word embeddings. It uses vectors for words instead of the traditional bag-of-words approach. While in LDA each topic is a distribution over the whole vocabulary, in ETM each topic is a vector. ETM has two vector spaces (semantic spaces): one in which the topics are developed and one in which the words are found. The distribution of topics over the terms is proportional to the scalar product of the word vectors and the topic vector, that is, the distance between the words and the topic.

Like LDA, ETM also has a modification to model the evolution of topics over time. This modification, called the Dynamic Embedded Topic Model (DETM), was developed by [19]. Instead of generating the vectors for the topics over the complete set of documents, it is done on the documents of each time slice. In the same way as in DTM, DETM creates the topics (vectors) for each time slice as an evolution of those of the previous time slice.

A very different approach is proposed by [20]. This author presented BERTopic. This model addresses the task of modelling topics using a classification approach based on the semantics of the documents. It uses SBERT [17] to employ pretrained language models. With these, it generates a vector representation of each document based on its content. These vectors are used by the dimensionality reduction algorithm and then by the density clustering algorithm to generate the topics. In this context, a topic is a group of documents created by the clustering algorithm. They are related because of their closeness, i.e. semantic affinity. The group of words that defines each topic is calculated by extracting all the words of the documents in the topic as if they were a bag of words, that is, as if they belonged to a single document; then, the most relevant words of this topic are calculated; and finally, the top words are taken as the topic. The main distinction BERTopic presents for modelling the evolution of a topic is chronology. BERTopic, unlike the others, generates topics for the latest date and modifies the words that define the topic according to the date to be observed. This representation is less detailed than the other methods, but the use of pretrained language models allows the topics to have a high semantic quality. This model has been used in other studies for big collections of texts [27].

However, apart from the individual drawbacks of each model, they all have a common one. The representation of topics is not addressed. Models that do not address temporal evolution use a bag of N most important words. However, with the time factor, this approach becomes more difficult because experts have to analyse a bag of words for each topic and for each date. Therefore, we want to improve the interpretation of these models by reducing the number of words to be analysed through the generation of labels. Because we assume that the fewer elements an expert has to analyse, the fewer difficulties it will have. These labels will also capture the temporal evolution of the topic to represent its evolution.

3 State of the art

The interpretation of a static topic involves difficulties and a high cognitive load in humans [8]. An effective way to reduce it is to generate representations of topics [8], commonly the top 10 terms of the topic [28]. However, this approach has been shown to cause difficulties for users in interpreting the topics [7, 29]. Other more recent approaches represented topics with labels, which could have the appearance of a word [28] (e.g. “finance”), n -grams [30] (e.g. “vaccinated-family”), images [7], phrases [28, 31, 32] (e.g. “production licence”), etc. If the interpretation of a static topic entailed difficulties, dynamic topics entail even more. Apart from the problem of interpreting the topic, it is also necessary to interpret its evolution. A dynamic topic changes and evolves over time. If a dynamic topic should be represented with a label, the label should change and evolve with the topic. In this way, the label would capture more accurately the content and meaning of the topic at each date. Such labels, following the nomenclature used in topic models, would be dynamic labels.

The databases Scopus¹ and WOS² were checked for techniques used for this type of topics, but nothing was found. In Scopus we carried out the query: (TITLE(dynamic*) AND TITLE(topic*) AND (TITLE(label*) OR TITLE(tag*))), while in WOS we performed the query: TI=(dynamic*) AND TI=(topic*) AND (TI=(label*) OR TI=(tag*)). These queries, on 26 October 2023, only return three different papers, and none of them labels dynamic topics or generates dynamic labels. This demonstrates that there is a gap in the literature that has not yet been filled. However, when searching the same databases for static topic labels, we did find results. Performing the query (TITLE(label*) OR TITLE(tag*)) AND TITLE(topic*) AND TITLE(model*) in the Scopus database we found more than 160 articles on 26 Octo-

¹ <https://www.scopus.com>.

² <https://webofscience.com>.

ber 2023. Because there were too many papers, the search was refined to look at the most recent papers written in English. The query that was carried out, on the same date, was: (TITLE(label*) OR TITLE(tag*)) AND TITLE(topic*) AND TITLE(model*) AND PUBYEAR < 2024 AND PUBYEAR > 2019 AND (LIMIT-TO(LANGUAGE, "English")).

The query returned 40 articles, of which 38 were accessible from our institution or provided by the authors. Among these, 7 were excluded from a preliminary review (focussing on analysing the title and abstract) as they did not pertain to the topic of topic modelling. The remaining 31 articles underwent a comprehensive assessment (completed article analysis). Twenty-two of them were eliminated because they did not present topic labelling techniques. Consequently, we identified nine articles that discussed topic labelling techniques published in the last 4 years. The various techniques were categorised on the basis of several criteria. In terms of content within the papers, we considered factors such as the labels' generation, the number of candidate labels generated before selecting the one corresponding to each topic, and whether the labels were extracted from internal sources (the documents utilised during the topic models' training) or external sources. Subsequently, the articles were organised based on the following identified patterns: their dependence or independence on the utilised topic model, whether the technique was supervised or unsupervised, and the method employed for label selection.

The types of methods identified were manual [33] and automatic [8, 32, 34–39]. Manual work was simple; they manually labelled 18 topics generated by an LDA model based on the top words of each topic [33]. The automatic ones could be of several types: distance-based [32, 34–36], semantic source-based [37–39], and neural [8]. Of these, half used an unsupervised approach [8, 32, 35, 36], and the other half used a semi-supervised approach [34, 37–39]. Three techniques depended on the chosen topic modelling algorithm and, in addition, generated only one label [35–37]. The rest were independent of the algorithm used to generate the topics and generated several candidate labels, which were then evaluated to choose the best [8, 32, 34, 38, 39]. When generating labels, three methods extracted them from the same documents with which the topic models had been trained [8, 35, 36], four from external sources [32, 37–39], and one both [34]. Most articles evaluated the performance of topic models and not labels [34–39]. Those who evaluated labels did so with a user study [32] or automatically using different metrics [8]. The use of metrics was a more objective approach than through users, allowing for more impartial benchmarking. In [8], they evaluated labels according to three metrics: relevance, coverage, and discrimination. Relevance measured the similarity between the label and the topic, coverage measured how many of the top 20 words of the topic appeared

on the label, and discrimination measured the differences between labels.

Distance-based Methods

Distance-based techniques encompass approaches that explore label retrieval by means of similarity functions applied to probability distributions and vectors. The methodologies examined encompass cosine similarity, Kullback–Leibler divergence, and TextRank.

Cosine Similarity-based Distance Some publications suggested using cosine similarity to give each topic the best label [35, 36]. Cosine similarity acts as a measure to evaluate the proximity between embeddings. The proximity of embeddings represents the relationship between labels and topics, the closer the more related. In [36], they generated sentences as a label. For that, they created a topic model that uses sentences as an atomic unit instead of words. They generated embeddings for sentences with the BERT model [13] and then applied clustering. Clusters are topics made up of sentences, and their centroid is the embedding of the topic. The label is assigned by extracting the sentence whose embedding has the highest cosine similarity to the embedding of the topic. In [35], they used n-grams to label the software. These authors extract source code and project information from GitHub projects (code, README, keywords, etc.) and use them as “document” for topic modelling. Within the extracted information are the software tags, which are the categories and topics that were manually assigned to the project by its developers (e.g. “deep learning”). These software tags are used as candidate labels. The procedure involves three important steps. First, the authors create embeddings for the candidate labels derived from the project-by-word matrix. Second, they generate topic embeddings with topic-by-project matrix vectors. Finally, the label of each topic is determined as the software label that has the greatest cosine similarity to the centroid of the topic. These techniques are dependent because one requires a specific topic modelling algorithm [36] and the other requires a previously labelled corpus [35].

Kullback–Leibler Divergence-based Distance The Kullback–Leibler (KL) divergence serves as a metric that quantifies the similarity between two probability distributions by assessing the information loss incurred when one distribution approximates the other. [34] proposed a topic labelling method for multilabel document classification with phrases. Candidate labels are derived from preprocessed textual data in the corpus and undergo manual validation. To determine the most appropriate labels for each topic, the authors employ a scoring mechanism. This function assigns a score to each candidate based on its KL divergence similarity for each topic regarding the similarity to the other topics. The three most ranked phrase candidates are considered labels. This technique is semi-supervised because it required human supervision.

TextRank TextRank [40] is a technique that determines the importance of text elements using graph-based distance measurements. It is used by [32] to discern the best labels from a candidate pool generated using two different methods: Explicit Semantic Analysis (ESA) and Yandex. ESA, which represents text content through a vector scheme based on TF-IDF, is used to convert Wikipedia article titles into embeddings. The topic embeddings are generated by extracting the distributions of the top 10 words of each. Then, ESA is queried to obtain a list of related titles based on the cosine distance from the topic embedding. These titles are then used as candidate labels for the ESA method. The second method uses Yandex. Yandex is a search engine that is used to search for document titles. For each topic, the top ten words are selected, and Yandex is asked separately for each set. The results show 30 title documents retrieved by the search engine for each topic. These titles are preprocessed and taken as candidate labels for the Yandex method. The TextRank algorithm is then modified so that it works with phrases instead of just one word and is applied to each method separately. The three phrase candidates with the highest value for each method are taken as label.

Semantic Source Methods

Multiple methods used semantic knowledge to label probabilistic topics. Some of these methods integrated knowledge graphs as external resources [37, 38], while others chose ontologies [39]. These strategies are semi-supervised because they deal with labelled data through unsupervised exploration.

Knowledge Graph-based The knowledge graphs consulted by [37, 38] are Wordnet [41] and the Microsoft Concept Graph (MCG) [42], respectively. In [38], the label was a word, with description and examples, while in [37] it was an n-gram. On the one hand, in [38], the semantic definition of the most relevant words for the topics was searched on Wordnet. Its definition was preprocessed, and candidate labels were produced by comprising noun phrases, bi-grams, and tri-grams. Following this, the candidate labels were ranked according to their WordNet similarity [43], with the label that had the highest value chosen. To enhance interpretability, the WordNet description of the selected label is subsequently incorporated, along with an illustrative example. On the other hand, in [37], a hierarchical topic model is constructed with a label hierarchy. The bag of words representing the topics were filtered to eliminate noisy elements, specifically semantically unrelated words. Subsequently, the system queried the Microsoft Concept Graph (MCG) to identify the concepts shared by these words. The method proceeded to construct a label hierarchy in the n-gram format, with the concepts shared by the words serving as labels (for example, “meat”). In cases where multiple shared concepts are identified (e.g. “meat” and “vegetables”), an effort is made to ascertain a common concept among them (e.g.

“food”). Alternatively, if no common concept is found, an n-gram label is formed (e.g. “meat-vegetables”). Although [38] is independent of the topic modelling algorithm, [37] requires a hierarchical topic modelling algorithm.

Ontology-based Onto_TML [39] used an ontology to label topics. The process involved extracting the most important terms from each topic and cross-referencing them with an ontology to determine their presence. If a term is found to exist in the ontology, its ancestor is recorded. The ontology used by the authors clearly defines a hierarchy of concepts. In this context, the ancestor is the parent class of the top word. The label selection procedure entails identifying the most frequently occurring ancestor among the top words of a given topic and subsequently utilising this particular ancestor as the label. The label can be a word or phrase depending on the ancestor.

Neural-based Methods

Neural-based methods used artificial neural networks and deep learning techniques to automatically assign labels or tags to topics. In [8], a two-phase BERT-based [13] topic labelling model with paired attention (TLPA) is proposed. TLPA is an attention mechanism that concentrates on two related sequences or data sets. To create topic embeddings, the top 500 words of each topic were taken into account. Then, all sentences in the corpus were filtered and weighted based on the Kullback–Leibler (KL) divergence between the topic embedding and the word distribution in each sentence. The 500 most similar sentences from each topic, called “documents”, were chosen as potential labels. These documents were then fed into the BERT model, which generated sets of sentence embeddings. Sentence embeddings, along with topic-related features, were given to a summarisation layer to select the most suitable candidate sentence for each topic.

4 Methodology

In this work, we faced two types of difficulties: technical and theoretical. On the one hand, working on such a large corpus has required addressing many technical challenges, since data processing and model generation are computationally heavy tasks. On the other hand, defining, creating, and validating a technique to label topics that can take into account the topics temporal evolution have been the main theoretical challenge. This section defines the terms used in the article and explains the decisions taken to process the texts, generate the models, label the topics, and validate the labels. The results of these steps are explained in the 5 evaluation section.

4.1 Definitions

In this section, we provide some relevant definitions that will be used in the rest of the paper.

Definition 1 (Topic word relevance) The relevance of a word is a measure used by topic modelling algorithms to rank words from most to least important for each topic. In dynamic topic models, topics consist of a group of words that change in relevance within the topic over time. Each model uses different algorithms to calculate the relevance between words and topics.

Definition 2 (Topic top words) The top N words of a topic are the N most relevant.

Definition 3 (Cross-domain document) A cross-domain document is a document that discusses several concepts and not exclusively one topic.

Definition 4 (Important words) The important words of a topic are the top 5 most relevant words. A word is important if it is within this group.

Definition 5 (Core word) A core word is a word that belongs among the five most relevant of the topic all dates.

Definition 6 (Trend word) A trend is a word that is important from a given moment and maintains its importance over time.

Definition 7 (Fashion word) A fashion is a word that is important at a given moment and then stops being important.

Definition 8 (Fleeting topic) A fleeting topic is one whose top five words are always fashion words.

Definition 9 (Label correspondence) The correspondence of a label with a topic is the correlation it has with the most relevant words and documents on that topic. We analysed the correspondence of the labels with the topics manually. We look for the appearance of the label in the abstracts of the most relevant documents on the topic.

Definition 10 (Topic coherence) A coherent topic requires that the different top words be related and mutually sustained. Therefore, the words cannot be contradictory and must belong to the same group [44–47].

Definition 11 (Topic interpretability) Interpretability is the human ability to understand the relationship between a group of words and the general concept to which they belong [48]. A topic is interpretable if its top five most relevant words belong to the same concept.

Definition 12 (Topic specificity) A topic-specific model removes ambiguities and provides a detailed representation of the corpus. A model that is over-specific may introduce noise and lose abstraction capacity [49].

Definition 13 (Topic label) A label is a statement that represents the semantic content of the topic [50]. It must be understandable to the user, capture the meaning of the topic, and distinguish it from others. The label can consist of one or more words.

4.2 Data and data preprocessing

Document corpus We work with the English-written abstracts of the COVID-19 Open Research Dataset, CORD-19 [51]. CORD-19 contains a mixture of articles and preprints on COVID-19 and similar coronaviruses such as SARS and MERS. From the 401,214 papers in the corpus, only 278,788 have a processable abstract³ and of those, 8,326 are not written in English. For those written in English and processable, 22,989 have been deleted because they do not have any metadata available. In addition, there are papers in the corpus with publication dates later than the corpus version. Given the distortion that these articles can generate in the models, it has been decided not to take them into account. (They are 116 papers.) Thus, only previously published articles are used.

Data preprocessing Data preprocessing is one of the most critical processes for generating topic models and working with them. It is important to remove as much noise as possible while keeping the relevant terms. The more the words we remove, the faster and more efficient the models will be, but more detail and specificity will be lost. For this reason, inspired by the guidelines given by [52], we have followed the following steps to process the texts.

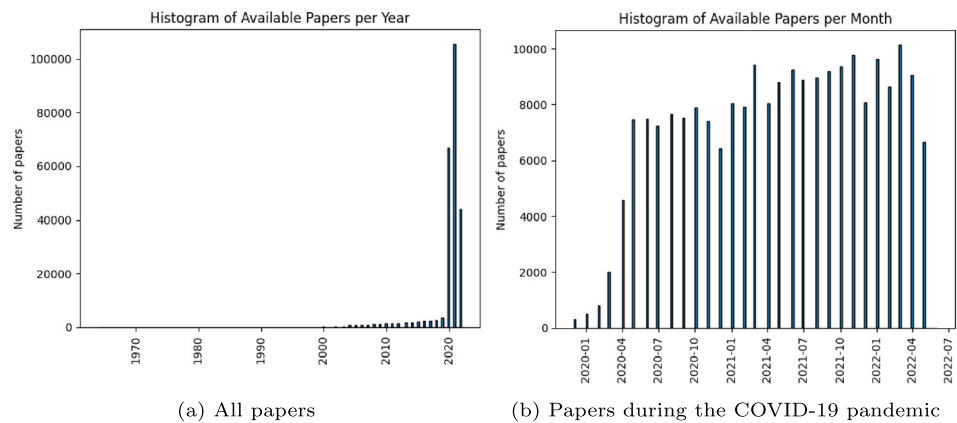
We lower-cased the text to remove ambiguities caused by the different ways of typing the words. We remove all stop words⁴ and punctuation characters present in the texts. Since a large vocabulary increases the resources needed for topic modelling, we reduced the vocabulary size. We remove irrelevant terms, acronyms, and abbreviations. The irrelevant terms do not add any value to the topics, and the acronyms and abbreviations are difficult to interpret outside the context. To do this, we decided to eliminate all words with less than four letters and to prune the vocabulary according to term frequency. This decision is based on the following three assumptions. First, the most difficult acronyms and abbreviations to interpret are those with fewer than four letters. Second, due to the low frequency of some words, they will be very specific and not relevant to the topics. And finally, due to the high frequency of some terms, they will be very general and will not provide any value to the topics.

Pruning was carried out by the following steps. First, the occurrences of vocabulary words were counted. Then two pruning thresholds were established; upper and lower. The

³ A processable abstract is one that is at least five words long.

⁴ <https://github.com/stopwords-iso/stopwords-en>.

Fig. 1 Date distribution of the usable papers



upper one is used to eliminate words that appear above a certain number of words. The lower one is used to eliminate words that appear less than a certain number. For pruning, relative percentages of the frequency of a term in relation to the volume of all the words in the corpus have been used. The upper threshold has been established at 80% and the lower threshold at 0.1%.

After the previous steps, we lemmatise the remaining words. We decided to use lemmatisation instead of stemming because the resulting words are easier to interpret. Since we removed many words from the texts, many texts are either wordless or contain very few words because these texts were full of noise (stopwords, rare words in the corpus, acronyms, etc.). This creates several problems. First, in topic modelling, it is not possible to work with empty documents. Second, it is necessary to have a minimum number of words in the texts to extract the topic to which it belongs. Since these are abstracts and the preprocessing has been rigorous and intensive, we have established that documents with fewer than five words (after being processed) will not be taken into account.

The resulting amount is that we have 232,483 articles in total of which 204,902 belong to the period from December 1, 2019, to June 2, 2022 (dates on which SARS-CoV-2 became a pandemic⁵). Figure 1a shows the distribution of articles by year, 88% of the usable articles belong to the period 2020–2022. Therefore, this corpus, due to the large amount of information generated from SARS-CoV-2, is biased. The monthly distribution of these articles is shown in Fig. 1b.

4.3 Model construction and evaluation

Model Creation

From the existing algorithms in the literature, we have taken DTM [18], DETM [19], and BERTopic [20] as three possible alternatives for the generation of dynamic topic

models. Although abstracts are short texts, we have found applications of DTM [53–55] and BERTopic [56, 57] for this type of documents. For the DETM model, we have not found works that exclusively use abstracts, but we have implemented it to compare its performance in this scenario. For DETM, we have tested two different embedding configurations: one with the word2vec [11] embeddings and the other with DeBERTa [16], so as to observe the differences between generating embeddings based on co-occurrence of terms and semantic correlation. In the original DETM [19] work, word2vec is used to generate embeddings. However, in our review of the literature, we have found no evidence of the use of DeBERTa in conjunction with DETM. Therefore, to the best of our knowledge, it is a novel approach.

Model Evaluation

We are particularly interested in assessing the quality of probabilistic topics. In the literature, the quality of topics is measured through the diversity and coherence values of the top words [19, 20, 26]. We will use the top 15 words to calculate these metrics. Because dynamic topics evolve over time, we evaluate the quality in all time slices. In this way, we will analyse how topics created using dynamic topic models vary in diversity and coherence according to the algorithm used (DTM [18], DETM [19], and BERTopic [20]). If they vary too much, the topics would be unstable, and their evolution would not be very representative.

Diversity Diversity is the percentage of unique words in the top N words of the topics. A value close to 0 indicates redundant topics, with many words repeated among them. Although a value close to 1 indicates singular topics, with many unique words.

Coherence

Abstracts are limited in length and summarise the most important information in the documents. After preprocessing, the information presented becomes more qualitative, and the word relationships are more relevant. Two different metrics will be used to measure coherence: UMass [58], and C_V [59]. Both measure the co-occurrence of all top word pairs

⁵ <https://www.who.int/es/news/item/27-04-2020-who-timeline---covid-19>.

of the topics in the documents. However, each one follows a different approach. UMass calculates coherence based on the direct relationship of words, whereas C_V is based on indirect relationship of words.

On the one hand, the direct relationship implies that both words coexist in a given space (e.g. a sentence, a paragraph, a document, etc.). On the other hand, the indirect relation implies that although the words do not coexist in the same space, they do coexist with the same words. For example, there are two famous car brands w_i and w_j that do not often coexist in any space. However, they frequently appear together with words like “wheel”, “parking”, “petrol”, etc. Because of this, w_i and w_j have a low direct relationship with each other, but a high relationship with the previous examples. However, since they are related to the same words, they have a high indirect relation.

Under these assumptions, UMass [58] is the metric used to measure the direct relationship. For each pair of top words, UMass weights the coherence based on the probability of the occurrence of both words in a document, knowing that at least one word is present in the document (Eq. 1). Given a pair of words w_i and w_j and, for a vocabulary of size N , $D(w_j)$ is the document frequency of w_j (i.e. the number of documents in which it is present). $D(w_i, w_j)$ is the number of documents that contain one or more times w_i and at least one w_j . The higher the probability that w_i occurs when w_j exists in the document, the higher the quality of the topic. Therefore, we measure coherence as coexistence in the documents.

$$C_{UMass} = \sum_{i=2}^N \sum_{j=1}^{i-1} \log \frac{D(w_i, w_j) + 1}{D(w_j)} \quad (1)$$

To evaluate the indirect relationship, the metric C_V [59] is used. It uses different techniques to evaluate topics. The first are the context windows. These windows extract words that are within a distance of ± 55 words in the texts from all the top words of the topics. Then, the context vector of each top word is constructed. This vector is constructed based on the relationship the word has with the rest of the vocabulary in the context windows. To calculate the relationship, C_V uses the normalised pointwise mutual information (NPMI) metric [60]. This metric is a modification of Pointwise Mutual Information (PMI) [61]. Although PMI uses windows of a different type, it weights the co-occurrence of a pair of top words (w_i and w_j) according to equation 2. $P(w_i)$ and $P(w_j)$ are the probability that w_i and w_j (respectively) are present in the window, while $P(w_i, w_j)$ is the probability that both are present in the window. NPMI, Eq. 3, adds normalisation to the PMI to keep the values between -1 and 1 . In addition, it potentiates the result so that there is a larger difference

between the low and high results.

$$\text{PMI}(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)} \quad (2)$$

$$\text{NPMI}(w_i, w_j)^\gamma = \left(\frac{\text{PMI}(w_i, w_j)}{-\log(P(w_i, w_j))} \right)^\gamma \quad (3)$$

The context vector is constructed according to Eq. 4. For a word w_i , its context vector is the NPMI it has with all words w_j of the vocabulary W . This process is carried out with all the top words from all topics. For the same topic, all its top words are taken, and the cosine similarity between their context vectors is calculated. Once calculated, all distances are aggregated, and the average is the topic coherence.

$$\vec{v}(w_i) = \left\{ \sum_{w_j \in W} \text{NPMI}(w_i, w_j)^\gamma \right\} \quad (4)$$

4.4 Labelling of dynamic topics

We assume that as a topic evolves over time, so does its label. The label must match the semantic content of the topic at any given time to be as descriptive as possible. In addition, it has to reflect the temporal evolution of the topic, which implies that part of the previous time slice label should be on the new label if relevant. For example, a topic A in 2020 has the top 5 words “vaccination”, “effects”, “covid-19”, “children”, and “foetuses”, and is assigned the label “vaccination-children”. In 2021, its top five words are “pregnant”, “mothers”, “children”, “vaccination”, and “effects”, and it is assigned the label “pregnant-vaccination-children”. Dynamic tags allow for more accurate descriptions of topics because they capture the granularity of the moment. On the contrary, if a static label was assigned to the topic, it would have to be more general, e.g. “vaccination”, which would cause a loss of detail. With dynamic labels, an expert could analyse the evolution of a topic without having to pay attention to the probability distribution over the vocabulary for each time slice, reducing difficulties. The problem is that to our knowledge, there is nothing similar in the literature. For a label to be dynamic it must include words that represent the core of the topic (if any) and the trends. There may be some dates when a word on the label stops being important. At that time, it is identified as fashion and disappears from the label.

The approach to generate dynamic labels developed in this paper (Fig. 2), is unsupervised, extracts labels from topics top words (internal source), is dependent on the topic modelling algorithm employed (need a dynamic topic model), generates a unique n-gram as label per topic, and is a distance-based method. Duplicate labels in the same time slice are not accepted because we assume that each topic identifies

different ideas. In accordance with this and to capture the details of each idea, we make each label as specific as possible while preserving the abstraction capacity. To label topics, we take the most important words. A word is important to the topic when it remains among the top five words. Under this assumption, a term that has been the top word in at least one time slice and remains among the top five in the rest will be considered as the core of the topic. Multiple words may form the core of the topic because they satisfy the same requirement. Moreover, a word that does not appear in the top five during all time slices, but that appears in the top five once it has been the top word of the topic, becomes a trend. Otherwise, if the word does not remain in the top five after being the top word, it will become a fashion.

The steps we have followed to generate the labels are specified in the flow chart in Fig. 9 and are as follows. (The image can be seen in a larger size in Annex A.) First, the most important word for each topic in the first-time slice is taken as a label (e.g. T_1 = “cell”). It is then checked for duplicate labels (e.g. $T_1 = T_5$ = “cell”). If so, the second most important word of each topic is added to the label (e.g. T_1 = “cell-lymphocyte” and T_5 = “cell-protein”). And it checks again for duplicate labels. If there are no duplicated labels, the process is completed for this time slice; if not, this step is repeated. For the next time slices, the process is similar, but has a previous step. For each topic in the time slice, we take its previous time slice label (e.g. T_1 = “cell-lymphocyte”). For each word that forms the label, we analyse whether they are still important to the topic. If so, they are kept on the label and, if not, removed from it (assuming that the word “cell” is still important but not “lymphocyte”, T_1 = “cell”). Once all label words are checked, the most important word of the topic for the current time slice is taken (e.g. “response” for T_1), and if it is not yet on the label, it is added (e.g. T_1 = “response-cell”). When this process has been performed for all topics in the time slice, we check whether there are duplicate labels. When there are no duplicates left, it moves on to the next time slice. This process is repeated iteratively for the remaining time slices. The results are a set of labels that vary according to the temporal slice in which they are observed, acquiring and losing detail according to the relevance of those terms in the topics.

Our approach takes into account the amount of words as a measure of the specificity and detail of a topic. This can be seen in EuroVoc,⁶ a thesaurus that explains the contents of EUR-Lex (the online publication service for legislative texts of the European Union). In EuroVoc, the term “transport” encompasses a variety of concepts, such as “transport policy”, which, in turn, includes concepts like “transport under customs control”. When consulting EUR-Lex, “transport under customs control” is more specific than “transport

policy”, which is more specific than “transport”. Another example can be found on the Twitter social network, where the use of a single hashtag is more general than when multiple hashtags are nested. A tweet with the hashtags #Elections #Spain #Independence is more specific than one with just #Elections. Therefore, we assume that the more words on the label, the more specific the topic is. Thus, we assume that the number of words in the description of a topic has a direct effect on its specificity. Even if the labels are very specific, a possible but very difficult case is that two topics might have the same label in different time slices. The above method ensures that there are no duplicate labels between topics in the same time slice, but it does not ensure that there are no duplicates between time slices. For example, the word “vaccination” may appear in topic A in 2017, with its top 3 words being “vaccination”, “measles”, and “children”. However, it may also appear on topic B in 2020, with its top 3 words being “vaccination”, “effects”, and “covid-19”. In these two topics, although the word “vaccination” is the label, the meaning is different. The label is only a representation of the topic that enhances its interpretability and captures its meaning. If there was only one topic for “vaccination”, granularity would be lost and greater abstraction would be achieved, and the problem could be oversimplified.

4.5 Label validation

The last contribution is the automatic label validation. This work aims to contribute to a knowledge graph generated in the work [62]. Labels will be introduced into the knowledge graph to make expert information searches more accurate and concrete. We are interested in checking that the labels correctly represent the topics and documents rather than their quality. The way of automatically evaluating the labels used in [8] would not be applicable to our labelling technique because the labels are formed from the top words of the topics. For example, coverage will always be complete. (The label will always be found in the top N words.) Instead of evaluating, we have decided to validate our labels taking into account the other two approaches used in this work. To check the relevance, we checked whether the label is related to the content of the topic. To assess discrimination, we checked whether the relationship between topics supports the semantic relationship between labels. To find out whether the label is related to the content of the topic for a date, the method is simple. The most relevant documents of the topic are taken for the date of the dynamic label, and it is automatically checked that the words of the label appear in the abstract. In this paper, we have also manually analysed these documents to justify the relationship with the label. To find out whether the topics support the semantic relationship between labels, the process is more complicated but similar. The aim is to find cross-domain documents in order to analyse them in the

⁶ <https://eur-lex.europa.eu/browse/eurovoc.html>.

Fig. 2 Dynamic Topic labelling procedure

same way as with the previous approach. In order to find these documents, we suggest a novel method.

The proposed method consists of using embeddings to generate a vector for each dynamic label. These vectors are distributed over a semantic space in which their distance represents the semantic similarity. This means that the closer two vectors are, the more semantically similar their dynamic labels will be. Similarly, the farther apart they are, the more different they are. A dynamic label has a vector for each date. Labels can be made up of one or more words. If it is composed of only one, the topic vector will be the embedding of that word. On the other hand, if it is formed by more than one, the vector will be the centre of the embeddings of all the words in the label. We will use word2vec model to generate word embeddings.

Once the label vectors have been generated, they are grouped by nearest neighbour. In this way, the semantically

most similar topics are found. When grouped, the relationship is tested by checking the existence of the labels on the cross-domain documents. In this paper, we have also manually analysed these documents to justify the relationship with the label. To define when a document is cross-domain, we define two thresholds. The first is the minimum probability with which it has to belong to each topic (individual threshold). The second is the minimum probability with which it has to belong to both (combined threshold). With respect to the number of topics in the model, we have defined the individual threshold as $\frac{1}{9}$ and the combined threshold as $\frac{1}{3}$. This means that, for a model with 100 topics, the document has to belong at least %11.11 to each and at least %33.33 to both. Documents that satisfy these requirements are considered cross-domain documents.

5 Evaluation

In this section, we explain how the topic models have been generated, which configurations we have used, and their performance. Once we have decided which model is the best for this study based on topic coherence and diversity, we take the topics generated by this model and analyse their evolution using the previously defined labelling strategy. To check how representative the labels are, we analysed the most representative documents of the topic for those dates. Finally, to find the most closely related topics, we will look for the nearest neighbour for each topic vector, and using cross-domain documents, we will analyse whether there is correlation.

5.1 Model performance analysis

In this section, we show the configurations of the models and their performance. To evaluate their performance, we will compare whether the evolution of the topics is well captured or not by calculating the coherences and diversity over several years. As calculating coherence is computationally expensive, it could not be calculated for all dates. We calculated it for the last five years and then five years at a time (see Annex B). We did this because the distribution of documents by year is very irregular (88% of the corpus we used is from the years 2020–2022). To clearly show the performance of the models, we summarised Tables 10, 11 and 12. We selected the dates 2022, 2019 and 2015 to show the coherence and diversity values of the models. Furthermore, we have provided a graph with all the values to support the summary table. The graphs can be seen zoomed in Annex B, Figs. 10, 11, 13, 12, and 14.

In order to understand the values in the tables, it is necessary to take the following into account. For indirect relationship coherence and diversity, the higher the value, the better. However, for direct relationship coherence, the closer the value is to 0, the better. Once we know which model best captures the evolution of the topics in the corpus, we will generate another one with the same configuration only on the dates that SARS-CoV-2 emerged, but with smaller time slices. In this way, we will be able to see through the labels the differences caused by the change in granularity of the topics and will have more detail on the knowledge generated during the period in which SARS-CoV-2 was developed.

5.1.1 DTM

As the number of existing topics in the document collection is uncertain, we have tested configurations with 100, 200, 300, and 400 topics. The model was implemented in *Python* with the package *Gensim* (v3.8.3).⁷ The best empirical results in

Table 2 DTM models performance summary

Topics	Time slices	C_V	UMass	Diversity
100	2022	0.6560	−2.7859	0.5320
	2019	0.6636	−2.7243	0.5307
	2015	0.6609	−2.7192	0.5347
200	2022	0.6323	−3.2246	0.5307
	2019	0.6330	−3.2318	0.5190
	2015	0.6323	−3.2158	0.5183
300	2022	0.5954	−3.5012	0.5127
	2019	0.5954	−3.5013	0.5082
	2015	0.5964	−3.4909	0.5049
400	2022	0.5688	−3.5940	0.4592
	2019	0.5689	−3.5982	0.4577
	2015	0.5683	−3.5973	0.4523

coherence and diversity were obtained by the following configuration for all models: seed 0; minimum iterations of LDA 6; maximum iterations of LDA 20; maximum EM iterations in LDA 10; alpha (α) 0.01; and evolution speed 0.005.

The results, summarised in Table 2 and Fig. 3, show that the coherence over years is maintained with very low variations. The full values are in the Table 10 and the zoomed graph in Fig. 10. On the one hand, one of the main differences between the models is that the 100 and 200 topics models have a clear difference in direct relationship coherence (*UMass*) and in topic diversity but a minimal difference in indirect relationship coherence (C_V). However, the models with 300 and 400 topics have almost no differences in coherence but do have differences in diversity. In addition, coherence (of both indirect and direct relationship) and diversity decrease as the number of topics increases. This means that, at a certain threshold, more topics introduce noise. On the other hand, another main difference is that, while between 100 and 200 topic models the difference is notable, from 200 topics there is not that much change. The best DTM configuration for this corpus is the 100-topic one, as it has the best performance for all the metrics used.

5.1.2 BERTopic

For the implementation of this model, we used the *BERTopic* v0.11.0⁸ package. All the models we created have followed the same configuration. We use the all-MiniLM-L6-v2⁹ sentence transformer to generate the document vectors and let BERTopic automatically define the rest of the parameters. As with the DTM model, the number of topics in the corpus

⁷ <https://github.com/RaRe-Technologies/gensim/releases/tag/3.8.3>.

⁸ <https://github.com/MaartenGr/BERTopic/releases/tag/v0.11.0>.

⁹ <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>.

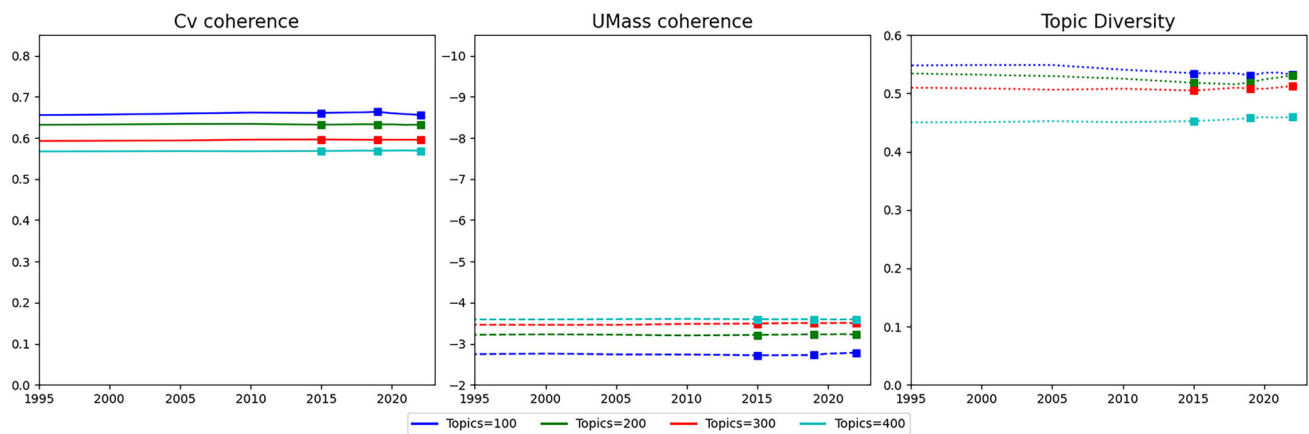


Fig. 3 DTM models performance

is unknown, so we have decided to generate models for the same configurations (100, 200, 300 and 400 topics).

The results we obtained with the BERTopic model, summarised in Table 3 and Fig. 4, are very good in indirect (C_V) and direct ($UMass$) coherence for some dates but are very unstable). The full values are given in Table 11, and the zoomed graph is shown in Fig. 11. Some of the patterns identified in the DTM model topics are also present in the BERTopic model topics. Again, as the number of topics increases, both coherence metrics (indirect and direct one) and the diversity metric decrease. While between the 100 and 200 topic models there are few differences in both coherence metrics, between the 300 and 400 topic models are less evident. Moreover, the topics generated are very similar and barely differ from each other. This is reflected in a low topic diversity over all years.

As noted above, BERTopic is not a specially developed model to handle the evolution of the topic. Unlike other models, BERTopic does not generate a time slice topic as an evolution of previous ones. Instead, BERTopic generates a topic model for the last time slice, from which it tries to generate the topics for the rest. This may be the reason why BERTopic is unable to correctly capture the evolution of topics. This becomes evident when coherence metrics decrease rapidly as the time slice gets older. In addition, the results vary a lot between dates, almost reducing the indirect relationship coherence by half and the direct relationship coherence by nearly three times in some of them. This behaviour shows that BERTopic suffers from problems in capturing the evolution of topics in this corpus. As mentioned above, about 88% of the articles belong to the period 2020–2022, the dates for which BERTopic is able to generate topics with low variance in indirect and direct relationship coherence. However, for the rest of the dates, it does not adapt, and the results are poor. Based on this, we assume that the best BERTopic configuration for this corpus is the 100-topic one.

Table 3 BERTopic models performance

Topics	Time slices	C_V	$UMass$	Diversity
100	2022	0.8068	−2.0985	0.2260
	2019	0.6007	−4.7835	0.2207
	2015	0.5132	−5.7133	0.2140
200	2022	0.8042	−2.2982	0.1967
	2019	0.5080	−6.1321	0.1842
	2015	0.4824	−6.0701	0.1983
300	2022	0.7575	−3.0198	0.1536
	2019	0.4730	−6.3694	0.1648
	2015	0.4851	−5.9573	0.1849
400	2022	0.6917	−3.9798	0.1243
	2019	0.4768	−6.3873	0.1550
	2015	0.4802	−6.0804	0.1725

5.1.3 DETM

An existing implementation of the DETM model¹⁰ has been used. In order to use this model, we need to extract the embeddings for the vocabulary words, and as we mentioned before, we experimented with embeddings generated with the DeBERTa and word2vec models. On the one hand, to generate embeddings with the DeBERTa model, we downloaded the original model¹¹ and retrained it with the corpus documents. Then, to extract the static vectors for each vocabulary word, we drawn all the vectors for each word and took the mean as its vector. On the other hand, to train the word2vec model, we have used the implementation in the *Gensim v4.2* library. The parameters used in this implementation are summarised in Table 4.

¹⁰ <https://github.com/quynhneo/detm>.

¹¹ <https://huggingface.co/microsoft/deberta-base>.

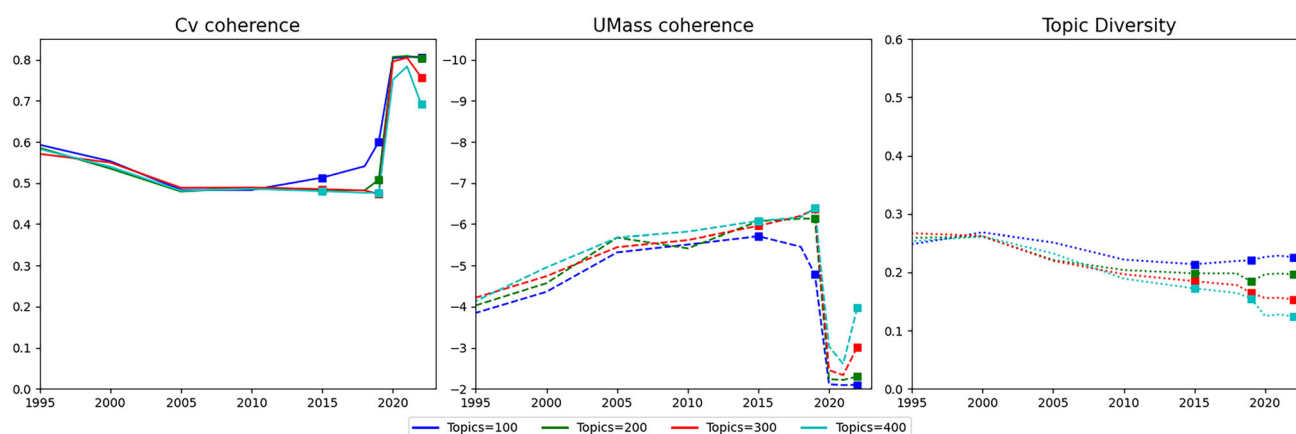


Fig. 4 BERTopic models performance

Table 4 Word2vec embedding model training parameters

Parameter	Value	Parameter	Value
Skip-gram window size	10	Learning rate (α)	0.025
Negative sampling	15	Minimum learning rate (α)	0.0001
Embedding dimensions	300	Seed	1
Epochs	5000		

We only generated models with 100 and 200 topic configurations because we found that 300 and 400 topic configurations do not generate good topics in previous models (DTM and BERTopic). The best empirical results in indirect (C_V) and direct ($UMass$) relationship coherence were obtained with the configuration in Table 6. These results, summarised in Table 5 and Figs. 5 and 6, show that, contrary to what we expected, DeBERTa did not perform well. The full values are in the Table 12 and the zoomed graphs in Figs. 12 and 13. We think this is because we are not using the contextual embedding for each word. After all, language models are designed to create dynamic embeddings depending on the context of the word. The way they are converted to static by using the average may be the cause of their low performance. The DETM model with DeBERTa embeddings achieves half the indirect and direct relationship coherence of the word2vec model. Although the results are not good, we can see that the DETM models generated a stable topic evolution with small changes in indirect relationship coherence metric. However, there is a lot of variation in diversity and direct relationship coherence metrics. The best DETM model obtained has been the one trained with word2vec embeddings because it has the best values all the metrics. Similarly and for the same reason, the best configuration for this model is the 100-topic configuration.

The generated topics cannot rival DTM and BERTopic. The results show that these models achieve higher diversity and coherence than the DETM. As in BERTopic, as the coherence of the indirect relationship decreases, so does the

coherence of the direct relationship. It should be mentioned that in the original work of the DETM model [19], DETM shows coherence scores worse than the DTM model in some cases. This can also be observed in another work in which DETM and DTM models are trained with different topic configurations on a subset of CORD-19 and again DTM performs better [63].

5.1.4 Comparison results

We can summarise the above evaluations in the following way: on the one hand, BERTopic is the model that generates the topics with higher indirect relationship coherence, but is unable to generate a correct stable evolution of the corpus. On the other hand, the DETM model strongly depends on the embeddings generation strategy. However, the evolution of the topics it generates is not as stable as DTM (although not as unstable as BERTopic). Finally, the DTM model seems to be the most stable across all dates, achieving good coherence and diversity. The comparison of the three models is shown in Fig. 7.

Based on this, we generate SARS-CoV-2 related topics with the DTM model following the same configuration. For this, as mentioned above, the time slices are shorter. This time, the granularity is by month and only documents from December 2019 to June 2022 are considered. The resulting topics are shown in Table 7 and Fig. 8. In Table 13, the full performance can be seen, and the zoomed graph is shown in Fig. 14. This model behaves as previously described,

Table 5 DETM models performance

Model	Topics	Time slices	C_V	UMass	Diversity
DETM DeBERTa	100	2022	0.2533	-6.6634	0.0853
		2019	0.2496	-6.9977	0.1867
		2015	0.2507	-7.1097	0.2220
DETM DeBERTa	200	2022	0.2536	-6.9542	0.0903
		2019	0.2552	-7.1930	0.1530
		2015	0.2744	-7.7431	0.2293
DETM Word2Vec	100	2022	0.5525	-3.2132	0.2120
		2019	0.5765	-4.0833	0.4059
		2015	0.5159	-5.0228	0.4667
DETM Word2Vec	200	2022	0.5471	-3.4626	0.1743
		2019	0.5335	-4.5557	0.2853
		2015	0.4713	-5.6803	0.3230

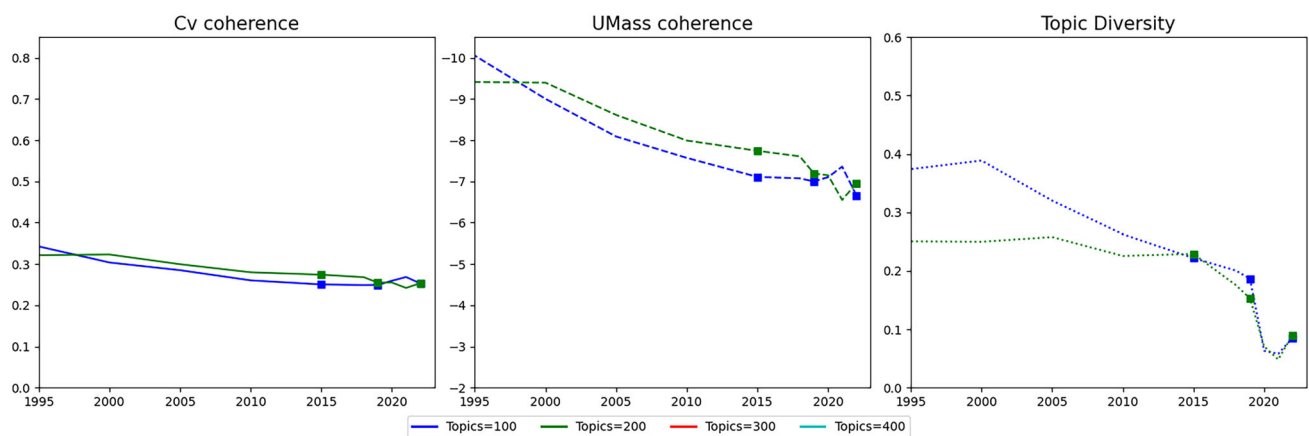
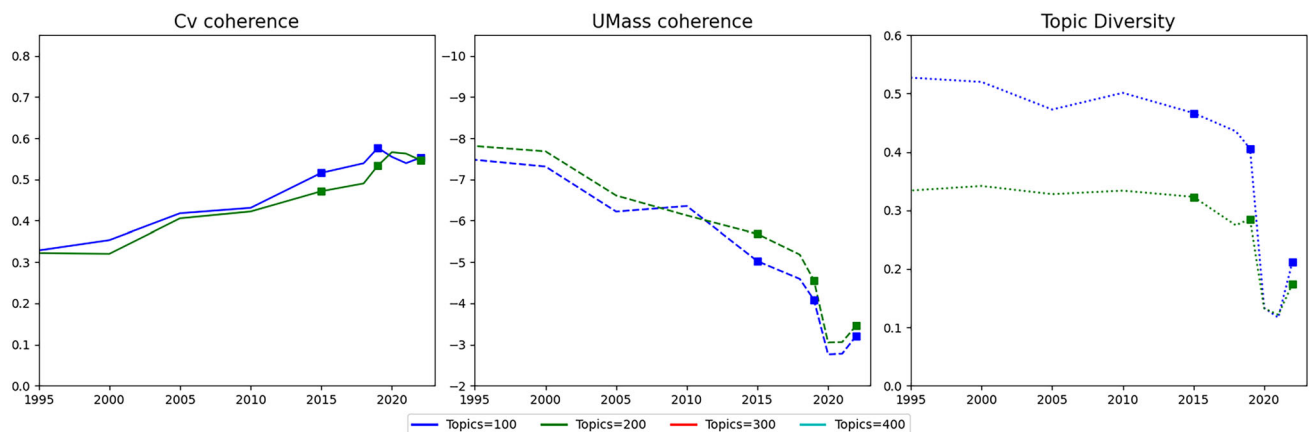
**Fig. 5** DETM DeBERTa models performance**Fig. 6** DETM Word2Vec models performance

Table 6 Parameters used to train the DETM models.

Embedding	Parameter	Value 100 topics	Value 200 topics
Deberta	Anneal the learning rate		False
	Batch size		512
	Dimensions of embeddings		768
	Dimensions of rho		768
	Encoder activation function		RELU
	Encoder dropout		0.1
	Encoder dimensions of hidden space		400
	Epochs		1200
	Eta dropout		0.12
	Eta number of layers		4
	Gradient clipping		2
	L2 regularization		0.012
	Learning rate		0.0015
	Learning rate decay factor		2
	Number of topics	100	200
	Optimizer		ADAM
	Prior variance		0.005
	Used seed	9720883	3306915
Word2vec	Anneal the learning rate		False
	Batch size		512
	Dimensions of embeddings		300
	Dimensions of rho		300
	Encoder activation function		RELU
	Encoder dropout		0.1
	Encoder dimensions of hidden space	400	300
	Epochs		1200
	Eta dropout		0.12
	Eta number of layers		4
	Gradient clipping		2
	L2 regularization		0.012
	Learning rate		0.0015
	Learning rate decay factor		2
	Number of topics	100	200
	Optimizer		ADAM
	Prior variance		0.005
	Used seed	1806363	209764

from 200 topics the indirect relationship coherence (C_V) and the direct relationship coherence ($UMass$) decreases significantly. Also, we show that the evolution of the topics is well captured because the differences between time slices are not notable. Based on the performance of the models in the three metrics, we assume that the best DTM configuration for this document collection is the 100-topic one.

5.2 Dynamic label validation

Our labelling strategy captures the evolution of the inherent topics in the text collection. To show this, we have applied our labelling strategy on the topics generated by the 100-topic DTM model trained on the whole corpus and the 100-topic DTM model trained on the COVID-19 period. Our technique may also be applicable for the other models. Some examples of the resulting labels are given in Tables 8 and 9, respectively. The complete tables can be found in the resources

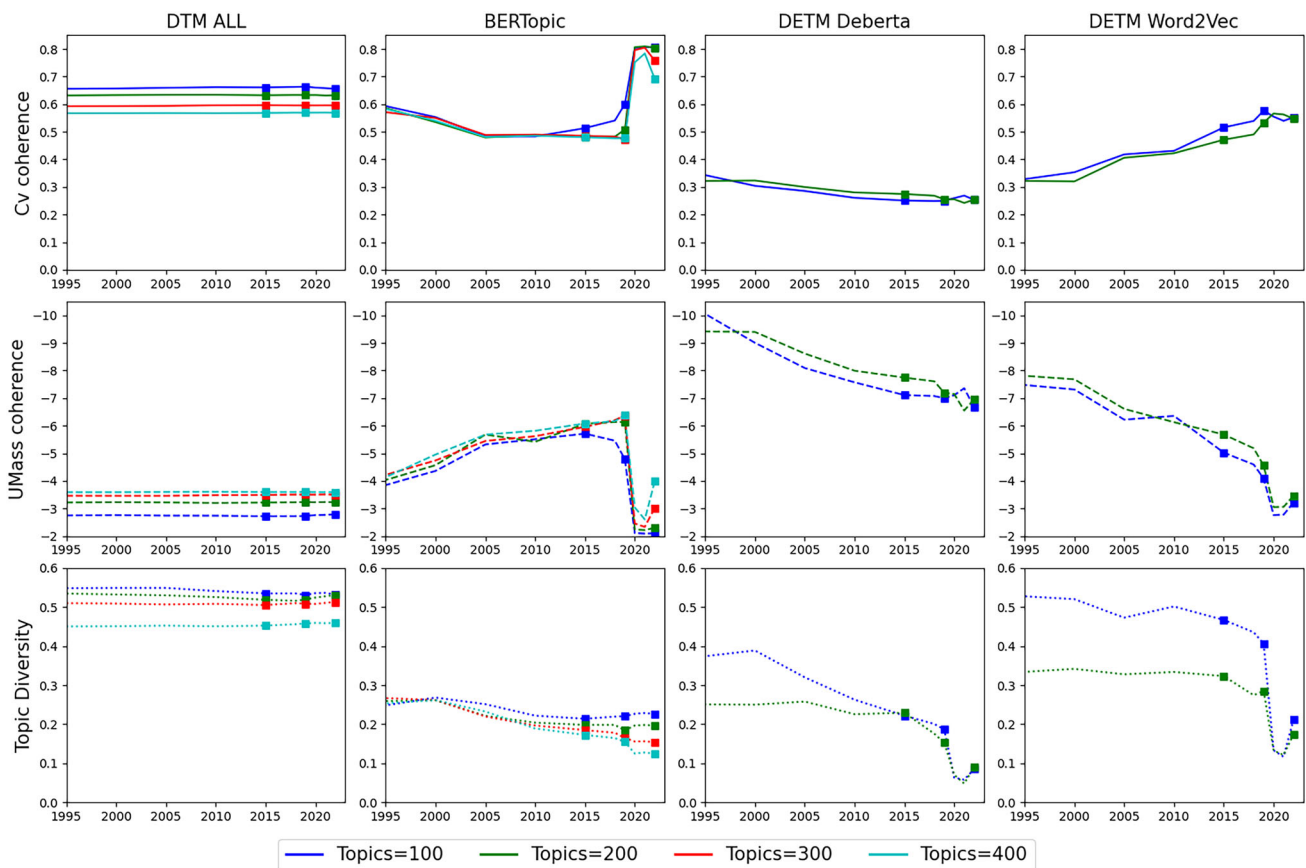


Fig. 7 All models performance

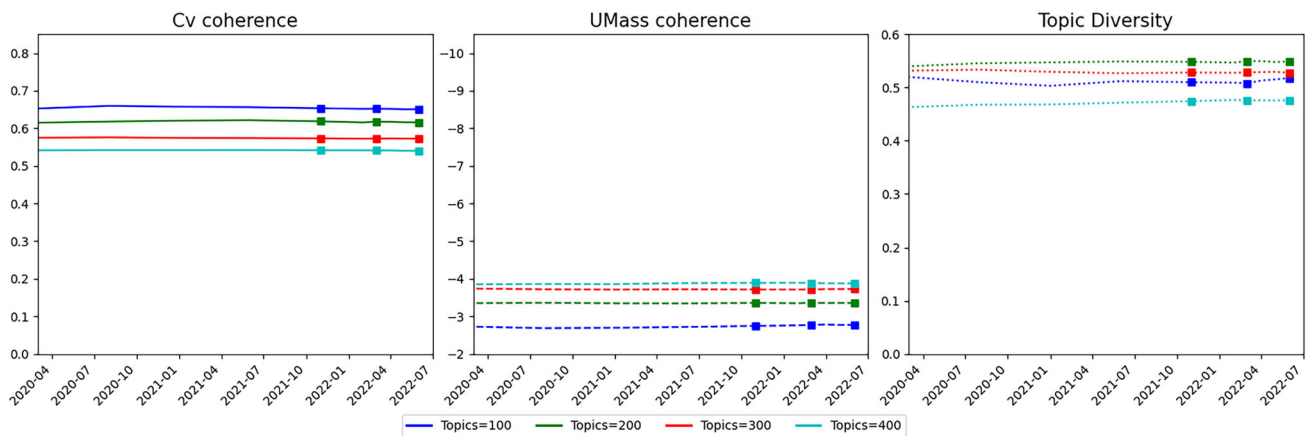


Fig. 8 DTM models performance for SARS-CoV-2 dates

published in Zenodo [10]. The main difference between these two approaches is the level of specificity. As given in the tables, the granularity directly affects the knowledge captured by the topics. An annual granularity on this entire corpus allows us to know the background, while a monthly granularity exclusively applied on COVID-19 dates allows us to know in more detail the development of this disease. For example, according to Table 9, the T_{32} shows the appearance

of heart failure and ventilators, a consequence and a treatment of the disease, respectively.

According to [1], the “interpretation of a topic takes into account not only the top N words, but also the top documents in a topic”. Therefore, we analyse whether the labels correspond to the knowledge captured by the topics by analysing the correspondence with the most relevant documents. For this task, we validate some examples of the entire corpus

Table 7 DTM models performance for SARS-CoV-2 dates

Topics	Time slices	C_V	UMass	Diversity
100	2022-06	0.6509	-2.7755	0.5180
	2022-03	0.6520	-2.7732	0.5080
	2021-11	0.6533	-2.7469	0.5100
200	2022-06	0.6158	-3.3570	0.5487
	2022-03	0.6175	-3.3657	0.5497
	2021-11	0.6185	-3.3593	0.5483
300	2022-06	0.5726	-3.7275	0.5276
	2022-03	0.5722	-3.7183	0.5282
	2021-11	0.5731	-3.7159	0.5282
400	2022-06	0.5401	-3.8803	0.4762
	2022-03	0.5409	-3.8891	0.4762
	2021-11	0.5414	-3.8912	0.4745

DTM model. The model that captures the annual evolution of the entire corpus.

The top documents on the topics are easy to obtain and are provided by the model. However, the cross-domain documents are not and require the following actions to be carried out. To extract cross-domain documents, the topic label vectors have to be generated. Vectors are generated through the words on their labels for a given date. The embeddings of each word are taken, and their corresponding centre becomes the representative vector of the topic (in case there is a single word in the label, its embedding will become the vector of the topic). To generate vectors for vocabulary words, as mentioned above, we used the word2vec model. The model we used was the one we used to generate the topics with the DETM model, and its configuration is given in Table 4. We analyse whether the vector representation is correct or not through the relationship of the nearest neighbour documents. This means that for each topic vector, we will look for its nearest neighbour, and using the cross-domain documents among them, we will analyse whether the semantic relationship between the labels corresponds to the real relationship.

Top Document Validation

As collected by the topic label T_{64} , there is a topic related to outbreaks of diseases because until 2018, the label “outbreak” remains unchanged. Some of the most significant papers up to that date are the works of [64], [65] and [66]. These articles discuss the real-time detection of mycobacteriosis outbreaks in swine slaughterhouses, the 2009 influenza A pandemic (caused by the H1N1 virus), and avian influenza infections in China during 2013 and 2017 (caused by the H7N9 virus), respectively. In 2019, the label evolves and acquires the term “data”, which implies that, in published articles, the information used is very important. The most important work on this topic to date is [67]. This work discusses the outbreaks of the Ebola virus in the northeast of

the Democratic Republic of Congo. The authors used up-to-date information from that time combined with case counts from previous outbreaks to estimate the short- and long-term development of the outbreak and, in addition, to estimate the impact of vaccination.

In 2020, the label changes and the disease “covid-19” appears on it, being [68] the most relevant work for this date, omitting preprints. The authors calculated the time to double COVID-19 cases in the USA before and after mobility restrictions, using disease surveillance data. In 2021, the label remains unchanged, implying that studies follow the same course as in [69]. In this investigation, the authors evaluated the effect of the policies taken by the Cyprus government during the second wave of the COVID-19 epidemic using different sources of information (sources from the Cyprus government). In 2022, the label continues to evolve and acquires a new term: “wave”. This makes it clear that much of the work developed during 2022 on coronavirus data analyses the different waves experienced. The most relevant work on this topic to date is [70]. In this work, the researchers created a model to predict COVID-19 outbreaks during the first and second waves in 5 different countries using official information provided by these countries.

Like T_{64} , the other topics also follow an evolution, some of them with more changes than others. For example, the topic T_{71} talked about “injury” until 2000; then introduced the word “clinical” until 2015. After that date, the word “lesion” disappears, and the word “clinical” remains the only label on the topic until 2018. The rest of the evolution can be observed in the table. Among the most representative documents on this topic, we find clinical reports of cases of diseases (some painful) in patients [71], [72], [73].

Another example is the topic T_{79} , its label remains “pneumonia” until 2019, after which COVID-19 appears. Due to the impact of COVID-19 and the relationship it has with

Table 8 An excerpt from the dynamic labels of the DTM model generated during the entire period of time

	T₆₄	T₇₁	T₇₉	T₉₃
2022	covid-19, data, wave	pain, patient, report	covid-19, imaging, pneumonia	cardiac, heart
2021	covid-19, data	clinical, patient, report	covid-19, pneumonia	cardiac, heart
2020	covid-19, data	clinical, report	covid-19, pneumonia	cardiac, heart
2019	data, outbreak	clinical, report	pneumonia	heart
2018	outbreak	clinical	pneumonia	heart

Table 9 An excerpt from the dynamic labels of the DTM model generated during the COVID-19 period

	T₃₂	T₃₅	T₄₄	T₆₆
June 2022	cardiac, failure, heart, patient	sars-cov-2, variant	lockdown, period	anxiety, mental, symptom
November 2021	cardiac, heart, patient	sars-cov-2, variant	lockdown, period	anxiety, mental, symptom
April 2021	patient, ventilation	protein, sars-cov-2	lockdown, period	anxiety, mental
August 2020	patient, ventilation	protein, sars-cov-2	lockdown, period	anxiety, mental
December 2019	patient, ventilation	protein, spike	period	mental

pneumonia, the label adds the term “COVID-19” to its definition. Furthermore, in 2022 the term “imaging” appears, which in the medical domain refers to images of the inside of the body for clinical analysis and medical interventions (computed tomography (CT) scans, X-rays, magnetic resonance imaging (MRI), etc.). Some of the most important works use CT scans as part of their research. Some applications are: detecting the distinctive characteristics of SARS-CoV-2 in patients with COVID-19 [74], differentiating the diagnosis of COVID-19 from other diseases [75], predicting the occurrence of diseases caused by COVID-19 [76], and studying the importance of pneumonia detected on CT scans but not chest radiographs [77].

One of the most changing topic labels is **T₉₃**, although it always followed the same guideline. Among the terms that have appeared on this label, before 2018 we find “recipient”, “patient”, and “transplant”; it is not until 2015 that “heart” appears on the label for the first time. This tells us that heart transplantation has gradually gained relevance in the literature. Some of the most relevant papers have been [78], [79], [80], and [81]. All of these investigations are related to heart transplants, either the consequences on patients or the possible diseases and infections derived from this process.

Cross-Domain Document Validation

For topic **T₆₄**, the closest topic in 2022, 2021, and 2020 is topic **T₉**, while for date 2019 the closest is **T₈** and for 2018 and earlier dates it is **T₈₉**. For the dates when they are neighbours, **T₉** is labelled as “covid-19” and “pandemic”. The most relevant cross-domain work for these periods is: the study of mortality rates in patients with COVID-19 during the first three waves in Milan [82], the effects of the COVID-19 pandemic on the environment [83], and the monitoring of the transmissibility and mortality of the COVID-19 pandemic

[84]. The **T₈** is labelled as “data” while the neighbouring **T₆₄** in 2019 and, the most relevant cross-domain document, is an investigation of the potential use of search queries and social networking data to detect and monitor human cases of avian influenza A (H7N9) in China [85]. Finally, for 2018 and earlier dates, **T₈₉**, labelled as “assay”, is the nearest neighbour of **T₆₄**. Some of the most relevant cross-domain documents for this set of dates are the use of a PCR assay to detect a wide range of human adenoviruses [86] and the comparison of two different immunoassays to detect respiratory virus antigens [87].

For topic **T₇₁**, the nearest neighbour until 2000 is topic **T₈₉**, but in 2021 it is again the nearest neighbour. On this date (2021), the topic **T₈₉** is labelled as “assay” and “detection” and the most important cross-domain document is the study of viral loads of SARS-CoV-2 in throat and exhalation samples from hospitalised patients by COVID-19 [88]. For the years 2001, 2002, 2016, and 2017, the closest topic is **T₁₇**. From 2003 to 2015 and from 2018 to 2020, the closest topic is **T₉₆**, and during these last dates (2018 to 2020), **T₉₆** is labelled as “clinical”, “drug”, and “trial”. Some of the most important cross-domain documents for these dates are a comparison of the effectiveness of two different types of enteral nutrition in critically ill patients [89] and a comparison of clinical manifestations between different age groups of patients with COVID-19 [90]. For 2022, the closest topic is **T₃₀**, and is labelled as “patient” and “symptom”. One of the most important cross-domain documents is a study of the development of a disorder in patients who had recovered from COVID-19 [91].

For topic **T₇₉**, the topic **T₈₉** is the closest neighbour until 2019 and, from this date, the closest is **T₉**. The **T₈₉** until 2018 is labelled as “assay”, and in 2019, it is also labelled as “detec-

tion”. One of the most important cross-domain documents for these domains is an assay to diagnose pneumococcal pneumonia [92]. During the time when T_9 is the nearest neighbour, it is labelled as “covid-19” and “pandemic”. One of the most relevant cross-domain documents for these domains is a study related to paediatric radiology changes during the COVID-19 pandemic [93].

The topic T_{93} has 4 different neighbours. The T_{89} until 1990, from 2008 to 2018, and the years 2020 and 2021; the T_{28} from 1990 to 2007; the T_{46} in 2019; and the T_{64} in 2022. Between domains T_{93} and T_{89} , the most relevant works are: a paper on community preparedness for cardiovascular emergencies during the COVID-19 pandemic [94]; and a comparison of arrhythmia diagnosis in human-dependent versus algorithm-dependent systems [95]. At the date of these articles (2021), as we mentioned previously, the labels of the T_{89} topic are “assay” and “detection”. During the period when the topic T_{28} is the neighbour of T_{93} , the label of T_{28} is: “patient” and “risk” during the dates 1990 to 1997; and then the label “risk” disappears from the topic. Meanwhile, the T_{93} topic label does not change during these dates, its label remains as “patient” and “recipient”. The most relevant cross-domain documents during this period for these domains are the results of stem cell transplantation for a rare disease [96], [97]; and bone stem cell transplantation for a blood disorder [98]. For 2020, the topic T_{46} was labelled as “medium” and “social”, and although there are no cross-domain documents for these dates, there are documents for another date on the effect of news coverage and online advertising on patients seeking information on aspirin, heart health, and stroke [99]. For the topic T_{64} and T_{93} , one of the most relevant cross-domain documents talks about the impact of COVID-19 on heart transplant recipients [100].

6 Conclusions and future work

We have shown that the evolution of concepts can be captured from dynamic topic labels. For a collection of documents of a particular domain with the terms that appear in texts. We generated several DTM, DETM, and BERTopic models and evaluated which generates better topics for the CORD-19 corpus, identifying the strengths and weaknesses of each one. They are available in Zenodo [9], with all models, the abstracts processed, and the script we used to evaluate the models, the code used to generate the labels, and a *jupyter notebook* where the different experiments carried out can be tested. Because the key value of this work is the labelling strategy, we have created another resource for it [10], where we have uploaded the labels corresponding to the models analysed. Please note that, due to the preprocessing decision to remove all numbers, terms such as “COVID-19” appear as

“covid-” and others such as “SARS-CoV-2” appear as “sars-cov-”.

Interpretability is a fundamental axis of this work and that is why we have emphasised it in the description of the techniques we have used and in the reports we have included among our resources. These resources prove to be useful for analysing the evolution of the CORD-19 corpus because of their simplicity. These techniques could be used with any other topic model if the topic-document and term-topic matrices are available. Models from other domains can be easily analysed. The main utility of our tagging system lies in two points. First, we make it easier for a domain expert to find the topic he/she is interested in, avoiding the need to study each topic individually. Second, with this representation, the temporal evolution of the topics is also captured by the labels, making it easier to understand how the collection of texts has evolved. The evaluation we did is incomplete, since we did not analyse all topics. However, the results are robust and support the labelling strategy. Even so, among the available resources, there are tools to perform the same analysis on any topic and document in the corpus, so that in case you want to study the relationships between other topics and their labels, you can easily do it.

In summary, our analyses have shown that the DTM model is the best for analysing the CORD-19 corpus and that our novel approach to labelling dynamic topics matches the content of each topic. We studied the labels of some topics and performed a supervised evaluation. In this evaluation, we found that the labels represent the most relevant documents for each topic and facilitated the interpretation of the topic content. Additionally, with these labels, we generate topic vectors to organise topics in a semantic space. We found that the semantic correspondence of labels, i.e. the distance between topic vectors, is strongly related to the cross-domain documents with the closest topic. In fact, there is no topic that does not have at least one document that is shared with its nearest neighbour.

We want to continue this work by making label evaluation an unsupervised task. We currently recognise that the number of topics that can exist in a text collection is too large to be analysed individually. Therefore, exploring different approaches that can assess the interpretability of labels and their correspondence with topics is one of our next steps. Moreover, although the words inside the labels are related, the relationship is not explicit, and in complex domains such as the medical domain, the same disease can be both a symptom and a consequence. Therefore, we want to work on the relationship between the words within the labels to improve the representation of the captured topic.

Appendix A: dynamic topic labelling flowchart Appendix B: Models performance in detail

See Fig. 9.

See Tables 10, 11, 12, 13 and Fig. 10, 11, 12, 13, 14



Fig. 9 Dynamic Topic labelling procedure

Table 10 Complete corpus generated DTM models performance

Topics	Time slices	C_V	UMass	Diversity
100	2022	0.6560	-2.7859	0.5320
	2021	0.6580	-2.7678	0.5359
	2020	0.6601	-2.7604	0.5353
	2019	0.6636	-2.7243	0.5307
	2018	0.6625	-2.7232	0.5347
	2015	0.6609	-2.7192	0.5347
	2010	0.6617	-2.7383	0.5407
	2005	0.6593	-2.7409	0.5487
	2000	0.6570	-2.7606	0.5487
	1995	0.6557	-2.7453	0.5480
200	2022	0.6323	-3.2246	0.5307
	2021	0.6317	-3.2347	0.5273
	2020	0.6331	-3.2271	0.5240
	2019	0.6330	-3.2318	0.5190
	2018	0.6334	-3.2248	0.5157
	2015	0.6323	-3.2158	0.5183
	2010	0.6341	-3.2008	0.5253
	2005	0.6339	-3.2201	0.5297
	2000	0.6329	-3.2285	0.5320
	1995	0.6318	-3.2176	0.5343
300	2022	0.5954	-3.5012	0.5127
	2021	0.5957	-3.5090	0.5104
	2020	0.5955	-3.5015	0.5078
	2019	0.5954	-3.5013	0.5082
	2018	0.5955	-3.5038	0.5098
	2015	0.5964	-3.4909	0.5049
	2010	0.5959	-3.4816	0.5080
	2005	0.5939	-3.4592	0.5064
	2000	0.5931	-3.4592	0.5087
	1995	0.5926	-3.4611	0.5098
400	2022	0.5688	-3.5940	0.4592
	2021	0.5695	-3.5918	0.4585
	2020	0.5693	-3.5951	0.4592
	2019	0.5689	-3.5982	0.4577
	2018	0.5692	-3.5969	0.4558
	2015	0.5683	-3.5973	0.4523
	2010	0.5675	-3.6044	0.4505
	2005	0.5679	-3.5991	0.4523
	2000	0.5674	-3.5918	0.4507
	1995	0.5672	-3.5916	0.4501

Table 11 BERTopic models performance

Topics	Time slices	C_V	UMass	Diversity
100	2022	0.8068	-2.0985	0.2260
	2021	0.8069	-2.0915	0.2287
	2020	0.8039	-2.1092	0.2267
	2019	0.6007	-4.7835	0.2207
	2018	0.5409	-5.4568	0.2195
	2015	0.5132	-5.7133	0.2140
	2010	0.4831	-5.5099	0.2219
	2005	0.4833	-5.3199	0.2513
	2000	0.5531	-4.3617	0.2682
	1995	0.5931	-3.8475	0.2483
200	2022	0.8042	-2.2982	0.1967
	2021	0.8094	-2.2177	0.1980
	2020	0.8068	-2.2362	0.1967
	2019	0.5080	-6.1321	0.1842
	2018	0.4824	-6.1329	0.1983
	2015	0.4824	-6.0701	0.1983
	2010	0.4883	-5.4179	0.2040
	2005	0.4793	-5.6793	0.2214
	2000	0.5350	-4.5696	0.2614
	1995	0.5856	-4.0310	0.2593
300	2022	0.7575	-3.0198	0.1536
	2021	0.8049	-2.3334	0.1564
	2020	0.7958	-2.4539	0.1558
	2019	0.4730	-6.3694	0.1648
	2018	0.4820	-6.2017	0.1781
	2015	0.4851	-5.9573	0.1849
	2010	0.4893	-5.6193	0.1967
	2005	0.4886	-5.4465	0.2198
	2000	0.5499	-4.7411	0.2622
	1995	0.5708	-4.2193	0.2667
400	2022	0.6917	-3.9798	0.1243
	2021	0.7837	-2.6038	0.1277
	2020	0.7516	-3.0344	0.1250
	2019	0.4768	-6.3873	0.1550
	2018	0.4762	-6.1656	0.1646
	2015	0.4802	-6.0804	0.1725
	2010	0.4857	-5.8169	0.1892
	2005	0.4812	-5.6779	0.2325
	2000	0.5396	-4.9553	0.2617
	1995	0.5827	-4.1344	0.2533

Table 12 DETM models performance

Model	Topics	Time slices	C_V	UMass	Diversity
DETM DeBERTa	100	2022	0.2533	−6.6634	0.0853
		2021	0.2687	−7.3585	0.0580
		2020	0.2597	−7.1033	0.0633
		2019	0.2496	−6.9977	0.1867
		2018	0.2492	−7.0751	0.2007
		2015	0.2507	−7.1097	0.2220
		2010	0.2606	−7.5717	0.2627
		2005	0.2853	−8.0903	0.3200
		2000	0.3040	−9.0051	0.3887
		1995	0.3430	−10.0540	0.3740
DETM DeBERTa	200	2022	0.2536	−6.9542	0.0903
		2021	0.2422	−6.5506	0.0483
		2020	0.2552	−7.1488	0.0707
		2019	0.2552	−7.1930	0.1530
		2018	0.2683	−7.6082	0.1757
		2015	0.2744	−7.7431	0.2293
		2010	0.2802	−7.9932	0.2257
		2005	0.2998	−8.6144	0.2580
		2000	0.3234	−9.3984	0.2500
		1995	0.3217	−9.4125	0.2507
DETM Word2Vec	100	2022	0.5525	−3.2132	0.2120
		2021	0.5395	−2.7765	0.1167
		2020	0.5547	−2.7606	0.1333
		2019	0.5765	−4.0833	0.4059
		2018	0.5393	−4.5920	0.4353
		2015	0.5159	−5.0228	0.4667
		2010	0.4311	−6.3555	0.5013
		2005	0.4180	−6.2201	0.4727
		2000	0.3533	−7.3136	0.5200
		1995	0.3284	−7.4762	0.5273
DETM Word2Vec	200	2022	0.5471	−3.4626	0.1743
		2021	0.5626	−3.0588	0.1213
		2020	0.5661	−3.0505	0.1327
		2019	0.5335	−4.5557	0.2853
		2018	0.4903	−5.1812	0.2743
		2015	0.4713	−5.6803	0.3230
		2010	0.4222	−6.1266	0.3337
		2005	0.4057	−6.6103	0.3277
		2000	0.3201	−7.6820	0.3417
		1995	0.3218	−7.8097	0.3339

Table 13 DTM models performance for SARS-CoV-2 dates

Topics	Time slices	C_V	UMass	Diversity
100	2022-06	0.6509	-2.7755	0.5180
	2022-05	0.6507	-2.7742	0.5153
	2022-04	0.6516	-2.7846	0.5133
	2022-03	0.6520	-2.7732	0.5080
	2022-02	0.6516	-2.7618	0.5093
	2021-11	0.6533	-2.7469	0.5100
	2021-06	0.6562	-2.7191	0.5120
	2021-01	0.6575	-2.6964	0.5033
	2020-08	0.6596	-2.6863	0.5100
	2020-03	0.6526	-2.7268	0.5200
200	2022-06	0.6158	-3.3570	0.5487
	2022-05	0.6160	-3.3588	0.5480
	2022-04	0.6172	-3.3554	0.5497
	2022-03	0.6175	-3.3657	0.5497
	2022-02	0.6156	-3.3506	0.5470
	2021-11	0.6185	-3.3593	0.5483
	2021-06	0.6215	-3.3467	0.5489
	2021-01	0.6203	-3.3492	0.5473
	2020-08	0.6179	-3.3636	0.5457
	2020-03	0.6149	-3.3522	0.5399
300	2022-06	0.5726	-3.7275	0.5276
	2022-05	0.5725	-3.7275	0.5296
	2022-04	0.5730	-3.7263	0.5291
	2022-03	0.5722	-3.7183	0.5282
	2022-02	0.5724	-3.7152	0.5278
	2021-11	0.5731	-3.7159	0.5282
	2021-06	0.5739	-3.7218	0.5269
	2021-01	0.5743	-3.7142	0.5296
	2020-08	0.5759	-3.7219	0.5336
	2020-03	0.5749	-3.7406	0.5316
400	2022-06	0.5401	-3.8803	0.4762
	2022-05	0.5403	-3.8812	0.4760
	2022-04	0.5409	-3.8850	0.4760
	2022-03	0.5409	-3.8891	0.4762
	2022-02	0.5412	-3.8940	0.4767
	2021-11	0.5414	-3.8912	0.4745
	2021-06	0.5421	-3.8872	0.4715
	2021-01	0.5418	-3.8607	0.4683
	2020-08	0.5418	-3.8631	0.4678
	2020-03	0.5412	-3.8547	0.4633

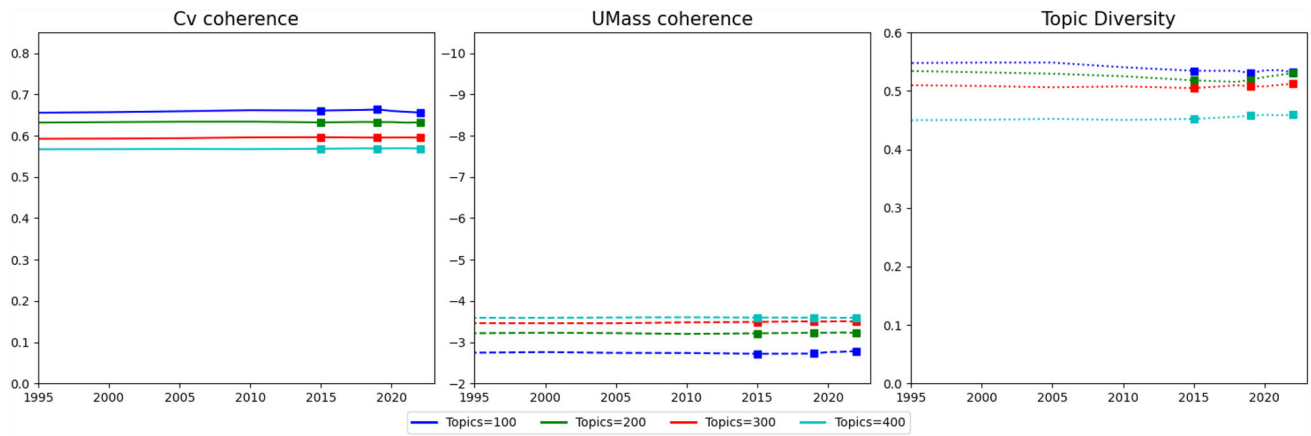


Fig. 10 DTM models performance

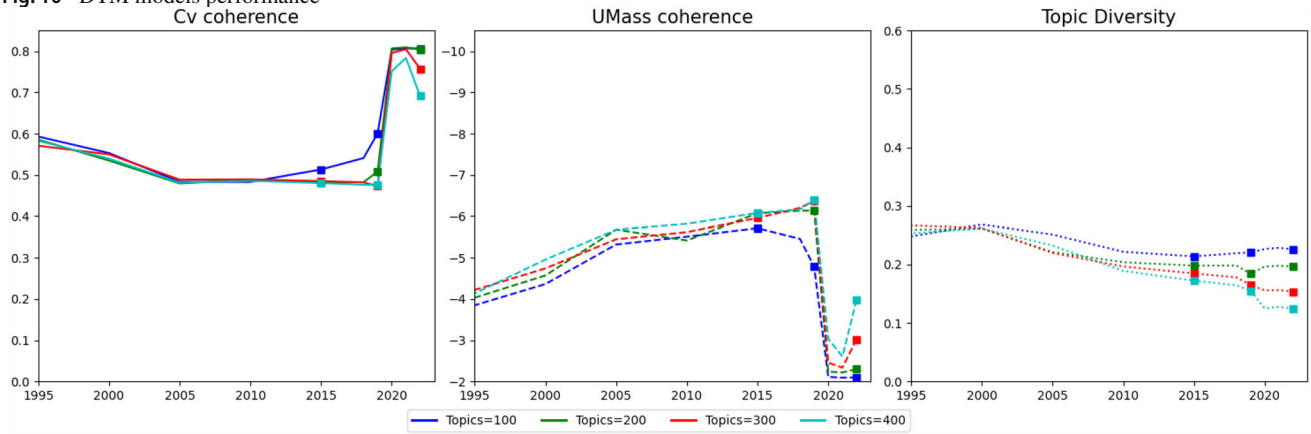


Fig. 11 BERTopic models performance

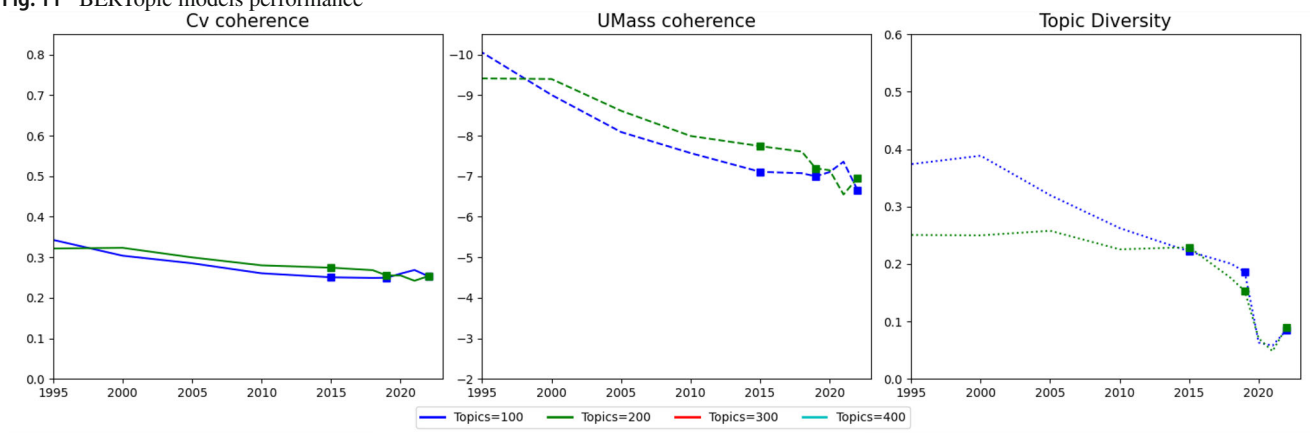


Fig. 12 DETM DeBERTa models performance

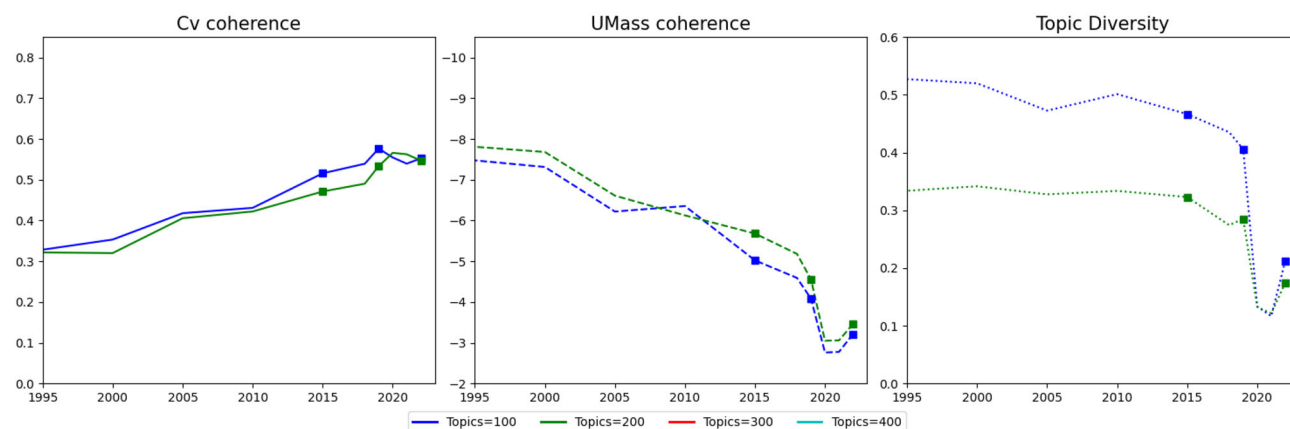


Fig. 13 DETM Word2Vec models performance

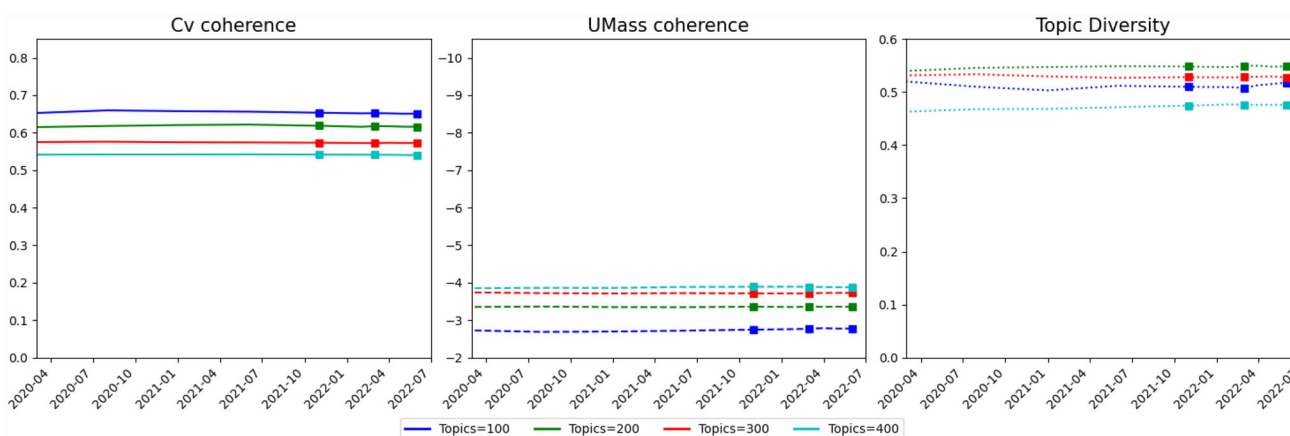


Fig. 14 DTM models performance for SARS-CoV-2 dates

Acknowledgements The authors gratefully acknowledge the Universidad Politécnica de Madrid (www.upm.es) for providing computing resources on Magerit Supercomputer. This work is supported by the DRUGS4COVID++ project, funded by Ayudas Fundación BBVA a equipos de investigación científica SARS-CoV-2 y COVID-19; and by the Predoctoral Grant (PIPF-2022/COM-25947) of the Consejería de Educación, Ciencia y Universidades de la Comunidad de Madrid, Spain.

Author Contributions Ibai Guillen-Pacho has developed and written the entire content of the article, which means that he has carried out the research, written the text and created the necessary tables, graphs and images. Carlos Badenes-Olmedo has played a crucial role in the process, guiding and reviewing the research at all times. This means that he provided guidance and advice to improve the content and ensure that it conforms to academic standards. He has also made some grammatical and structural corrections to the text. Oscar Corcho has contributed his expertise to the article, which means that he reviewed the content to ensure its accuracy and relevance. In addition, he structured the article in a way that is easy to understand for the intended audience and ensured that the conventions of writing style appropriate to the type of article are followed.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work is supported by the DRUGS4COVID++ project, funded by Ayudas Fundación BBVA a

equipos de investigación científica SARS-CoV-2 y COVID-19; and by the Predoctoral Grant (PIPF-2022/COM-25947) of the Consejería de Educación, Ciencia y Universidades de la Comunidad de Madrid, Spain.

Declarations

Conflict of interest The authors declare that they have no conflicting interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Isoaho, K., Gritsenko, D., Mäkelä, E.: Topic modeling and text analysis for qualitative policy research. *Policy Stud. J.* **49**(1), 300–324 (2019). <https://doi.org/10.1111/psj.12343>
- Zha, D., Li, C.: Multi-label dataless text classification with topic modeling. *Knowl. Inf. Syst.* **61**(1), 137–160 (2019). <https://doi.org/10.1007/s10115-018-1280-0>
- Li, Y., Nair, P., Wen, Z., Chafi, I., Okhmatovskaia, A., Powell, G., Shen, Y., Buckeridge, D.: Global surveillance of COVID-19 by mining news media using a multi-source dynamic embedded topic model. In: *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. ACM, Atlanta (2020). <https://doi.org/10.1145/3388440.3412418>
- Greene, D., Cross, J.P.: Exploring the political agenda of the European parliament using a dynamic topic modeling approach. *Polit. Anal.* **25**(1), 77–94 (2017). <https://doi.org/10.1017/pan.2016.7>
- Hu, J., Sun, X., Lo, D., Li, B.: Modeling the evolution of development topics using dynamic topic models. In: *2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, pp. 3–12. IEEE, Macao SAR (2015). <https://doi.org/10.1109/saner.2015.7081810>
- Zou, C.: Analyzing research trends on drug safety using topic modeling. *Expert Opin. Drug Saf.* **17**(6), 629–636 (2018). <https://doi.org/10.1080/14740338.2018.1458838>
- Aletras, N., Baldwin, T., Lau, J.H., Stevenson, M.: Representing topics labels for exploring digital libraries. In: *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries. JCDL '14*, pp. 239–248. IEEE Press, London, United Kingdom (2014). <https://doi.org/10.1109/JCDL.2014.6970174>
- He, D., Ren, Y., Khattak, A.M., Liu, X., Tao, S., Gao, W.: Automatic topic labeling model with paired-attention based on pre-trained deep neural network, vol. 2021–July. Institute of Electrical and Electronics Engineers Inc., (2021). <https://doi.org/10.1109/IJCNN52387.2021.9534093>. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85116421761&doi=10.1109%2FIJCNN52387.2021.9534093&partnerID=40&md5=3f8bc9f37b65f8255fcbf9b0773defe2>
- Guillén-Pacho, I.: Dynamic Topic Models of “Dynamic Topic Modelling for Exploring the Scientific Literature on Coronavirus: An Unsupervised Labelling Technique”, version 1.0.0. Zenodo (2024). <https://doi.org/10.5281/zenodo.12750327>
- Guillén-Pacho, I.: Topic Labels of “Dynamic Topic Modelling for Exploring the Scientific Literature on Coronavirus: An Unsupervised Labelling Technique”, version 1.0.0. Zenodo (2024). <https://doi.org/10.5281/zenodo.12752032>
- Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. arXiv (2013). <https://doi.org/10.48550/ARXIV.1301.3781>
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. arXiv (2013). <https://doi.org/10.48550/ARXIV.1310.4546>
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv (2018). <https://doi.org/10.48550/ARXIV.1810.04805>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. arXiv (2017). <https://doi.org/10.48550/ARXIV.1706.03762>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv (2019). <https://doi.org/10.48550/ARXIV.1907.11692>
- He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv (2020). <https://doi.org/10.48550/ARXIV.2006.03654>
- Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv (2019). <https://doi.org/10.48550/ARXIV.1908.10084>
- Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120 (2006). <https://doi.org/10.1145/1143844.1143859>
- Dieng, A.B., Ruiz, F.J.R., Blei, D.M.: The Dynamic Embedded Topic Model. arXiv (2019). <https://doi.org/10.48550/ARXIV.1907.05545>
- Grootendorst, M.: BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv (2022). <https://doi.org/10.48550/ARXIV.2203.05794>
- Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**(Jan), 993–1022 (2003)
- Hofmann, T.: Probabilistic latent semantic indexing. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57 (1999). <https://doi.org/10.1145/312624.312649>
- Avasthi, S., Chauhan, R., Acharjya, D.P.: Extracting information and inferences from a large text corpus. *Int. J. Inf. Technol.* **15**(1), 435–445 (2022). <https://doi.org/10.1007/s41870-022-01123-4>
- Bai, Y., Jia, S., Chen, L.: Topic evolution analysis of COVID-19 news articles. In: *Journal of Physics: Conference Series*, vol. 1601(5), p. 052009 (2020). <https://doi.org/10.1088/1742-6596/1601/5/052009>
- Avasthi, S., Chauhan, R., Acharjya, D.P.: Information extraction and sentiment analysis to gain insight into the COVID-19 crisis. In: *Advances in Intelligent Systems and Computing*, pp. 343–353. Springer (2021). https://doi.org/10.1007/978-981-16-2594-7_28
- Dieng, A.B., Ruiz, F.J.R., Blei, D.M.: Topic modeling in embedding spaces. *Trans. Assoc. Comput. Linguist.* **8**, 439–453 (2020). https://doi.org/10.1162/tacl_a_00325
- Messan, K.S., Sulima, P.P., Ghosh, D., Nye, J.: The research foundation for COVID-19 vaccine development. *Front. Res. Metr. Anal.* **8**, 1078971 (2023). <https://doi.org/10.3389/frma.2023.1078971>
- Lau, J.H., Newman, D., Karimi, S., Baldwin, T.: Best Topic Word Selection for Topic Labelling. In: *Coling 2010: Posters*, pp. 605–613. Coling 2010 Organizing Committee, Beijing, China (2010)
- Lee, T.Y., Smith, A., Seppi, K., Elmqvist, N., Boyd-Graber, J., Findlater, L.: The human touch: How non-expert users perceive, interpret, and fix topic models. *Int. J. Hum. Comput. Stud.* **105**, 28–42 (2017). <https://doi.org/10.1016/j.ijhcs.2017.03.007>
- Weston, S.J., Shryock, I., Light, R., Fisher, P.A.: Selecting the number and labels of topics in topic modeling: a tutorial. *Adv. Methods Pract. Psychol. Sci.* **6**(2), 25152459231160104 (2023). <https://doi.org/10.1177/25152459231160105>
- Bhatia, S., Lau, J.H., Baldwin, T.: Automatic Labelling of Topics with Neural Embeddings. arXiv (2016). <https://doi.org/10.48550/arXiv.1612.05340>
- Mitrofanova, O., Kriukova, A., Shulginov, V., Shulginov, V.: E-hypertext media topic model with automatic label assignment. *Communications in Computer and Information Science* 1357 CCIS, 102–114 (2021). https://doi.org/10.1007/978-3-030-71214-3_9
- Sumpeno, S., Purwitasari, D., Farandy, B., Navastara, D.A., Purnomo, M.H.: Graph model and deep learning for topic labels in classifying short texts of scientific article titles. *Int. J. Intell. Eng. Syst.* (2022). <https://doi.org/10.22266/ijies2022.0430.41>
- Zheng, B., Li, H.: Automatic Labeling of SDN Controller Defect Text Based on Neural Topic Model, pp. 402–410. Institute

- of Electrical and Electronics Engineers Inc. (2022). <https://doi.org/10.1109/QRS-C57518.2022.00065>. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85152626795&doi=10.1109%2fQRS-C57518.2022.00065&partnerID=40&md5=b34627485e934d235bad74f376fe96f>
35. Bouziane, Y., Abdi, M.K., Sadou, S.: Automatically labelled software topic model. *Int. J. Open Source Softw. Process.* **11**(1), 57–78 (2020). <https://doi.org/10.4018/IJOSSP.2020010104>
 36. Kozbagarov, O., Mussabayev, R., Mladenovic, N.: A new sentence-based interpretative topic modeling and automatic topic labeling. *Symmetry* **13**(5), 837 (2021). <https://doi.org/10.3390/sym13050837>
 37. Ogawa, T., Saga, R.: Text-based causality modeling with a conceptual label in a hierarchical topic structure using Bayesian rose trees. In: *Proceedings of the Annual Hawaii International Conference on System Sciences*, vol. 2020-January, pp. 1101–1110 (2021)
 38. Rahman, S., Hossain, S.S., Arman, M.S., Rawshan, L., Toma, T.R., Rafiq, F.B., Badruzzaman, K.B.M.: Assessing the effectiveness of topic modeling algorithms in discovering generic label with description. In: Arai, K., Kapoor, S., Bhatia, R. (eds.) *Advances in Information and Communication. Advances in Intelligent Systems and Computing*, pp. 224–236. Springer International Publishing, Cham (2020)
 39. Kinariwala, S., Deshmukh, S.: Onto_tml: auto-labeling of topic models. *J. Integr. Sci. Technol.* **9**(2), 85–91 (2021)
 40. Mihalcea, R., Tarau, P.: TextRank: Bringing Order into Text. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404–411. Association for Computational Linguistics, Barcelona (2004)
 41. Miller, G.A.: WordNet: a lexical database for English. *Commun. ACM* **38**(11), 39–41 (1995). <https://doi.org/10.1145/219717.219748>
 42. Ji, L., Wang, Y., Shi, B., Zhang, D., Wang, Z., Yan, J.: Microsoft concept graph: mining semantic concepts for short text understanding. *Data Intell.* **1**(3), 238–270 (2019). https://doi.org/10.1162/dint_a_00013
 43. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: *32nd Annual Meeting of the Association for Computational Linguistics*, pp. 133–138. Association for Computational Linguistics, Las Cruces, New Mexico (1994) <https://doi.org/10.3115/981732.981751>
 44. Lewis, C.I.: An analysis of knowledge and valuation. *Mind* **57**(225), 71–85 (1946)
 45. BonJour, L.: *The Structure of Empirical Knowledge*. Harvard University Press, Cambridge (1985)
 46. Bovens, L., Hartmann, S.: *Bayesian Epistemology*. Oxford University Press, Oxford (2004)
 47. Douven, I., Meijs, W.: Measuring coherence. *Synthese* **156**(3), 405–425 (2007). <https://doi.org/10.1007/s11229-006-9131-z>
 48. Mavrin, A., Filchenkov, A., Koltcov, S.: Four keys to topic interpretability in topic modeling. In: Ustalov, D., Filchenkov, A., Pivovarov, L., Žižka, J. (eds.) *Artificial Intelligence and Natural Language. Communications in Computer and Information Science*, pp. 117–129. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-01204-5_12
 49. Chemudugunta, C., Smyth, P., Steyvers, M.: Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. In: *Advances in Neural Information Processing Systems*, vol. 19. MIT Press (2006)
 50. Mei, Q., Shen, X., Zhai, C.: Automatic labeling of multinomial topic models. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 490–499. ACM, San Jose California USA (2007). <https://doi.org/10.1145/1281192.1281246>
 51. 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.: CORD-19: The COVID-19 open research dataset. In: *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. Association for Computational Linguistics, Online* (2020). <https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1>
 52. Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., Adam, S.: Applying LDA topic modeling in communication research: toward a valid and reliable methodology. *Commun. Methods Meas.* **12**(2–3), 93–118 (2018). <https://doi.org/10.1080/19312458.2018.1430754>
 53. Gao, Q., Huang, X., Dong, K., Liang, Z., Wu, J.: Semantic-enhanced topic evolution analysis: a combination of the dynamic topic model and word2vec. *Scientometrics* **127**(3), 1543–1563 (2022). <https://doi.org/10.1007/s11192-022-04275-z>
 54. Lou, W., Meng, J.: The diversity of canonical and ubiquitous progress in computer vision: a dynamic topic modeling approach. *Inf. Process. Manag.* **60**(3), 103238 (2023). <https://doi.org/10.1016/j.ipm.2022.103238>
 55. Rumbut, J., Fang, H., Wang, H.: Topic modeling for systematic review of visual analytics in incomplete longitudinal behavioral trial data. *Smart Health* **18**, 100142 (2020). <https://doi.org/10.1016/j.smhl.2020.100142>
 56. Atzeni, D., Bacciu, D., Mazzei, D., Prencipe, G.: A systematic review of Wi-Fi and machine learning integration with topic modeling techniques. *Sensors* **22**(13), 4925 (2022). <https://doi.org/10.3390/s22134925>
 57. Shausan, A., Vuorinen, A.: Thirty-six years of contributions to queueing systems: a content analysis, topic modeling, and graph-based exploration of research published in the QUESTA journal. *Queueing Syst.* **104**(1), 3–18 (2023). <https://doi.org/10.1007/s11134-023-09876-w>
 58. Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 262–272. Association for Computational Linguistics, Edinburgh (2011). <https://aclanthology.org/D11-1024>
 59. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. WSDM '15*, pp. 399–408. Association for Computing Machinery, New York (2015). <https://doi.org/10.1145/2684822.2685324>
 60. Bouma, G.: Normalized (pointwise) mutual information in collocation extraction. In: *Proceedings of the Biennial GSCL Conference*, pp. 31–40 (2009)
 61. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. *Comput. Linguist.* **16**(1), 22–29 (1990)
 62. Badenes-Olmedo, C., Corcho, O.: Lessons learned to enable question answering on knowledge graphs extracted from scientific publications: a case study on the coronavirus literature. *J. Biomed. Inform.* **142**, 104382 (2023). <https://doi.org/10.1016/j.jbi.2023.104382>
 63. Kumar, A., Esmaili, N., Piccardi, M.: A temperature-modified dynamic embedded topic model. In: Park, L.A.F., Gomes, H.M., Doborjeh, M., Boo, Y.L., Koh, Y.S., Zhao, Y., Williams, G., Simoff, S. (eds.) *Data Mining Communications in Computer and Information Science*, pp. 15–27. Springer Nature, Singapore (2022). https://doi.org/10.1007/978-981-19-8746-5_2
 64. Hsieh, Y.-H., Cheng, K.-F., Wu, T.-N., Li, T.-C., Chen, C.-Y., Chen, J.-H., Lin, M.-H.: Transmissibility and temporal changes

- of 2009 pH1N1 pandemic during summer and fall/winter waves. *BMC Infect. Dis.* **11**(1), 1–12 (2011). <https://doi.org/10.1186/1471-2334-11-332>
65. Adachi, Y., Makita, K.: Real time detection of farm-level swine mycobacteriosis outbreak using time series modeling of the number of condemned intestines in abattoirs. *J. Vet. Med. Sci.* **77**(9), 1129–1136 (2015). <https://doi.org/10.1292/jvms.14-0675>
 66. Guo, Z., Xiao, D., Li, D., Wang, Y., Yan, T., Dai, B., Wang, X.: The temporal distribution of new h7n9 avian influenza infections based on laboratory-confirmed cases in mainland china, 2013–2017. *Sci. Rep.* **8**(1), 4051 (2018). <https://doi.org/10.1038/s41598-018-22410-w>
 67. Worden, L., Wannier, R., Hoff, N.A., Musene, K., Selo, B., Mossoko, M., Okitolonda-Wemakoy, E., Tamfum, J.J.M., Rutherford, G.W., Lietman, T.M., Rimoin, A.W., Porco, T.C., Kelly, J.D.: Projections of epidemic transmission and estimation of vaccination impact during an ongoing Ebola virus disease outbreak in Northeastern Democratic Republic of Congo, as of Feb. 25, 2019. *PLoS Negl. Trop. Dis.* **13**(8), 0007512 (2019). <https://doi.org/10.1371/journal.pntd.0007512>
 68. Lurie, M.N., Silva, J., Yorlets, R.R., Tao, J., Chan, P.A.: Coronavirus disease 2019 epidemic doubling time in the united states before and during stay-at-home restrictions. *J. Infect. Dis.* **222**(10), 1601–1606 (2020). <https://doi.org/10.1093/infdis/jiaa491>
 69. Kakoullis, L., Eliades, E., Papachristodoulou, E., Parperis, K., Chra, P., Constantinidou, A., Chatzittofis, A., Sampsonas, F., Panos, G.: Response to COVID-19 in cyprus: policy changes and epidemic trends. *Int. J. Clin. Pract.* **75**(4), e13944 (2021). <https://doi.org/10.1111/ijcp.13944>
 70. Hoque, A., Malek, A., Zaman, K.M.R.A.: Data analysis and prediction of the COVID-19 outbreak in the first and second waves for top 5 affected countries in the world. *Nonlinear Dyn.* **109**(1), 77–90 (2022). <https://doi.org/10.1007/s11071-022-07473-9>
 71. Manlongat, E.V., Allen, W.C.: Incomplete Kawasaki disease presenting as bilateral hip synovitis. *J. Paediatr. Child Health* **53**(9), 912–914 (2017). <https://doi.org/10.1111/jpc.13586>
 72. ALHetheli, G.I., Aljarbou, O.Z., Alhumidi, A., AlJasser, M.I.: Atypical palmar involvement with erythema elevatum diutinum as a sole manifestation: A report of two cases. *Clin. Cosmet. Investig. Dermatol.* **13**, 529–535 (2020). <https://doi.org/10.2147/ccid.s241960>
 73. Lee, I.H., Cho, S., Ahn, D.J., Kim, M.-K.: Systemic lupus erythematosus presenting as hyponatremia-associated rhabdomyolysis. *Medicine* **100**(39), 27390 (2021). <https://doi.org/10.1097/md.00000000000027390>
 74. Saha, M., Amin, S.B., Sharma, A., Kumar, T.K.S., Kalia, R.K.: AI-driven quantification of ground glass opacities in lungs of COVID-19 patients using 3d computed tomography imaging. *PLoS ONE* **17**(3), 0263916 (2022). <https://doi.org/10.1371/journal.pone.0263916>
 75. Shi, W.-Y., Hu, S.-P., Zhang, H.-L., Liu, T.-F., Zhou, S., Tang, Y.-H., Zhang, X.-L., Shi, Y.-X., Zhang, Z.-Y., Xiong, N., Shan, F.: Differential diagnosis of COVID-19 pneumonia from influenza a (H1N1) pneumonia using a model based on clinicoradiologic features. *Front. Med.* **8**, 651556 (2021). <https://doi.org/10.3389/fmed.2021.651556>
 76. Mirza-Aghazadeh-Attari, M., Zarrintan, A., Nezami, N., Mohammedi, A., Zarrintan, A., Mohebbi, I., Pirnejad, H., Khademvatani, K., Ashkavand, Z., Forughi, P., Arasteh, A., Attari, J.A.: Predictors of coronavirus disease 19 (COVID-19) pneumonitis outcome based on computed tomography (CT) imaging obtained prior to hospitalization: a retrospective study. *Emerg. Radiol.* **27**(6), 653–661 (2020). <https://doi.org/10.1007/s10140-020-01833-x>
 77. Upchurch, C.P., Grijalva, C.G., Wunderink, R.G., Williams, D.J., Waterer, G.W., Anderson, E.J., Zhu, Y., Hart, E.M., Carroll, F., Bramley, A.M., Jain, S., Edwards, K.M., Self, W.H.: Community-acquired pneumonia visualized on CT scans but not chest radiographs. *Chest* **153**(3), 601–610 (2018). <https://doi.org/10.1016/j.chest.2017.07.035>
 78. Kittleson, M.M., Kobashigawa, J.A.: Management of the ACC/AHA stage d patient. *Cardiol. Clin.* **32**(1), 95–112 (2014). <https://doi.org/10.1016/j.ccl.2013.09.004>
 79. de Nattes, T., Litzler, P.-Y., Gay, A., Nafeh-Bizet, C., François, A., Guerrot, D.: Hemolysis induced by left ventricular assist device is associated with proximal tubulopathy. *PLoS ONE* **15**(11), 0242931 (2020). <https://doi.org/10.1371/journal.pone.0242931>
 80. Desai, P.S., Fanous, E.J., Tan, W., Lee, J., Trinh, T., Rafique, A.M., Parikh, R.V., Press, M.C.: Trajectory of cardiac catheterization for acute coronary syndrome and out-of-hospital cardiac arrest during the COVID-19 pandemic. *Cardiol. Res.* **12**(1), 47–50 (2021). <https://doi.org/10.14740/cr1149>
 81. Liu, Y.-H., Cao, Z.-Y., Dai, Y.-N., Zeng, L.-H., Zhang, Y.-S., Fan, H.-L., Duan, C.-Y., Tan, N., He, P.-C.: Association of proton pump inhibitor and infection and major adverse clinical events in patients with ST-elevation myocardial infarction: A propensity score matching analysis. *Front. Med.* **9**, 882341 (2022). <https://doi.org/10.3389/fmed.2022.882341>
 82. Giacomelli, A., Ridolfo, A.L., Pezzati, L., Oreni, L., Carrozzo, G., Beltrami, M., Poloni, A., Caloni, B., Lazzarin, S., Colombo, M., Pozza, G., Pagano, S., Caronni, S., Fusetti, C., Gerbi, M., Petri, F., Borgonovo, F., D'Aloia, F., Negri, C., Rizzardini, G., Antinori, S.: Mortality rates among COVID-19 patients hospitalised during the first three waves of the epidemic in Milan, Italy: a prospective observational study. *PLoS ONE* **17**(4), 0263548 (2022). <https://doi.org/10.1371/journal.pone.0263548>
 83. Elsaid, K., Olabi, V., Sayed, E.T., Wilberforce, T., Abdelkareem, M.A.: Effects of COVID-19 on the environment: an overview on air, water, wastewater, and solid waste. *J. Environ. Manag.* **292**, 112694 (2021). <https://doi.org/10.1016/j.jenvman.2021.112694>
 84. Yuan, J., Li, M., Lv, G., Lu, Z.K.: Monitoring transmissibility and mortality of COVID-19 in Europe. *Int. J. Infect. Dis.* **95**, 311–315 (2020). <https://doi.org/10.1016/j.ijid.2020.03.050>
 85. Chen, Y., Zhang, Y., Xu, Z., Wang, X., Lu, J., Hu, W.: Avian influenza a (H7N9) and related internet search query data in china. *Sci. Rep.* **9**(1), 10434 (2019). <https://doi.org/10.1038/s41598-019-46898-y>
 86. Wong, S., Pabbaraju, K., Pang, X.L., Lee, B.E., Fox, J.D.: Detection of a broad range of human adenoviruses in respiratory tract samples using a sensitive multiplex real-time PCR assay. *J. Med. Virol.* **80**(5), 856–865 (2008). <https://doi.org/10.1002/jmv.21136>
 87. Scalia, G., Halonen, P.E., Condorelli, F., Mattila, M.L., Hierholzer, J.C.: Comparison of monoclonal biotin-avidin enzyme immunoassay and monoclonal time-resolved fluoroimmunoassay in detection of respiratory virus antigens. *Clin. Diagn. Virol.* **3**(4), 351–359 (1995). [https://doi.org/10.1016/0928-0197\(94\)00050-5](https://doi.org/10.1016/0928-0197(94)00050-5)
 88. Malik, M., Kunze, A.-C., Bahmer, T., Herget-Rosenthal, S., Kunze, T.: SARS-CoV-2: viral loads of exhaled breath and oronasopharyngeal specimens in hospitalized patients with COVID-19. *Int. J. Infect. Dis.* **110**, 105–110 (2021). <https://doi.org/10.1016/j.ijid.2021.07.012>
 89. Hrdy, O., Vrbica, K., Strazevska, E., Suk, P., Souckova, L., Stepanova, R., Sas, I., Gal, R.: Comparison of continuous versus intermittent enteral nutrition in critically ill patients (COINN): study protocol for a randomized comparative effectiveness trial. *Trials* **21**(1), 1–10 (2020). <https://doi.org/10.1186/s13063-020-04866-2>
 90. Han, Y., Luo, Z., Zhai, W., Zheng, Y., Liu, H., Wang, Y., Wu, E., Xiong, F., Ma, Y.: Comparison of the clinical manifestations between different age groups of patients with overseas imported COVID-19. *PLoS ONE* **15**(12), 0243347 (2020). <https://doi.org/10.1371/journal.pone.0243347>

91. Dhanasekararaja, P., Soundarrajan, D., Kumar, K.S., Pushpa, B.T., Rajkumar, N., Rajasekaran, S.: Aggressive presentation and rapid progression of osteonecrosis of the femoral head after COVID-19. *Indian J. Orthop.* **56**(7), 1259–1267 (2022). <https://doi.org/10.1007/s43465-022-00635-2>
92. van Schaik, M.L., Duijkers, R., Paternotte, N., Jansen, R., Roze-meijer, W., van der Reijden, W.A., Boersma, W.G.: Feasibility of a quantitative polymerase chain reaction assay for diagnosing pneumococcal pneumonia using oropharyngeal swabs. *Mol. Biol. Rep.* **46**(1), 1013–1021 (2018). <https://doi.org/10.1007/s11033-018-4558-0>
93. Machado-Rivas, F., Gallo-Bernal, S., Briggs, D., Panykh, O., Gee, M.S.: Characterization of pediatric imaging trends and likelihood of exam cancellation in the COVID-19 pandemic. *Acad. Radiol.* **29**(4), 508–513 (2022). <https://doi.org/10.1016/j.acra.2021.12.011>
94. Robles, M.C., Corches, C.L., Bradford, M., Rice, T.S., Sukul, D., Springer, M.V., Bailey, S., Oliver, A., Skolarus, L.E.: Understanding and informing community emergency cardiovascular disease preparedness during the COVID-19 pandemic: Stroke ready. *J. Stroke Cerebrovasc. Dis.* **30**(2), 105479 (2021). <https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105479>
95. Willcox, M.E., Compton, S.J., Bardy, G.H.: Continuous ECG monitoring versus mobile telemetry: A comparison of arrhythmia diagnostics in human-versus algorithmic-dependent systems. *Heart Rhythm O2* **2**(6), 543–559 (2021). <https://doi.org/10.1016/j.hroo.2021.09.008>
96. Boelens, J.J., Wynn, R.F., O'Meara, A., Veys, P., Bertrand, Y., Souillet, G., Wraith, J.E., Fischer, A., Cavazzana-Calvo, M., Sykora, K.W., Sedlacek, P., Rovelli, A., Uiterwaal, C.S.P.M., Wulfraat, N.: Outcomes of hematopoietic stem cell transplantation for hurler's syndrome in Europe: a risk factor analysis for graft failure. *Bone Marrow Transpl.* **40**(3), 225–233 (2007). <https://doi.org/10.1038/sj.bmt.1705718>
97. Bai, L.Y., Chiou, T.J., Liu, J.H., Yen, C.C., Wang, W.S., Yan, M.H., Hsiao, L.T., Chao, T.C., Chen, P.M.: Hematopoietic stem cell transplantation for severe aplastic anemia—experience of an institute in Taiwan. *Ann. Hematol.* **83**(1), 38–43 (2003). <https://doi.org/10.1007/s00277-003-0781-3>
98. Gaziev, D., Galimberti, M., Lucarelli, G., Polchi, P., Giardini, C., Angelucci, E., Baronciani, D., Sodani, P., Erer, B., Biagi, M.D., Andreani, M., Agostinelli, F., Donati, M., Nesci, S., Talevi, N.: Bone marrow transplantation from alternative donors for thalassemia: HLA-phenotypically identical relative and HLA-nonidentical sibling or parent transplants. *Bone Marrow Transplant.* **25**(8), 815–821 (2000). <https://doi.org/10.1038/sj.bmt.1702242>
99. Southwell, B.G., Eder, M., Finnegan, J., Luepker, R.V., Duval, S., Russell, C., Graves, R.N., Namboodri, A.: News coverage and online advertising effects on patient-led search for aspirin, heart health, and stroke information and educational tool use. *Patient Educ. Couns.* **104**(3), 663–665 (2021). <https://doi.org/10.1016/j.pec.2020.08.004>
100. Scolari, F.L., Hastenteufel, L.C.T., Einsfeld, L., Bueno, J., Orlandin, L., Clausell, N., Goldraich, L.A.: Impact of COVID-19 infection among heart transplant recipients: a southern Brazilian experience. *Front. Med.* **9**, 814952 (2022). <https://doi.org/10.3389/fmed.2022.814952>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.