Capturing research literature attitude towards Sustainable Development Goals: an LLM-based topic modeling approach

Francesco Invernici[1], Francesca Curati[1], Jelena Jakimov[1], Amirhossein Samavi[1], Anna Bernasconi[1*]

[1]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Via Ponzio 34/5, Milano, 20133, Italy.

*Corresponding author(s). E-mail(s): anna.bernasconi@polimi.it

Abstract

The world is facing a multitude of challenges that hinder the development of human civilization and the well-being of humanity on the planet. The Sustainable Development Goals (SDGs) were formulated by the United Nations in 2015 to address these global challenges by 2030.

Natural language processing techniques can help uncover discussions on SDGs within research literature. We propose a completely automated pipeline that 1) fetches content from academic literature and prepares datasets dedicated to five groups of SDGs; 2) performs topic modeling, a statistical technique used to identify topics in large collections of textual data; and 3) enables topic exploration through keywords-based search and topic frequency time series extraction.

For topic modeling, we leverage the stack of BERTopic scaled up to be applied on large corpora of textual documents (we find hundreds of topics on hundreds of thousands of documents), introducing i) a novel LLM-based embeddings computation for representing scientific abstracts in the continuous space, and ii) a hyperparameter optimizer to efficiently find the best configuration for any new dataset. We additionally produce the visualization of results on interactive dashboards reporting topics' temporal evolution. Results are made inspectable and explorable, contributing to the interpretability of the topic modeling process.

The proposed LLM-based topic modeling pipeline allows users to capture insights on the evolution of the attitude toward SDGs within scientific abstracts in the 2006-2023 time span. All the results are reproducible by using our system; the workflow can be generalized to be applied at any point in time to any large corpus of text data.

Keywords: Topic modeling, Text embeddings, LLM, Sustainable Development Goals, Textual data analysis, Temporal trends

Introduction

Sustainable Development Goals (SDGs) are 17 United Nations' global objectives identified to address some of the biggest challenges of human civilization [1]. These goals include issues such as gender equality and education, poverty and hunger, health, and climate change. Each goal is designed to address a specific issue or a set of strongly related issues; however, all goals should work together to create a better and more sustainable future for humanity. We use keywords that describe SDGs as our point of access to a scientific literature landscape that is typically very vast and for which easy, flexible exploration is problematic. We accessed academic research outcomes through the Elsevier Scopus database, which stores a rich content of abstracts along with their metadata, via their RESTful API [2], focusing on the years' range 2006-2023.

For the analysis, we follow an unsupervised statistical approach based on natural language processing, specifically focused on topic modeling [3]. Unsupervised Topic Modeling is used to discover and analyze latent topics within a document, without leveraging pre-existing labels or supervision. This method works under the assumption that each document represents a single topic, or at least that one topic is preponderant, so as to exclude encompassing multiple topics at the same time.

In our work, we frame topic modeling as a clustering task [4] over the latent space of embeddings, differently from other approaches that train end-to-end models for topic modeling, either based on classical methods [5] or language models [6]. Egger and Yu [7] surveyed four topic modeling techniques, namely latent Dirichlet allocation, non-negative matrix factorization, Top2Vec, and BERTopic [8]. In line with their analysis and the suggestions of a more recent survey by Abdelrazek *et al.* [9], we selected the neural model BERTopic to implement our approach for topic modeling from document clustering. Neural topic models are particularly appropriate to guarantee scalability (both in terms of model and data), flexibility (i.e., the ability to adapt to different tasks like, in our case, dynamic topic modeling), and the possibility of being embedded in end-to-end data pipelines; these aspects are particularly important in our scenario. Thanks to these characteristics, BERTopic has already been successfully used in social sciences [10–12], while other architectures were more popular in the previous years [13, 14].

We propose to use BERTopic in a different domain: SDGs have generated much interest as a key to understanding the general attitude (both research-driven and general-public) toward high-stakes themes related to many transverse continents and socioeconomic groups. SDGs have been investigated through several different techniques either comprehensively [15, 16] or individually [17, 18]. Some work focused on extracting SDG-related topics of discussion on social media comment threads [19, 20] or on online news [21]. Saheb *et al.* [22] targeted a small corpus of 182 research abstracts focused on a specific area (artificial intelligence solutions for sustainable energy), while Raman *et al.* [23] selected a small corpus of 448 research abstracts on green/sustainable AI. Even if, to a small extent, the employed techniques and the domain of interest overlap with our focus, all mentioned works significantly differ from ours in the scale of their elaboration. Indeed, typically, they are based on small datasets (a few hundred documents) and consequently build very small topic models (e.g., [22] identifies 8 topics, [21] 10 topics, [20] 17 topics, and [23] 5 topics). The work by Smith *et al.* [24]

is more similar to ours in spirit; here, about 30k abstracts related to SDG 3 (Good Health and Well-being) are analyzed, and about 200 topics are identified. Our innovation is to make this kind of analysis completely reproducible on any large dataset and to expose it on a user-friendly interface. In parallel, this allows us to complement previous efforts by providing a complete overview of all SDG-related keywords.

Here, we propose to adopt an LLM-based topic modeling pipeline named TETYS (standing for 'Topics Evolution That You See'), which has the following characteristics:

- it can be run on big-text datasets in a completely automated mode;
- it enhances BERTopic [8] default configuration with an LLM-based embedding computation;
- it employs an innovative parameters' optimization mechanism that randomly searches the parameters' space to optimize a Density-Based Clustering Validation (DBCV) score – thus making running the same pipeline on multiple big datasets practical;
- it allows us to build interpretable topic models for big corpora of complex (i.e., scientific/technical) text documents; and
- it builds a Web platform providing a complete overview of the topics, with interactive exploration of topics' representation over time.

In this manuscript, we deliver the results of applying TETYS on five groups of documents (called macro-areas) that concern a collection of SDG-related keywords (respectively on Basic Human Needs and Well-being; Environmental Sustainability; Economic Development and Employment; Equality and Social Inclusion; and Global Partnerships and Peace). The pipeline was optimized to run on each of these groups of documents. Our TETYS platform, exposed at http://gmql.eu/tetys/, is a Web interface that makes results explorable for any stakeholder.

Materials and Methods

We overview the preparation of the text corpora and then describe the TETYS pipeline, divided into its sub-pipeline for building and fitting the topic model and its sub-pipeline dedicated to topic exploration artifacts.

Datasets preparation

We extracted abstracts and metadata of research publications from Scopus, one of the largest repositories for academic peer-reviewed documents, including journal articles and conference proceedings. Scopus was established by the publisher Elsevier [25] and is considered relatively more comprehensive than Web of Science [26]. Scopus has enabled many text mining approaches, also using topic modeling [27] in very specific domains such as personal information privacy [28] or public procurement [29].

Next, we detail how we grouped the SDGs to define five overarching macro-areas that include a significant number of abstracts to be analyzed with our approach. Then, we describe the strategy to retrieve abstracts and their metadata from Scopus API and, finally, we detail the data cleaning process.

M	Included SDGs	Keywords	#abst.
M1	1 No Poverty 2 Zero Hunger 3 Good Health and Well-being 4 Quality Education 6 Clean Water and Sanitation	Poverty alleviation; Food security; Public health; Education access; Water quality; Sanitation infrastructure; Healthcare provision.	333,901 (original) **320,798 (final)**
M2	7 Affordable and Clean Energy 11 Sustainable Cities and Communities 12 Responsible Consumption and Production 13 Climate Action 14 Life Below Water 15 Life on Land	Renewable energy; Urban sustainability; Sustainable consumption; Climate change mitigation; Marine biodiversity; Ecosystem conservation; Energy efficiency.	399,922 (original) **339,949 (final)**
M3	8 Decent Work and Economic Growth 9 Industry, Innovation, and Infrastructure	Economic growth; Innovation ecosystems; Infrastructure development; Entrepreneurship support; Industrialization strategies; Industrial Innovation; Labor market dynamics.	50,482 (original) **41,218 (final)**
M4	5 Gender Equality 10 Reduced Inequality	Gender empowerment; Social equity; Inclusive policies; Women's rights; Minority rights; Income inequality; Social justice.	33,769 (original) **25,017 (final)**
M5	16 Peace, Justice, and Strong Institutions 17 Partnerships for the Goals	Legal institutions; International cooperation; Peace efforts; Sustainable development cooperation; Global governance; Justice systems; Multilateral agreements.	56,275 (original) **33,769 (final)**

Table 1: Description of five macro-areas (M) grouping the SDGs. M1 = Basic Human Needs and Well-being; M2 = Environmental Sustainability; M3 = Economic Development and Employment; M4 = Equality and Social Inclusion; M5 = Global Partnerships and Peace. Numbers of abstracts are reported as i) number of original abstracts, and ii) number of abstracts after deduplication and data cleaning (in bold type).

Definition of SDG macro-areas

We grouped the initial SDGs into macro-areas to make it easier to identify big topics, trends, and relationships, thereby providing a clearer picture of sustainable development as a whole. We chose not to exceed some hundred thousand documents, as this proved effective in previous works [30] and is recommended in BERTopic documentation [31].

We queried ChatGPT [32] with an appropriately crafted prompt asking to group the 17 SDGs into 5 macro-areas, each concisely described through 7 keywords, which are likely to be selected by the authors of the scientific papers; the output was carefully checked by dedicated domain experts for each macro-area, to avoid potential biases in keyword selection [33] – final keywords are described in Table 1.

Data and metadata retrieval

Due to its extensive coverage and being one of the most trusted databases in the academic field, we selected Scopus as our data analysis source.

We accessed programmatically the corpus of literature data provided by Scopus, by employing its Academic Research APIs [34] through two endpoints:

(1) Scopus Search API enables users to submit queries to the Scopus index and retrieve relevant metadata in user-specific text formats and the link to the corresponding abstracts.

(2) Abstract Retrieval API allows us to retrieve an abstract after searching its link using the first endpoint.

The endpoint (1) uses a `query` parameter that allows a boolean search with field restriction; we employ the fields `pubstage` set to “final” to exclude preprints, `pubyear` starting from 2006 up to 2023 included, `language` to include English-language abstracts, and `key` for specifying keywords related to the abstracts (contained in author-specified keywords or automatically-indexed keywords). By enclosing terms to be searched in double quotation marks, we employ a similarity-based “search for a loose or approximate phrase” exposed by API. While (1) fetches the identifiers of documents of interest, the actual abstracts with their metadata are retrieved by calling (2), one paper at a time.

Data cleaning

For each macro-area, we obtained a dataset of ten-to-hundred thousands of documents (see numbers in the last column of Table 1), each equipped with a set of 20 metadata fields. We removed from the metadata set the rows that did not have a corresponding abstract document, or that lacked a Digital Object Identifier (DOI), title, or publication date (see Figure 1 for the distribution of missing values per each metadata field).

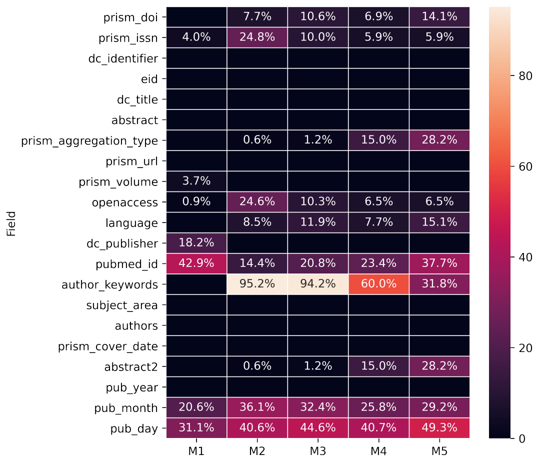

Fig. 1: Heatmap representing the percentage of missing metadata API fields (rows) per macro-area (columns). Cells with no number indicate that the metadata field is present in all records. Lighter colors indicate the metadata field is heavily lacking.

Then, we performed data deduplication for rows with the same digital object identifier and/or internal Scopus identifier. Finally, we enforced the time window of interest for the publication date, keeping only abstracts published between 2006 and 2023 (included), and converted the dates into the Python DateTime format. At the end of the stage, we enforced the selection of abstracts written in English. Refer again to Table 1 (last column, second value) for counts of papers after the deduplication.

In Figure 2, we present the distribution of abstracts published each year, in the considered period, for each macro-area (M1 to M5). The trend shows a general increase, which confirms aspects such as the increased global awareness of sustainability issues, the development of technology, and the growing number of researchers. Interestingly, M1 (Basic Human Needs and Well-being) and M4 (Equality and Social Inclusion) show a spike during the period 2020-2023, likely due to the COVID-19 pandemic, while M5 (Global Partnerships and Peace) exhibits a less right-skewed distribution w.r.t. others.

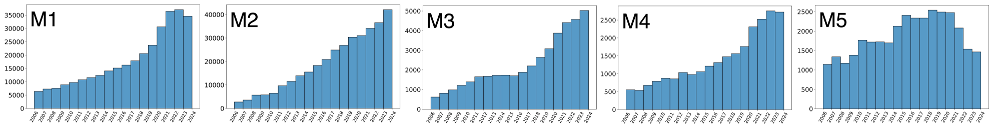

Fig. 2: Data distribution over the years for all five macro-areas.

TETYS Pipeline

Our pipeline consists of two sub-pipelines (see Figure 3), one for topic modeling and one for topic exploration. The first sub-pipeline is dedicated to building a solid topic model and fitting it to the current dataset, arranging for an interpretable model representation. The second sub-pipeline is concerned with extending the information within the topic model, allowing exploration via keyword-based search and adding simple distance metrics and time series on which statistical tests can be drawn.

Every step in the two sub-pipelines is performed on five different datasets (each based on one of the previously defined macro-areas); each process produces, as a result, a topic model that can be explored in a Web-based dashboard. The pipeline instances are completely separated; when appropriate, others could be generated independently from one another as the data architecture, the backend, and the frontend are general and can be configured based on need.

Topic modeling

We base our work on BERTopic [8], a topic modeling framework that leverages six steps to achieve unsupervised latent topic identification and textual representation learning. It requires 1) converting documents into embeddings, 2) reducing the dimensionality of the embeddings; 3) clustering the reduced embeddings; 4) tokenizing documents; 5) using a word-weighting scheme; and 6) optionally tuning the obtained topic representation.

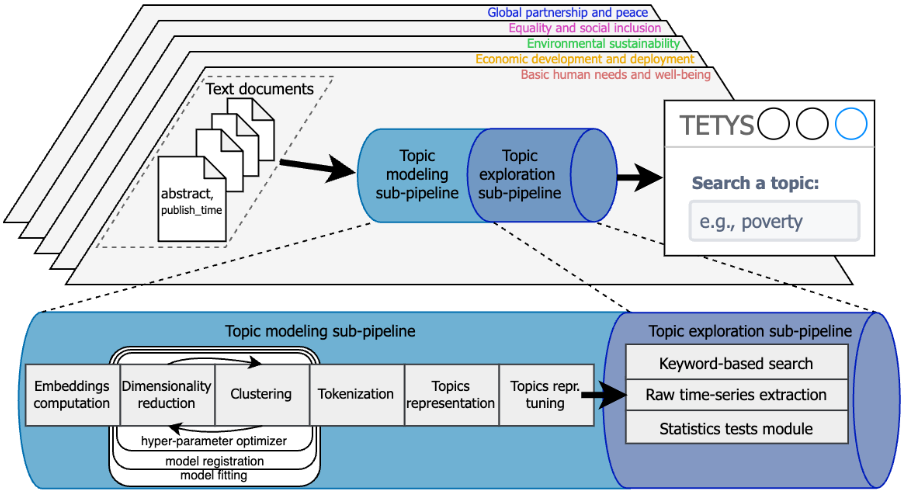

Fig. 3: TETYS pipeline architecture.

The default configuration employs, respectively, in the first five steps: the sentence-transformer BERT (SBERT [35]); the Uniform Manifold Approximation and Projection (UMAP) dimension reduction technique [36]; the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [37]; the word tokenizer CountVectorizer [38]; and a class-based term frequency–inverse document frequency (c-TF-IDF) model [39].

Since BERTopic's first conception, several enhancements have been introduced. Thanks to its modular structure and the possibility of completely customizing its pipeline, we searched for the best possible configuration given each macro-area domain and dataset at hand. With respect to a standard configuration of the BERTopic stack, TETYS introduces several contributions:

- we replaced the default SBERT with a Large Language Model (LLM) for the computation of embeddings;
- we designed an innovative systematic *optimizer* for the two hyperparameter-tuning steps of the pipeline (dimensionality reduction of embeddings and their clustering) – this mechanism allows us to evaluate multiple configurations with different parameters, quickly converging to a (local) optimal one;
- we implemented a model *registration* functionality, to persist the output of the optimization phase and the consequent model fitting.

In the following, we discuss more in-depth these three novelties, followed by a brief description of the classical steps offered by BERTopic (including the tokenization and the representation of topics with its tuning).

LLM-based embeddings computation

In order to learn the latent topic structure of a dataset, we map each abstract to a point in an embedding representation, leveraging LLMs.

On June 20th, 2024, we inspected the Massive Text Embedding Benchmark (MTEB) leaderboard [40] and selected the general-purpose model that maximized the average performance over a set of criteria listed by the leaderboard, while satisfying the memory constraints of our setup (more details in Results, '*Execution and time performances*').

We selected the second release of the Salesforce embedding model (SFR-Embedding-2 R LLM [41]). The model was trained on abstracts concatenated with the corresponding paper title, producing 4096-dimensional embedding representations. This choice replaced the default component SBERT proposed in [8] (which featured a much lower dimensional space). The selected SFR model is known to bring enhancements across all downstream tasks, with particularly notable improvements in clustering and classification tasks, making it a top-performance model on the HuggingFace MTEB benchmark leaderboard, at the time of our development.

In the absence of documentation for the SFR-Embedding-2 R model, we referred to the SFR-Embedding-Mistral [42] model, its closest documented ancestor model. This is trained on a variety of data from different tasks. For clustering tasks, it utilizes data sourced from the preprint repositories arXiv, bioRxiv, and medRxiv, while applying filters to exclude development and testing sets.

Loading the SFR-Embedding-2 R model and dataset into GPU memory was non-trivial. Due to its large size, it was impossible to simultaneously load the model and dataset and encode the abstracts into embedding vectors. We exploited the `transformers.pipelines` API [43] and its built-in mechanisms for lazy loading and on-demand processing, which efficiently manage memory usage. The pipeline processes the data in manageable chunks, not requiring the whole data to be loaded in the GPU memory, only the necessary parts of the model and data are loaded when needed.

Hyperparameter optimizer

In order to evaluate the goodness of the intermediate topic models that are generated (each one based on a specific configuration of the parameters set), we introduce an optimization mechanism. In our previous work [30], we had proposed to optimize hyperparameters by performing a grid search, i.e., trying all the possible combinations to maximize the clusters' one-to-one relative density connection using the Density-Based Clustering Validation (DBCV) [44] score (spanning -1 for lowest quality to 1 for highest quality). The DBCV score is a performance metric for clustering algorithms; however, we leveraged this metric for all our hyperparameters as DBCV not only assesses the quality of the clusters but also provides valuable insights into the cohesiveness and separation of topics. Note that, here, our guiding principle was to select a validation metric that aligns with the nature of our clustering method and the constraints of our unsupervised, model-agnostic setup.

Clearly, with grid search, we can always achieve the optimal configuration, even if at the cost of spending a significantly longer time. Here, we experiment with a random search, which involves sampling a fixed number of hyperparameter combinations (much smaller than the total number of possible configurations). With this option, we obtain satisfactory results, allowing us to scale our approach up to any number of TETYS execution pipelines; specifically, we propose the following steps:

(1) We generate the parameters' space including four parameters for dimensionality reduction (UMAP) and four parameters for clustering the embeddings (see Table 2 for the parameters ranges including the tested ⟨ start, end, step ⟩ scheme).
(2) We define a finite number of random search steps (empirically, we appreciated that –once around the 100th step– the local maximum solution found by the random search typically approximates the global maximum one found with the grid search approach).
(3) Until the number of steps identified in (2) is reached, we experiment with one configuration at a time as follows:
   (i) Draw one configuration in the parameters' space (see Table 2).
   (ii) Run UMAP and HDBSCAN with the selected configuration on a validation subset of the current dataset (a randomly sampled 20% of the dataset).
   (iii) Calculate the corresponding DBCV score.
   (iv) If the DBCV score **is not** the current best (local) maximum one, discard the configuration and proceed to the next one. If **it is** the current best one, proceed with *Model registration* and *Model fitting*.
(4) The model with the highest DBCV (once the random search steps are concluded) is considered the best one and employed for the following BERTopic steps.

Step	Parameter name	Parameter range	M1	M2	M3	M4	M5
UMAP	n_neighbors	(1, 100, 5)	20	20	100	50	100
	min_dist	(0, 1, 0.05)	0	0	0	0	0
	n_components	(5, 50, 5)	5	10	10	28	35
	metric	('euclidean')			— 'euclidean' —		
HDBSCAN	min_samples	(10, 100, 10)	75	75	10	10	15
	min_cluster_size	(25, 100, 5)	25	25	25	25	25
	cluster_selection_method	('eom', 'leaf')			— 'eom' —		

Table 2: For each step and parameter, we report the value ranges ⟨start, end, step⟩ tested by the *optimizer* of the hyperparameters of the dimensionality reduction and the clustering steps. The last five columns report, for each of the five macro-areas, which parameters configuration led to the best DBCV performance, thus used for the model fitting.

In the best run for each macro-area, we obtained DBCV scores of, respectively, 0.52, 0.76, 0.39, 0.46, and 0.38 using the parameters' values reported in the last five columns of Table 2.

Model fitting and registration

Once the optimizer has selected the final parameters set, we run the *Model registration* and *Model fitting* components.

During *Model registration* we save the model in two formats: (i) `pickle`, a binary object for quality checks during this optimization process; (ii) `safetensors`, a PyTorch model [45] ready to be used for future inference on new data that the model has not seen. This component is designed to address the challenges posed by the stochastic nature of the HDBSCAN algorithm. It ensures that the best model found during hyperparameter optimization is saved immediately and preserved for future use. The main advantage comes from the fact that reinitializing the BERTopic model, even with the same hyperparameters, can yield different results, due to the randomness involved in HDBSCAN initialization. This variability can lead to a model that underperforms compared to the one identified during the hyperparameter optimization phase. By incorporating the *registration* component, we not only ensure that the integrity of the best-performing model is preserved, but also that any subsequent analysis or application of the model is based on a consistent and reproducible version. A disadvantage of this approach is that we need to store multiple models; since we do not know in advance which model will perform the best among all the models found, we need to keep track of several versions, consistently increasing memory usage. Additionally, the time required to fit the model can be significant for certain parameter configurations. To address this issue and avoid saving models that will not be useful, we added the possibility of saving the model only if 1) it is the best model found thus far, and 2) its DBCV score is greater than a 0.30 threshold limit, which we identified empirically through manual inspection of preliminary results.

Then, *Model fitting* involves exploiting the hyperparameters corresponding to the current DBCV score. With these parameters, we run UMAP and HDBSCAN on the whole dataset (100%). Note that, while in UMAP the parameters correspond to hyperparameters observed during validation, for clustering we need to fit the model –with its selected hyperparameters– to the data and compute the actual parameters (e.g., number of clusters, center of clusters, etc.). As an outcome of running this component, we build the final models, on which subsequent steps of BERTopic are applied.

Topic representation and tuning

The three remaining steps in the BERTopic pipeline contribute to achieving interpretable, synthetic representations of topics. The first step involves an abstract vectorization (performed with the default `scikit-learn` [38] CountVectorizer).

Second, we fit the c-TF-IDF model with the `reduce_frequent_words` parameter set, which considers the square root of the normalized frequency of the terms (i.e., words). With this model, we obtain the most relevant terms (i.e., topics) per class, with their frequency. This corresponds to a textual, human-understandable representation for each cluster. The most important topics can be retrieved using the TF-IDF representations.

Third, to improve our topic representation, we target the reduction of similar keyword repetition, such as those with the same root word or variations (e.g., singular and plural forms of the same word). More distinct and meaningful keywords, without redundancy, ensure that each keyword adds value to the overall representation and understanding of the topic. For this, we employ Maximal Marginal Relevance (MMR), which selects keywords for topic representation, based on their relevance score and their dissimilarity to previously selected items. The goal is to maximize the relevance score while minimizing redundancy. MMR allows us to get a clearer, more accurate picture of the keywords, where topics are more distinct and meaningful, while making them easier to understand and interpret.

Topic exploration

While the first sub-pipeline essentially allows us to systematize the customization of a BERTopic-like process, the second sub-pipeline creates a set of support data structures and representations useful to make topic exploration possible on dedicated visual dashboards.

First, we adopt the `word_cloud` [46] package to generate word clouds with the most frequent terms of each topic, thereby providing a visual representation to inspect the topic content.

Second, we enable a keyword-based search, by exploiting the `find_topics` function implementation in BERTopic [8], which essentially allows inputting a simple search term (possibly including spaces) to retrieve a list of similar topics equipped with their score of similarity w.r.t. the input term.

Third, we compute per-topic time-series, representing the counts of papers published during the observed period 2006-2023. Our approach builds time-series using a parametric number of months in each *time bin*. For each abstract, we consider the date when it was published and the topic it belongs to; then, given a time granularity (1-month, 3-month, 6-month, or year), we compute bins corresponding to the requested timeframe. As an output, we obtain tuples of the form $\langle topic\_id, (bin\_id, start\_date), \#abstracts\_in\_bin \rangle$. This method resembles the Dynamic Topic Modeling techniques proposed within BERTopic [8]; essentially, we add run-time computation of features that are useful for analyzing time-series: i) binning; ii) absolute/relative frequency (we normalized the count w.r.t. the number of total abstracts published in that bin); and iii) ranking. In this way, we can interpret the values as pointwise measures of the intensities of the topic, as other previous works on dynamic topic modeling [3]. Taking advantage of these time-series, we generate line plots for the counts of abstracts per bin.

Finally, we implement two statistical tests. To check if the trend difference of *two* periods of the same topic is significant, we use the non-parametric Kruskal-Wallis test [47], typically employed for comparing sample medians (checking if two groups are sampled from the same population). The test produces a p-value, enabling the acceptance or rejection of the simple null hypothesis "there is no significant difference in the topic representation in periods T1 versus T2" (we adopt the library `SciPy.stats.kruskal` [48]). We use the 5% p-value as the threshold for significance; lower p-values allow the rejection of the null hypothesis [47]. To check if the trend difference of *multiple* periods of the same topic is significant, we apply Kruskal-Wallis to all intervals and verify if at least one interval is significantly different from the others. To understand which interval deviates from others, we use the Dunn test [49] with multiple testing corrections.

Results

In the following, we describe the five obtained topic models, evaluate them by comparison with those obtained using a baseline pipeline, and finally propose the topic exploration dashboard.

Extracted topics overview

In the five macro-areas we found, respectively, 550 topics (Basic Human Needs and Well-being), 856 topics (Environmental Sustainability), 181 topics (Economic Development and Employment), 136 topics (Equality and Social Inclusion), and 167 topics (Global Partnerships and Peace). The number of identified topics is roughly proportional to the number of abstracts for each macro-area (see Table 1). M1 and M2 are the biggest macro-areas, as they also include more Sustainable Development Goals compared to M3-M5.

For a quick overview, in Figure 4 we present diagrams illustrating the distribution of topics, only including the top 30 topics based on their abstracts' counts. The y-axis maximum values are 5,000 for M1, 2,500 for M2, 1,400 for M3, 600 for M4, and 1,400 for M5.

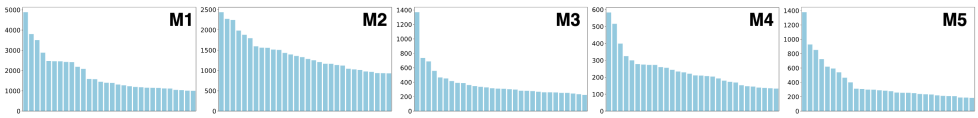

Fig. 4: Distribution of the 30 largest topics based on the number of abstracts associated with each of them for each macro-area and configuration.

Figure 5 shows, for each macro-area, its intertopic distance map. This map places the topics in two dimensions, where the Euclidean distance between any two of them represents their similarity: the closer they are, the more semantically similar they are. Topics are represented as circles, and their size depends on the number of abstracts they gather. Due to the projection from a higher-dimensional space to two dimensions, we observe several overlaps in the map. In the figure, the five largest topics for each area are connected to their corresponding word-clouds.

In M1 (Basic Human Needs and Well-being), the 'pollutants' and 'bacteriological sanitation' topics are likely related to the *Clean Water and Sanitation* goal (SDG 6). Topics on 'cancer' and 'smoking' are closely connected with the *Good Health and Well-being* goal (SDG 3). The topic related to 'health and diets' is probably derived from publications related to *Zero Hunger* (SGD 2);

In M2 (Environmental Sustainability), the terms 'workloads, virtualisation and energy-aware' seem related to the optimization of computing resources, and probably are in connection with energy consumption in data centers. The 'electric vehicle and charging' topic can also be related to the same goals. Hydrogen is considered a clean energy carrier [50] and is often connected with clean and renewable energy, thus, topics related to it can be connected to both *Affordable and Clean Energy* and *Responsible Consumption and Production goals* (SDGs 7 and 12). The topic with the 'watersheds, urbanising and ecosystems' terms seems closely related to the *Sustainable Cities and Communities* goal (SDG 11). The terms 'levelized, microgrids, and hybrid' are often associated with sustainable energy problems and solutions, which are closely related to the *Affordable and Clean Energy* goal (SDG 7).

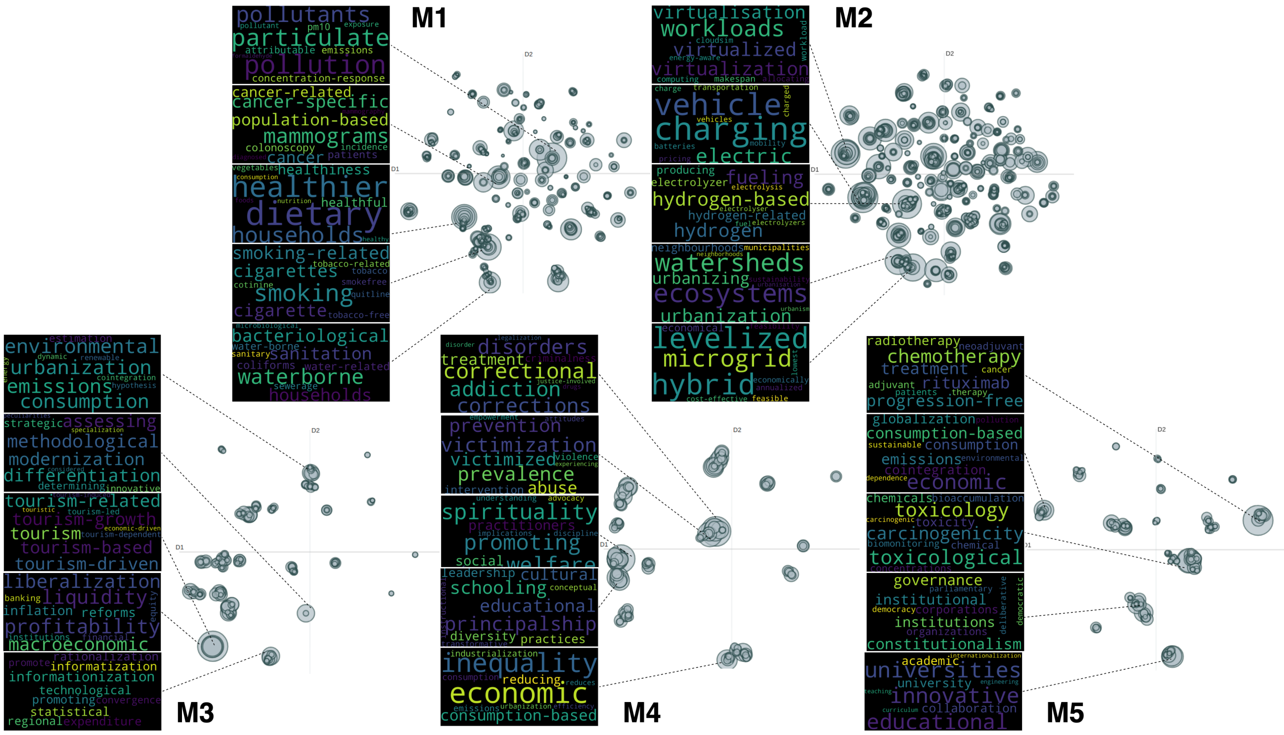

Fig. 5: The biggest and most interesting topics from the five macro-areas

In M3 (Economic Development and Employment), the topics on 'liquidity, macroeconomic, and profitability' and on 'tourism-related' are connected to the *Decent Work and Economic Growth* goal (SDG 8). Instead, the topics on 'methodological modernization' and 'environmental urbanization' appear related to the *Industry, Innovation, and Infrastructure* goal (SDG 9).

In M4 (Equality and Social Inclusion), the topic of 'victimization and abuse' is related to the *Gender Equality* goal (SDG 5), while other topics can be connected more generally to the *Reduced Inequality* goal (SDG 10).

In M5 (Global Partnerships and Peace), topics look very versatile, possibly because the concept of "partnership" can include many different ideas and realizations.

Execution and time performances

The primary computational cost is in embedding generation and the subsequent model fitting phase; all remaining processing is negligible in comparison. To run our topic modeling sub-pipeline, we employ a virtual machine equipped with an NVIDIA A100 (40GB) GPU [51], 32 virtual CPUs, 64 GB RAM, 60 GB SSD, and 500 GB HDD.

Note that the NVIDIA A100 (40GB) represents, *de facto*, the minimal commodity-grade GPU suitable for LLM inference. In cases where equivalent or superior hardware is unavailable, lighter models, such as MiniLMs and ParagraphLMs - available in the

SBERT library [35] - can operate on lower-end configurations, including CPU-only setups. However, these models are architecturally outdated and less performing.

Empirical evidence suggests that LLM inference with reasonable memory for context management requires approximately 2.5 times the full-precision model size in VRAM. In our case, inference and fitting with the 7B SFR-Embedding-2 R model (13.5 GB) on the NVIDIA A100 was feasible for input lengths comparable to a typical document (i.e., literature abstract text), using 27 GB when loaded into the GPU memory (out of 40 GB available) and 26.49 GB (fp32) memory (out of 64 GB RAM available) – at initialization time. The use of quantized models or reduced input length can lower hardware requirements, though often at the cost of reduced accuracy.

Execution time scales linearly with dataset size for embedding computation and has a variable impact on dimensionality reduction (UMAP) and clustering (HDBSCAN). For scalability reference, a small dataset such as M5 (33,769 abstracts) required 2h 48m for embeddings calculation and 23m for model fitting; a larger dataset such as M2 (339,949 abstracts) required 25h 18m for embeddings calculation and 1h 51m for model fitting.

Evaluation of topic modeling results

We proposed a customized implementation of the BERTopic pipeline, where a local optimal configuration can be found by exploiting our hyperparameters optimization and model registration mechanisms. This procedure is necessary due to the high quantity of data and the need to use many different models (e.g., five in our case) to be trained and fitted at the same time. A quantitative evaluation of a model can be obtained at each single iteration (with a new candidate hyperparameter configuration) by leveraging the DBCV score. Then, the final selected configuration is assessed through a manual evaluation, as described next.

For evaluating the TETYS pipeline, we compared two different configurations used in our specific use case:

- **Baseline**: Allenai-SPECTER Embedding Model (this is a non-LLM model developed by AllenAI [52]), with hyperparameters (exact) grid search method.
- **TETYS**: SFR-Embedding 2 R Embedding Model ([41]), with hyperparameters random search method.

Note that the Baseline configuration leverages Scientific Paper Embeddings using Citation-informed TransformERs (SPECTER), a pre-defined model developed to learn general-purpose vector representations of scientific documents. It builds on the architecture of Transformer-based language models, in particular, SciBERT [53], an adaptation of the BERT model architecture [54] to the scientific domain. The model is trained on abstracts concatenated with the corresponding paper title; it produces 768-dimensional embedding representations. This configuration is much smaller and faster to fine-tune, thus, we use a grid search strategy for hyperparameter tuning to iterate over all combinations of parameters. The embedding model is specialized for scientific documents, which perfectly corresponds to our task.

On the other hand, the TETYS configuration is the novel one proposed in this work, as described in the 'Materials and Methods' section. This configuration is larger and very time-consuming for the fitting process. Since we introduced model registration in the original pipeline –storing the best-performing model identified at any point– it became impractical to try all possible combinations for models, as fitting some models for certain macro-areas can take a long time (i.e., approximately exceeding an hour). For this reason, a random search strategy was used to avoid excessive computation time. Due to this, we may not find the best possible model (rather, one that achieves a local maximum of the DBCV score). Note that the embedding model is more general and optimized for a broader range of tasks (differently from SPECTER). Supplementary Table 1, in the Appendix, presents the values of the hyperparameters for the best models obtained using the two configurations in the five macro-areas scenarios.

Quantitative assessment

The Density-Based Clustering Validation (DBCV) quantitatively evaluates the quality of the topics' structure identified by the model. It provides an overall score that allows us to assess embeddings computation and hyperparameter search (for dimensionality reduction and clustering), providing one optimal configuration for a given dataset (macro-area).

In addition to DBCV, we considered other topic modeling-specific metrics that hint at the quality of the resulting models. These include topic cohesion and diversity scores. For coherence, we compute three standard measures: C_v, C_{UMass}, and C_{NPMI}, as proposed in [55], using the gensim [56] implementation. For topic diversity, we use the implementation provided in OCTIS [57]. These metrics assess how interpretable and distinct the topics are. In all cases, *higher values indicate better quality*. Note that, as reported by Stevens *et al.* [58], these metrics often correlate with other factors such as the number of topics in the model and the noise in the labels of the topics. Empirically, models producing too many or poorly defined topics tend to have lower coherence, while noisy textual representations tend to have higher diversity scores. While DBCV remains our reference metric, these topic modeling-specific metrics helped to validate the semantic and structural quality of the topics generated downstream.

Table 3 provides an overview of the number of topics and corresponding metrics' values for the five macro/areas. DBCV scores are also compared in the radar plot in Figure 6, showing an overall consistent improvement in the LLM-based configuration.

We note that for M1 (Basic Human Needs and Well-being) and M2 (Environmental Sustainability), the LLM-based configuration model produced a significantly greater number of topics compared to the non-LLM-based configuration model. Surprisingly, for M5 –probably the most heterogeneous dataset (as observed in the analysis of the five largest topics of Figure 5)– the number of topics found with the Baseline configuration is consistently greater than the one with the TETYS configuration. This is possibly due to the particular combination of `n_components` and the `n_neighbors` parameter values in TETYS: we are using a higher-dimensional space and forcing the model to look for a much larger neighborhood, resulting in fewer bigger clusters (w.r.t. the Baseline configuration).

Manual analysis

As an unsupervised technique, topic modeling attempts to identify topics within collections of documents without leveraging any other information, labels, or predefined

	Baseline						TETYS					
	#topics	DBCV	C_v	C_{UMass}	C_{NPMI}	Diversity	#topics	DBCV	C_v	C_{UMass}	C_{NPMI}	Diversity
M1	301	0.44	0.563	-1.487	0.068	0.801	550	**0.52**	0.611	-1.412	0.088	0.637
M2	424	0.72	0.496	-2.310	0.041	0.752	856	**0.76**	0.552	-2.056	0.052	0.563
M3	98	0.36	0.455	-0.731	-0.059	0.739	181	**0.39**	0.499	-0.772	-0.017	0.615
M4	42	0.44	0.428	-0.910	-0.040	0.852	136	**0.46**	0.476	-0.879	-0.041	0.646
M5	291	0.37	0.425	-0.929	-0.074	0.838	167	**0.38**	0.568	-0.928	-0.001	0.736

Table 3: For each macro-area/configuration, overview of number of topics identified by the models and maximum achieved DBCV score, C_v, C_{UMass}, C_{NPMI}, and Diversity.

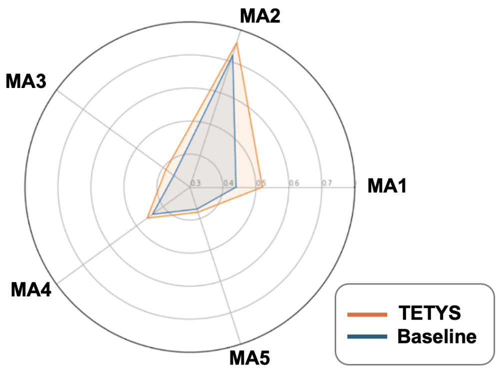

Fig. 6: DBCV scores for both model configurations for five macro-areas

topics. Then, evaluating the quality of topic models becomes a challenging task that requires domain knowledge and expertise in the fields covered by the scientific papers under consideration.

The goal of our evaluation is to determine whether the LLM-based topic model is better at assigning topics (as we postulated), given the enhanced potential of the employed embedding model. Our manual evaluation was carried out for two macro-areas, i.e., M1 (Basic Human Needs and Well-being) and M2 (Environmental Sustainability), which are the largest ones and encompass the greatest number of Sustainable Development Goals.

By employing the same two configurations of the previously described quantitative assessment, we performed the inference on a test set of 50 abstracts for each macro-area; these abstracts were new, i.e., not seen by the models in the training phase (thus, here, we speak about 'inference' rather than 'fitting'). In these datasets, we did not include any abstract assigned to the special topic "-1" (i.e., that does not belong to any valid topic) by any of the two configurations.

After classifying the abstracts with both configurations, we asked two researchers who are experts respectively in the domains of M1 and M2, to manually assess each abstract. They were equipped with a spreadsheet whose rows represent single articles; for each article, we provided the abstract, doi, and additional metadata (such as the author-defined keywords and the subject category). For both the *Baseline configuration* and the *TETYS configuration* we provided the topic ID, topic probability, topic name, number of abstracts assigned to the topic, and the list of the ten most represented terms in the topic (along with their frequency). Given this information, the evaluators were asked to indicate the identifiers of the most suitable topic among: 1) the ones available in the Baseline configuration; and 2) the ones available in the TETYS configuration. By comparing the evaluators' choice with the ones derived from the automatic configurations, we computed the Precision, Recall, and F1-scores (see Table 4). TETYS achieves better results in all the indicators—specifically, the F1-weighted score shifts from ∼70% to ∼90% in both M1 and M2 cases. We note that the Baseline performed better in M1, which is consistent with the fact that the targeted macro-area M1 'Basic Human Needs and Well-being' is semantically closer to the training set focus of SPECTER, as opposed to M2 'Environmental Sustainability'.

		Baseline			TETYS		
	Avg type	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
M1	Micro	0.800	0.800	0.800	0.920	0.920	0.920
	Macro	0.719	0.725	0.713	0.861	0.881	0.868
	Weighted	0.805	0.800	0.719	0.897	0.920	0.905
M2	Micro	0.700	0.700	0.700	0.920	0.920	0.920
	Macro	0.675	0.714	0.686	0.859	0.870	0.862
	Weighted	0.660	0.700	0.672	0.910	0.920	0.913

Table 4: Precision, recall, and F1 scores for both configurations run on M1 and M2. Note that, in case of multi-class imbalanced data classification tasks, like the one we are resolving, the Micro average is considered the most appropriate and meaningful one.

Moreover, we asked our evaluators to declare a preference between the assignment obtained using the Baseline configuration versus the one obtained using the TETYS configuration. Here, we allowed three possible choices:

- the evaluator concludes that the assignment obtained by the **Baseline** configuration is superior;
- the evaluator concludes that the assignment obtained by the **TETYS** configuration is superior;
- none of the assignments is clearly superior w.r.t. the other one (**undefined**).

Evaluator's choice	**M1 Percentage**	**M2 Percentage**
Baseline	18%	16%
TETYS	56%	60%
undefined	26%	24%
McNemar's test result (h_0: baseline > TETYS)	p-value 0.0025 statistic 9.0	p-value 0.0004 statistic 8.0

Table 5: Ballot comparison, with statistical evidence that TETYS configuration is strongly preferable to the Baseline in both M1 and M2.

Table 5 reports the number of each selected option in percentage. We statistically tested the preference of one configuration over the other; along the guidelines indicated in Schuff *et al.* [59], we performed the non-parametric McNemar statistical test [60] (used for paired nominal data), ignoring the 'undefined' cases. For both macro-areas, we observed a *strong statistical preference* for the TETYS configuration over the Baseline one, rejecting the null hypothesis in both cases, with -respectively- 0.0025 and 0.0004 p-values.

From this small experiment, we conclude that the LLM-based configuration is slightly better or at least as good as the non-LLM-based configuration. We expect that such restrained improvement is due to the use of the random search strategy for the LLM-based model, which means that we likely settled for a model that is not the best possible one.

By manual inspection of topics, we also observed that the TETYS configuration allowed us to achieve better quality, interpretability, and diversity [9]. TETYS also improved flexibility over the Baseline, because the LLM has more knowledge about different domains, while SPECTER was specific for the dataset (used for training) covering the medical/biological domain; note that SPECTER-AllenAI can be considered a very strong baseline for the requested task, as it is specifically designed for scientific literature.

Dashboard for interactive exploration

The results of the TETYS pipeline are made available through a Web application (http://gmql.eu/tetys/), demoed in [61], that allows users to appreciate the topics (resulting from the topic modeling sub-pipeline) and their characteristics, including their representation in time (resulting from the topic exploration sub-pipeline).

Figure 7 represents the system architecture divided into a *frontend* and a *backend*. The frontend contains a Web application working as a **Client** with functionalities that allow users to select a macro-area of interest, filter the content of the topic model using keywords or a specific publication's DOI, visualize the content, and download it (through plots and tables).

The backend contains four modules. Data persistence is taken care of in the **Database** (collecting publications metadata and information describing the topics, like their trends over time, stored as time-series data) and in the **ML Model registry**, which stores the topics models of the project as large *pickle* objects. The database is implemented with DuckDB [62], an in-process analytical database, that we use to exploit the efficiency in data storage and retrieval of the Apache Parquet format [63]. These two modules can be queried by the central **Server**, i.e., the orchestrator of TETYS: this includes a project registry along with services to perform keyword-based search and similarity-based search over the five different projects (one per macro-area). In each project, we allow analysis (i.e., statistical testing) and results download. Keyword-based search is exploited to find ranked topics that are close (i.e., relevant) to specific keywords. Similarity-based search is exploited to find ranked topics that are relevant to a specific point in the embedding space, i.e., one abstract – identified through its DOI. These search procedures make use of the **External service** of Crossref APIs [64], which retrieve all the papers' information that can be visualized in the

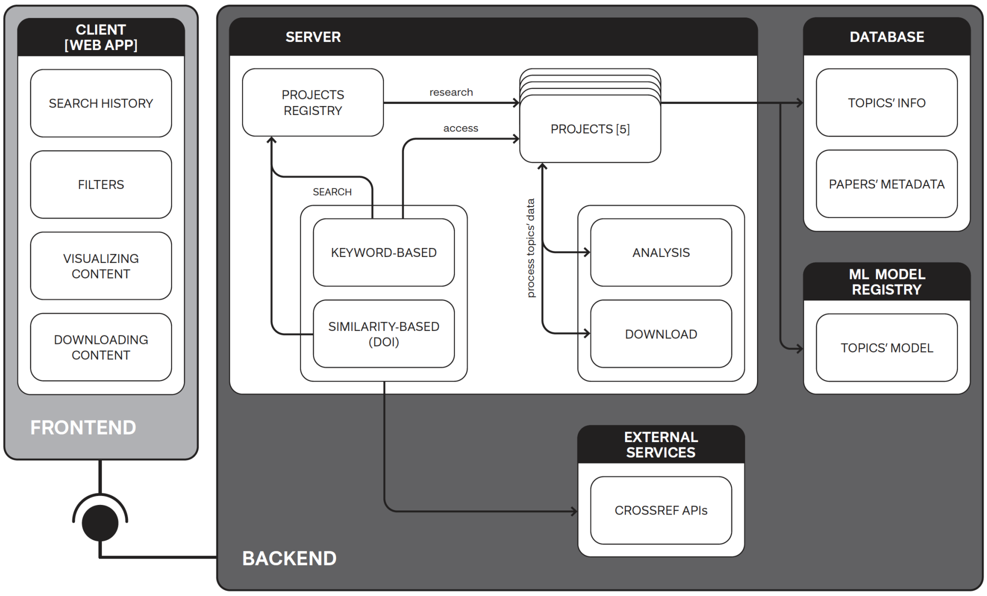

Fig. 7: System architecture

application. Note that the Model registry contains the models that, for each project, infer the most relevant topics for any query, both keyword-based and DOI-based.

The TETYS dashboard allows us to directly inspect the results obtained by our pipeline, supporting users in exploring topics, which would be a tedious and time-consuming task if performed manually. Users are asked to select one macro-area out of the five offered. For each macro-area, they can either select one of the trending (i.e., biggest) topics shown in a scrollable gallery or start their search using a keyword or a specific DOI. These two possibilities allow them to access two possible pages: the *Single Topic* page (see Figure 8, Panel A) or the *Topic Comparison* page (see Figure 8, Panels B/C). Panel A shows a descriptive card of the topic with its wordcloud and star diagram, a component for performing two-interval or multi-interval comparisons between user-selected time spans of the topic time series, and a downloadable list of publications that are assigned to the topic. Panel B shows a set of topics selected by the user from a pool of topics related to the searched keyword; topics (max. 5) can also be selected during multiple consecutive searches (as shown in Panel C). Their corresponding time series are shown on the same graph, where users can (de)select tracks as needed and use a slider to focus on a time span of interest. Different time resolutions can be set; the relative frequencies of the topics in one specific time instant can be visualized on hover.

Fig. 8: Pages of the TETYS Dashboard of M1 (Basic Human Needs and Well-being). A) Single Topic page; B) Topic Comparison page with a single keyword search session; C) Topic Comparison Page with multiple keyword search sessions.

Methods Literature Review

Topic modeling and clustering

A growing body of work challenges the conventional boundary between topic modeling and clustering, highlighting the feasibility of clustering-based approaches as viable, often superior alternatives. Thompson and Mimno [65] demonstrated that clustering token-level contextualized embeddings from pre-trained language models (PLMs) like BERT and GPT-2 can yield topic-like structures with performance equivalent to or better than traditional Latent Dirichlet allocation, especially in capturing polysemy and scaling across topic numbers. Sia *et al.* [4] further support this direction; they show how clustering word embeddings using hard and soft clustering algorithms that are more sophisticated than k-means, combined with appropriate feature reduction techniques, can produce coherent, computationally efficient topics. Early work on whole-document embeddings, such as SPECTER, also exhibits proof that PLMs encode topical information and are aware of cross-topic relatedness; e.g., Engineering, Mathematics, and Computer Science are close to each other, whereas Business and Economics are close to each other.

In addition to the curse of dimensionality caused by the high-dimensional embeddings produced by LLMs, clustering algorithms selected for topic modeling should also be robust against non-convex-shaped clusters, since, as indirectly demonstrated by Petukhova *et al.* [66], document clusters may not exhibit convex shapes. In these cases, density plays a critical role, reinforcing the importance of choosing clustering methods that adapt to these structural realities.

For these reasons, density-based techniques are particularly well-suited for document grouping and topic modeling tasks. DBSCAN, its hierarchical version HDBSCAN, and other optimized variants, such as QuickDBScan and KDTreeDBSCAN [67], are gaining traction due to their ability to detect clusters of arbitrary shape and manage outliers effectively in this context.

These developments underscore the growing relevance of density-based clustering in topic modeling, suggesting that future efforts can benefit substantially from integrating these modern, non-parametric strategies.

Clustering high-dimensional data

Dimensionality reduction techniques for clustering high-dimensional data, such as document embeddings, were extensively explored to improve both the quality and the efficiency of clustering. While traditional techniques like Principal Components Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) cover most of the use cases, more recent research highlights the advantages of UMAP in preserving both local and global structures when reducing dimensionality. UMAP's adaptability across embedding spaces from various language models makes it especially useful for document clustering tasks. Notably, Allaoui *et al.* [68] demonstrated that applying UMAP as a preprocessing step significantly boosts the performance of standard clustering algorithms like k-Means and HDBSCAN, leading to more coherent clusters and faster computation.

Extraction of labels for topics

There has been extensive research on weighting schemas that significantly impact model performance in representing topics. Foundational approaches include basic occurrence counting, term frequency (TF), and Term Frequency-Inverse Document Frequency (TF-IDF) [69], along with their variations. Okapi BM25 extends TF-IDF by normalizing weights based on document length relative to the corpus average, addressing biases in longer documents [70]. Several studies have systematically compared these weighting schemas to assess their influence on the topic model's performance [71].

For short texts specifically, to address the challenge of sparsity in building word co-occurrence statistics, Zuo *et al.* [72] developed pseudo-document approaches, noting that semantically irrelevant terms can disproportionately influence topic identification – a problem also marginally addressed by weighting schemas like Okapi BM25. Other solutions mitigated this tendency by applying feature engineering techniques to extract high-quality vocabularies from the initial corpus before the actual topic modeling process and by providing *contextual cues* to the topic model [73].

Another significant direction in term-weighting research focuses on entropy-based approaches that identify informative words for topic modeling. Li *et al.* [74] introduced an entropy weighting (EW) scheme that leverages conditional entropy, measured through word co-occurrences, to automatically assign higher weights to informative words and lower weights to less meaningful terms. Bridging traditional and neural approaches, Dieng *et al.* [75] introduced the Embedding Topic Model (ETM), which represents both words and topics as vectors in a shared embedding space. ETM places words into topics based on vector proximity, leveraging semantic relationships beyond co-occurrence patterns while maintaining a generative process similar to LDA but with a logistic-normal distribution. This neural network-optimized approach draws words from their semantic context, resulting in more coherent topic representations than possible with traditional term weighting schemas.

Beyond document-level weighting, researchers have also explored class-level or cluster-level approaches beyond traditional document-level weighting, as exemplified by cTF-IDF implemented in BERTopic. While our research focuses on longer, single-language documents, where text preprocessing techniques like stop-words removal are sufficient for extracting coherent topics' labels, it is worth noting emerging directions that involve large language models and transcend traditional weighting approaches, such as PromptTopic [76] and Mu *et al.* [77], which leverage advanced prompting techniques, topic seeding, and summarisation. However, these advanced methods come with significant computational costs that must be weighed against their potential benefits.

Discussion

Main contributions. The proposed system presents a series of innovations that include the possibility of applying the BERTopic pipeline in a customized way on big data corpora, the optimization of the hyperparameter search, and the storage of intermediate models to obviate the stochastic nature of HDBSCAN. From the technological point of view, our system poses the basis for applying the pipeline to many diverse domains and text corpora, provided that the constraints of our setup are observed.

Design notes. Regarding the use of LLMs for embedding computation, we observed that BERTopic models developed with LLM-based embedding models typically identified more topics than models developed with non-LLM-based embedding models. One of the likely reasons is the dimensionality of the embedding vectors, which is much larger in the case of LLM-based embedding models (4096 >> 768). In a larger latent space, the model has a better capacity to distinguish between similar but different topics, which can be difficult for models in a small latent space. In addition to the larger dimensionality of the latent space and better semantic representation, LLM-based embedding models are, in their essence, more powerful, since they are pre-trained on much larger and extensive text data, on top of using more advanced learning techniques and fine-tuning.

Regarding the choice of clustering quality-driven optimization, we chose the Density-Based Clustering Validation (DBCV) score as our primary optimization criterion as it is designed to evaluate clusters of arbitrary shapes and varying densities, accounting for cluster compactness and density separation (high density vs low density areas) – key properties of the clustering structures we investigated in our unsupervised setting and high-dimensional latent space representations. Other typical topic metrics are not specifically suited for high-dimensional density-based clustering. We did not select the Silhouette Score or Calinski-Harabasz Index, as both rely on assumptions of convex cluster shapes and separation that are not suitable for density-based methods in high-dimensional spaces. We excluded the Adjusted Rand Index (ARI), which assumes hard partitions, making it incompatible with our soft clustering setup. Finally, the Davies-Bouldin Index, while occasionally adapted for density-based methods, was discarded as limited in its ability to handle clusters of arbitrary shapes, as also noted in the original DBCV paper [44].

Limitations. A limitation in the current approach is related to the representation of topics. Since we run topic modeling as an unsupervised task on a high-dimensional latent space, given topics may appear not to be precisely separated from a textual perspective – as they can share terms in their representations. Through manual investigation, we verified that this is not due to limitations in the topics' identification process; instead, the problem rather pertains to representation extraction. We are confident that this issue will be solved with the application of new language models that are fine-tuned for this purpose.

Moreover, in our evaluation, we did not discuss the stability and efficiency [9] of our topic model, as they are not integral to our process. Note that, after the first fitting of the topic model, we reuse the model and update it with new entries during inference, without being affected by concerns of stability or efficiency.

Finally, TETYS currently supports only English as the input language, and input texts are limited in length, approximately equal to typical abstract size, due to GPU memory constraints of our experimental setup (single A100 40GB) and embedding model limitations, because texts are embedded as a whole.

Impact and future challenges. Regarding the specific working instance exposed in the TETYS dashboard, focusing on SDGs-related literature, we believe the system can be useful to a very broad range of stakeholders, including users such as students, researchers, or professionals who are interested in deepening their knowledge on an area of research and need a fast way to grasp a general idea of the main topics and their evolution in the last twenty years. Possibly, one such dashboard could be extended into a product useful to funding bodies, universities, or research centers.

Future work includes evaluating whether a single embedding for a full-length document is meaningful or if it is better to split texts into chunks and embed each separately, as commonly done in retrieval-augmented generation (RAG) tasks.

Conclusion

The TETYS pipeline is based on BERTopic; we enhanced it by using LLMs for the embedding computation. Then, for each data corpus at hand, we can find a local maximum in the random search space of the hyperparameter configuration that regards dimensionality reduction and clustering. This configuration is used for model registration and fitting. Given a corpus of text documents in input, eventually, our pipeline builds a valuable trade-off between the best and "fastest-to-find" topic model possible. We measure the goodness of configurations one by one by leveraging DBCV, while we assess the overall arrangement with a thorough manual evaluation.

This arrangement is particularly fit for big data corpora; we additionally enrich the pipeline result by enabling keyword-search and dynamic topic modeling with time series exploration using configurable time-bins and relative frequencies (with ranking). The final result exposes a rich computational model and associated metadata to the users, making topics' exploration interactive and possible on a large scale.

In this work, we demonstrated the power of the TETYS pipeline by running it on five different text document corpora generated from the Scopus database by collecting research abstracts that are pertinent to a specific set of Sustainable Development Goals, as defined by the United Nations. This approach allowed us to automatically uncover the attitude and the topic trends found in research literature about themes of interest for the general community.

Moving beyond research-related text, we expect to reapply the pipeline and its paradigm to other application domains that may particularly benefit from this kind of data analytics, i.e., analyzing topics' evolution of legislative text from different countries and systems.

Appendix

Macro area	Config.	Pipeline step	Parameter	Value
M1	Baseline 301 topics	UMAP	n_neighbors	20
			n_components	5
			min_dist	0.0
		HDBSCAN	min_cluster_size	25
			min_samples	100
	TETYS 550 topics	UMAP	n_neighbors	20
			n_components	5
			min_dist	0.0
		HDBSCAN	min_cluster_size	25
			min_samples	75
M2	Baseline 424 topics	UMAP	n_neighbors	50
			n_components	10
			min_dist	0.0
		HDBSCAN	min_cluster_size	25
			min_samples	50
	TETYS 856 topics	UMAP	n_neighbors	20
			n_components	10
			min_dist	0.0
		HDBSCAN	min_cluster_size	25
			min_samples	75
M3	Baseline 98 topics	UMAP	n_neighbors	20
			n_components	10
			min_dist	0.0
		HDBSCAN	min_cluster_size	25
			min_samples	50
	TETYS 181 topics	UMAP	n_neighbors	100
			n_components	10
			min_dist	0.0
		HDBSCAN	min_cluster_size	25
			min_samples	10
M4	Baseline 42 topics	UMAP	n_neighbors	50
			n_components	28
			min_dist	0.0
		HDBSCAN	min_cluster_size	25
			min_samples	50
	TETYS 136 topics	UMAP	n_neighbors	50
			n_components	28
			min_dist	0.0
		HDBSCAN	min_cluster_size	25
			min_samples	10
M5	Baseline 291 topics	UMAP	n_neighbors	5
			n_components	5
			min_dist	0.0
		HDBSCAN	min_cluster_size	25
			min_samples	10
	TETYS 167 topics	UMAP	n_neighbors	100
			n_components	35
			min_dist	0.0
		HDBSCAN	min_cluster_size	25
			min_samples	15

Table 1: Hyperparameter values for macro areas M1 to M5, in both configurations Baseline (non-LLM with grid search) and TETYS (LLM with random search), considering the UMAP and HDBSCAN steps. From both steps, we omit `metric = 'euclidean'` as it is always the same value for both UMAP and HDBSCAN, as well as `cluster_selection_method = 'eom'` as it is always the same value for HDBSCAN.

Declarations

List of abbreviations.
DBCV: Density-Based Clustering Validation
DOI: Digital Object Identifier
HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise
LLM: Large Language Model
MMR: Maximal Marginal Relevance
MTEB: Massive Text Embedding Benchmark
SBERT: Sentence Bidirectional Encoder Representations from Transformers
SDG: Sustainable Development Goal
SPECTER: Scientific Paper Embeddings using Citation-informed TransformERs
TETYS: Topics Evolution That You See
UMAP: Uniform Manifold Approximation and Projection

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Availability of data and materials. The original dataset (on which topic models have been built) was downloaded from Scopus API between June 1 and 30, 2024 via http://api.elsevier.com and http://www.scopus.com. The pipeline is packaged as a Python module and organized as a series of scripts. The code is available and documented on the GitHub repository https://github.com/FrInve/TETYS, under the BSD-3-Clause license; it contains requirements to reproduce the conda environment and code/data of intermediate steps. The TETYS web application is available at http://gmql.eu/tetys.

Competing interests. The authors declare that there are no conflicts of interest.

Funding. This research is supported by the TEThYS project, a beneficiary of the NGI Search 2nd Open Call (Number of the Sub-grant Agreement SEARCH OC2_18), funded by the European Union. The NGI Search project has received funding from the European Union's Horizon Europe research and innovation programme under the grant agreement 101069364 and it is framed under the Next Generation Internet Initiative

NGI SEARCH

Funded by the European Union

Authors' contributions. F.I. and A.B. jointly conceptualized the work and its methodology and revised/edited the manuscript; F.I. designed and implemented the data pipeline, the backend, and the frontend, while supervising the work of F.C., J.J., and A.S.; F.C. contributed to the User interface and UX design; J.J. contributed to the datasets' preparation and the topic modeling/exploration pipeline and ran the inference evaluation; A.S. contributed to the frontend of the dashboard; A.B. supervised the team, acquired funding, coordinated the project, and wrote the original draft.

Acknowledgements. The authors are grateful to Stefano Ceri, Lorenzo Mari, and Mark J. Carman for the useful advice during the course of the research.

References

[1] United Nations, Department of Economic and Social Affairs.: Sustainable Development Goals. Last accessed: Nov 24th, 2024. https://sdgs.un.org/goals.

[2] Elsevier.: Elsevier Developer Portal – Academic Research. Last accessed: Nov 24th, 2024. https://dev.elsevier.com/academic_research_scopus.html.

[3] Krause A, Leskovec J, Guestrin C. Data association for topic intensity tracking. In: Proceedings of the 23rd international conference on Machine learning; 2006. p. 497–504.

[4] Sia S, Dalmia A, Mielke SJ. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020. p. 1728–1736.

[5] Moody CE. Mixing Dirichlet Topic Models and Word Embeddings to Make Lda2vec. arXiv:160502019. 2016;https://doi.org/10.48550/arXiv.1605.02019.

[6] Meng Y, Zhang Y, Huang J, Zhang Y, Han J. Topic discovery via latent space clustering of pretrained language model representations. In: Proceedings of the ACM web conference 2022; 2022. p. 3143–3152.

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

[8] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:220305794. 2022;https://doi.org/10.48550/arXiv.2203.05794.

[9] Abdelrazek A, Eid Y, Gawish E, Medhat W, Hassan A. Topic modeling algorithms and applications: A survey. Information Systems. 2023;112:102131.

[10] Falkenberg M, Galeazzi A, Torricelli M, Di Marco N, Larosa F, Sas M, et al. Growing polarization around climate change on social media. Nature Climate Change. 2022;12(12):1114–1121.

[11] Ebeling R, Sáenz CAC, Nobre JC, Becker K. Analysis of the influence of political polarization in the vaccination stance: the Brazilian COVID-19 scenario. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 16; 2022. p. 159–170.

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

[13] Petrescu A, Truică CO, Apostol ES. Sentiment analysis of events in social media. In: 2019 IEEE 15th International Conference on Intelligent Computer

Communication and Processing (ICCP). IEEE; 2019. p. 143–149.

[14] Mitroi M, Truică CO, Apostol ES, Florea AM. Sentiment analysis using topic-document embeddings. In: 2020 IEEE 16th international conference on intelligent computer communication and processing (ICCP). IEEE; 2020. p. 75–82.

[15] Yin H, Aryani A, Lambert G, White M, Salvador-Carulla L, Sadiq S, et al. Leveraging artificial intelligence technology for mapping research to sustainable development goals: A case study. arXiv preprint arXiv:231116162. 2023;.

[16] Kharlashkin L, Macias M, Huovinen L, Hämäläinen M. Predicting Sustainable Development Goals Using Course Descriptions–from LLMs to Conventional Foundation Models. Journal of Data Mining & Digital Humanities. 2024;.

[17] Corallo L, Li G, Reagan K, Saxena A, Varde AS, Wilde B. A Framework for German-English Machine Translation with GRU RNN. In: EDBT/ICDT Workshops; 2022. .

[18] Puri M, Varde AS, de Melo G. Commonsense based text mining on urban policy. Language Resources and Evaluation. 2023;57(2):733–763.

[19] Verma A, Nayak JK. Understanding public sentiments and misbeliefs about Sustainable Development Goals: a sentiment and topic modeling analysis. Journal of Information, Communication and Ethics in Society. 2024;22.

[20] Roldan-Alvarez D, Martínez-Martínez F, Martin E, Haya PA. Understanding discussions of citizen Science around sustainable development goals in Twitter. IEEE Access. 2021;9:144106–144120.

[21] Fitri H, Widyawan W, Soesanti I. Topic modeling in the news document on sustainable development goals. IJITEE (International Journal of Information Technology and Electrical Engineering). 2021;5(3):82–89.

[22] Saheb T, Dehghani M, Saheb T. Artificial intelligence for sustainable energy: A contextual topic modeling and content analysis. Sustainable Computing: Informatics and Systems. 2022;35:100699.

[23] Raman R, Pattnaik D, Lathabai HH, Kumar C, Govindan K, Nedungadi P. Green and sustainable AI research: an integrated thematic and topic modeling analysis. Journal of Big Data. 2024;11(1):55.

[24] Smith TB, Vacca R, Mantegazza L, Capua I. Discovering new pathways toward integration between health and sustainable development goals with natural language processing and network science. Globalization and Health. 2023;19(1):44.

[25] Elsevier.: Elsevier Portal. Last accessed: Nov 24th, 2024. https://www.elsevier.com/.

[26] Mongeon P, Paul-Hus A. The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics. 2016;106:213–228.

[27] Grisales A AM, Robledo S, Zuluaga M. Topic modeling: Perspectives from a literature review. IEEE Access. 2023;11:4066–4078.

[28] Choi HS, Lee WS, Sohn SY. Analyzing research trends in personal information privacy using topic modeling. Computers & Security. 2017;67:244–253.

[29] Rejeb A, Rejeb K, Appolloni A, Kayikci Y, Iranmanesh M. The landscape of public procurement research: a bibliometric analysis and topic modelling based on Scopus. Journal of Public Procurement. 2023;23(2):145–178.

[30] Invernici F, Bernasconi A, Ceri S. Exploring the evolution of research topics during the COVID-19 pandemic. Expert Systems with Applications. 2024;252:124028.

[31] Grootendorst M.: BERTopic – Best practices on large datasets. Last accessed: Nov 24th, 2024. https://github.com/MaartenGr/BERTopic/issues/491.

[32] OpenAI.: ChatGPT. Last accessed: Nov 24th, 2024. https://chatgpt.com/.

[33] Dai S, Xu C, Xu S, Pang L, Dong Z, Xu J. Bias and unfairness in information retrieval systems: New challenges in the llm era. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2024. p. 6437–6447.

[34] Elsevier.: Elsevier Developer Portal. Last accessed: Nov 24th, 2024. https://dev.elsevier.com/.

[35] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Inui K, Jiang J, Ng V, Wan X, editors. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019. p. 3982–3992.

[36] McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:180203426. 2018;https://doi.org/10.48550/arXiv.1802.03426.

[37] McInnes L, Healy J, Astels S. hdbscan: Hierarchical Density Based Clustering. Journal of Open Source Software. 2017;2(11):205.

[38] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12(85):2825–2830.

[39] Ceri S, Bozzon A, Brambilla M, Della Valle E, Fraternali P, Quarteroni S. Information Retrieval Models. In: Web Information Retrieval. Berlin, Heidelberg: Springer; 2013. p. 27–37.

[40] Muennighoff N, Tazi N, Magne L, Reimers N. MTEB: Massive text embedding benchmark. arXiv preprint arXiv:221007316. 2022;https://doi.org/10.48550/arXiv.2210.07316.

[41] Meng R, Liu Y, Joty SR, Xiong C, Zhou Y, Yavuz S.: SFR-Embedding-2: Advanced Text Embedding with Multi-stage Training. Last accessed: Nov 24th, 2024. https://huggingface.co/Salesforce/SFR-Embedding-2_R.

[42] Salesforce AI Research.: SFR-Embedding-Mistral. Last accessed: Nov 24th, 2024. https://huggingface.co/Salesforce/SFR-Embedding-Mistral.

[43] Hugging Face.: Pipelines. Last accessed: Nov 24th, 2024. https://huggingface.co/docs/transformers/en/main_classes/pipelines.

[44] Moulavi D, Jaskowiak PA, Campello RJ, Zimek A, Sander J. Density-Based Clustering Validation. In: Proceedings of the 2014 SIAM international conference on data mining. SIAM; 2014. p. 839–847.

[45] The PyTorch Foundation.: PyTorch. Last accessed: Nov 24th, 2024. https://pytorch.org/.

[46] Mueller AC.: Wordcloud. Last accessed: Nov 24th, 2024. https://github.com/amueller/word_cloud.

[47] Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. Journal of the American statistical Association. 1952;47(260):583–621.

[48] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature methods. 2020;17(3):261–272.

[49] Dunn OJ. Multiple comparisons using rank sums. Technometrics. 1964;6(3):241–252.

[50] Lubitz W, Tumas W. Hydrogen: an overview. Chemical reviews. 2007;107(10):3900–3903.

[51] NVIDIA Corporation.: NVIDIA A100 Tensor Core GPU. Last accessed: Nov 24th, 2024. https://www.nvidia.com/en-eu/data-center/a100/.

[52] Cohan A, Feldman S, Beltagy I, Downey D, Weld D. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 2270–2282.

[53] Beltagy I, Lo K, Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. In: Inui K, Jiang J, Ng V, Wan X, editors. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019. p. 3615–3620.

[54] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–4186.

[55] 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; 2015. p. 399–408.

[56] Rehurek R, Sojka P. Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic. 2011;3(2).

[57] Terragni S, Fersini E, Galuzzi BG, Tropeano P, Candelieri A. OCTIS: Comparing and Optimizing Topic Models is Simple! In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics; 2021. p. 263–270. Available from: https://www.aclweb.org/anthology/2021.eacl-demos.31.

[58] Stevens K, Kegelmeyer P, Andrzejewski D, Buttler D. Exploring topic coherence over many models and many topics. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning; 2012. p. 952–961.

[59] Schuff H, Vanderlyn L, Adel H, Vu NT. How to do human evaluation: A brief introduction to user studies in NLP. Natural Language Engineering. 2023;29(5):1199–1222.

[60] McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153–157.

[61] Invernici F, Bernasconi A, Curati F, Jakimov J, Samavi A, et al. TETYS: Configurable Topic Modeling Exploration for Big Corpora of Text Documents. In: Proceedings of the 28th International Conference on Extending Database Technology (EDBT). vol. 28; 2025. p. 1114–1117.

[62] Raasveldt M, Mühleisen H. Duckdb: an embeddable analytical database. In: Proceedings of the 2019 International Conference on Management of Data; 2019. p. 1981–1984.

[63] Apache Software Foundation.: Apache Parquet. Last accessed: Nov 24th, 2024. https://parquet.apache.org/.

[64] Crossref.: Documentation - Metadata Retrieval - REST API. Last accessed: Nov 24th, 2024. https://www.crossref.org/documentation/retrieve-metadata/rest-api/.

[65] Thompson L, Mimno D. Topic modeling with contextualized word representation clusters. arXiv preprint arXiv:201012626. 2020;.

[66] Petukhova A, Matos-Carvalho JP, Fachada N. Text clustering with large language model embeddings. International Journal of Cognitive Computing in Engineering. 2025;6:100–108.

[67] Radulescu IM, Truica CO, Apostol ES, Boicea A, Radulescu F, Mocanu M. Performance Evaluation of DBSCAN With Similarity Join Algorithms. In: The 34th International-Business-Information-Management-Association (IBIMA) Conference, Madrid, Spain; 2019. .

[68] Allaoui M, Kherfi ML, Cheriet A. Considerably improving clustering algorithms using UMAP dimensionality reduction technique: a comparative study. In: International conference on image and signal processing. Springer; 2020. p. 317–325.

[69] Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation. 1972;28(1):11–21.

[70] Robertson S, Zaragoza H, et al. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval. 2009;3(4):333–389.

[71] Truica CO, Radulescu F, Boicea A. Comparing different term weighting schemas for topic modeling. In: 2016 18th international symposium on symbolic and numeric algorithms for scientific computing (SYNASC). IEEE; 2016. p. 307–310.

[72] Zuo Y, Wu J, Zhang H, Lin H, Wang F, Xu K, et al. Topic modeling of short texts: A pseudo-document view. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. p. 2105–2114.

[73] Truica CO, Apostol ES, Leordeanu CA. Topic modeling using contextual cues. In: 2017 19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC). IEEE; 2017. p. 203–210.

[74] Li X, Zhang A, Li C, Ouyang J, Cai Y. Exploring coherent topics by topic modeling with term weighting. Information Processing & Management. 2018;54(6):1345–1358.

[75] Dieng AB, Ruiz FJ, Blei DM. Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics. 2020;8:439–453.

[76] Wang H, Prakash N, Hoang NK, Hee MS, Naseem U, Lee RKW. Prompting large language models for topic modeling. In: 2023 IEEE International Conference on Big Data (BigData). IEEE; 2023. p. 1236–1241.

[77] Mu Y, Dong C, Bontcheva K, Song X. Large language models offer an alternative to the traditional approach of topic modelling. arXiv preprint arXiv:240316248. 2024;.