TETYS: Configurable Topic Modeling Exploration for Big Corpora of Text Documents

Francesco Invernici
Politecnico di Milano
Milan, Italy
francesco.invernici@polimi.it

Anna Bernasconi
Politecnico di Milano
Milan, Italy
anna.bernasconi@polimi.it

Francesca Curati
Politecnico di Milano
Milan, Italy

Jelena Jakimov
Politecnico di Milano
Milan, Italy

Amirhossein Samavi
Politecnico di Milano
Milan, Italy

ABSTRACT

Fast exploration of vast text corpora is typically heavily time-consuming. Topic modeling allows for discovering key concepts in massive text datasets without requiring prior knowledge of their content. We built TETYS, an end-to-end topic modeling pipeline, easily configurable for processing and visualizing datasets. We demonstrate its use when applied to five datasets encompassing research on Sustainability Development Goals, defining the world's most pressing social, economic, and environmental challenges. TETYS is based on neural topic modeling and exploits LLMs to be proficient in many domains, including research publications that range from human sciences to engineering and technology. In this demo, participants will be able to interact with the dashboard to discover insights about the datasets and appreciate/test temporal trends in their research topics.

Tool: http://gmql.eu/tetys. **Video**: https://tinyurl.com/tetys-video. **Code**: https://github.com/FrInve/TETYS.

1 INTRODUCTION

Gaining insights into the contents of large corpora of documents can be overwhelming and ineffective without a clear starting point. Identifying the anatomy of the dataset is crucial in many knowledge retrieval and management applications. Topic modeling, interpreted as a clustering task [18] over the latent space of embeddings, offers a key solution to this challenge. Additionally, using the temporal dimension makes it possible to shape the dynamics of topics' evolution precisely, providing a deeper understanding of trends and shifts in a dataset's content, which are essential for effective decision-making and strategic planning.

In this paper, we demonstrate the use of TETYS (Topics' Evolution That You See) –designed in the context of the homonymous project [2]– the first open-source automated LLM-based topic modeling pipeline that produces interactive dashboards allowing users to grasp and analyze big corpora's content quickly. We showcase the platform by considering -as input- five datasets of scientific publications concerning Sustainability Development Goals.

Topic modeling has thus far been exploited for *ad hoc* analyses, while a general architecture that exploits it has not been previously proposed. There exist domain-specific dashboards (e.g., for tweets [9] or technological trends [16]), and programmatic composable dashboards –for expert users not oriented to user-friendly experience (e.g., [6, 14]), but none of these resemble TETYS, which automatically generates intuitive dashboards for exploring the content of any textual dataset; a working prototype on COVID-19-related literature has been proposed previously [13].

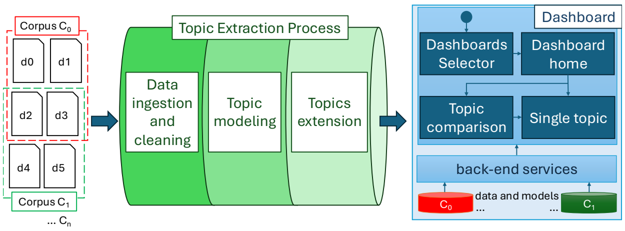


Figure 1: Architecture overview of the TETYS pipeline.

2 ARCHITECTURE AND IMPLEMENTATION

TETYS is implemented as an end-to-end pipeline that, from the input corpora, automatically extracts and analyzes topics, which are made available for examination in a configurable web dashboard. Two are its main architectural components (Fig. 1): the *extraction process*, which is run only once for each corpus of text documents, and the *interactive dashboard*, which collects and displays the results for all the corpora that went through the extraction process.

Topic extraction process. We include three phases (each implemented with a separate Python script): data ingestion and cleaning; topic modeling; and topic extension (enriching the model and enabling interactive exploration).

In the *data ingestion and cleaning* phase, the raw datasets of scientific documents are ingested and transformed into highly efficient Parquet files [7] with a predetermined output format that, for each tuple, includes an identifier (e.g., a paper's DOI), title, abstract, publication date, authors, and other optional domain-specific metadata that can be useful for visualization purposes (concatenated in a single field). During this phase, cleaning operations are applied to the raw datasets when necessary. By default, we perform record deduplication, removal of incomplete entries, selection of a preferred time window, and detection of the language of the abstract, with related filtering.

In the *topic modeling* phase, an unsupervised topic model is fitted on the ingested corpus of documents: a topic is assigned to each document, while a hierarchy of topics is derived from the structure of the latent semantic space of the whole corpus. This is achieved by adopting the transformer-based BERTopic [10] technique, with three novelties (discussed in our work [12]): we support the use of Large Language Models (LLMs) for embedding

© 2025 Copyright held by the owner/author(s). Published in Proceedings of the 28th International Conference on Extending Database Technology (EDBT), 25th March-28th March, 2025, ISBN 978-3-89318-099-8 on OpenProceedings.org. Distribution of this paper is permitted under the terms of the Creative Commons license CC-by-nc-nd 4.0.

Dashboard title	Describing keywords	#docs	#topics
Basic Human Need and Well-being	Poverty alleviation; Food security; Public health; Education access; Water quality; Sanitation infrastructure; Healthcare provision.	320,798	550
Economic Development and Employment	Economic growth; Innovation ecosystems; Infrastructure development; Entrepreneurship support; Industrialization strategies; Industrial Innovation; Labor market dynamics.	41,218	181
Environmental Sustainability	Renewable energy; Urban sustainability; Sustainable consumption; Climate change mitigation; Marine biodiversity; Ecosystem conservation; Energy efficiency.	339,949	856
Equality and Social Inclusion	Gender empowerment; Social equity; Inclusive policies; Women's rights; Minority rights; Income inequality; Social justice.	25,017	136
Global Partnership and Peace	Legal institutions; International cooperation; Peace efforts; Sustainable development cooperation; Global governance; Justice systems; Multilateral agreements.	33,769	167

Table 1: Overview of the five datasets processed by TETYS and their corresponding dashboards shown in this demonstration.

the documents; we implement an optimizer for automatically finding the best-performing hyper-parameters; and we add a model registration mechanism to increase the efficiency of the fitting process. Expert users have the ability to indicate a specific LLM instance suitable to their domain of use (in Hugging Face format [11]). In our default configuration, we adopt the `SFR-Embedding-2_R` [15] LLM, trained for semantic similarity.

In the *topic extension* phase, topics are equipped with time-series data. These include absolute and relative intensities, which are extracted from the temporal metadata of the ingested data, as well as other key performance indicators (KPIs) –computed on top of the intensities– such as rankings of differences over time and growth rates, which provide deeper insights into patterns and trends over time. All these data are stored in Parquet format, efficiently accessible by the other TETYS components.

Configurable dashboard. The TETYS dashboard makes the model and data generated from the ingested corpus available for display and exploration. It is designed as a Web application that relies on back-end services that expose –via RESTful APIs– the extended topic models. The user interface is developed using Angular.js and JavaScript data visualization libraries such as chart.js [5] and amCharts [1]. The backend services are developed in Python using FastAPI [19] and employ DuckDB [17] as an analytical database to query and retrieve all the topics' data computed during the Topic Extraction Process, which are stored in Parquet files.

The TETYS dashboard is easily adaptable to display the topics' information and trends of any corpus of documents processed by the pipeline. For each corpus, users need to complete a configuration file, termed *Project Card*, specifying the paths of the produced topic model and, optionally, the documents' metadata with their inferred topic. The backend exposes a *Project* object for each card found in its resource path.

The user interface is designed to be easily repurposed for any corpus of documents by copying a template dashboard, as those available in TETYS's repository, and changing the `project` parameter to the name set in the corresponding project card. An automated mechanism is currently under development to handle -at the time of launch- the addition of new dashboards for all newly available projects in the backend services.

User interface. The dashboard presents four types of pages for user interaction (see Fig. 1 for their navigation schema):

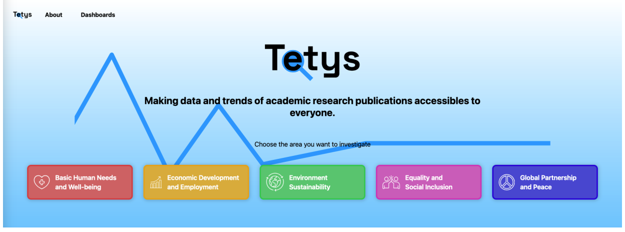
- **Dashboard selector.** Since the extraction process generates one dashboard for each project (i.e., an input corpus), we expose a landing page to select the one of interest.
- **Dashboard home.** For the chosen project, we show a carousel of interesting topics (e.g., the most represented ones) that can be selected. Users can alternatively start their exploration by using keywords or specific identifiers (e.g., DOI for scientific papers), prompting a search of the 10 corresponding most relevant topics.
- **Topic comparison.** Users can start up to five *keyword-search boxes*, each retrieving the 10 -related- most relevant topics. Across all the search components, a set of ≤5 topics can be selected for comparative display. The corresponding time series are shown in an interactive line plot, with options for selecting timespan and resolution. On hover, data points show ranked relative intensities.
- **Single topic.** When a topic is selected (from the carousel in a Dashboard home or the keyword-search boxes), a dedicated page is shown with several components: 1) a topic card reporting the topic title, number of assigned documents (and other available KPIs), and descriptive word cloud; 2) an interactive diagram of the terms' relevance in the topic; 3) a statistical toolbox for running two-interval or multi-interval comparison tests by selecting an overall timespan and specific enclosed intervals – we run a Kruskal-Wallis test, with Dunn's post-hoc test and Bonferroni p-value correction for multi-comparisons tests; 4) a downloadable list of documents included in the topic with their metadata.

3 DEMONSTRATION

For this demonstration, we applied TETYS to research papers published from 2006 to 2023 on Sustainability Development Goals (SDGs [20]), as we described in [12]. As our target was to build dashboards for inspection of big data corpora, we grouped SDGs by their macro-area of interest. We used the keywords specified by the authors in the papers' metadata to identify which SDG they addressed. We obtained five datasets covering all of the 17 Goals; Table 1 reports their title and describing keywords for the datasets, as well as their size in terms of number of documents and topics. By running TETYS on each dataset, we obtained five dashboards, each focused on a specific macro-area. In the following paragraphs of this demonstration, we refer to this specific instance of TETYS.

Scenarios. We cover all the main functionalities of TETYS Dashboards with three scenarios, which exemplify how to use TETYS to: (1) select a project regarding a corpus of documents and discover a subset of topics of interest; (2) inspect the characteristics of one topic; (3) compare different topics' trends.

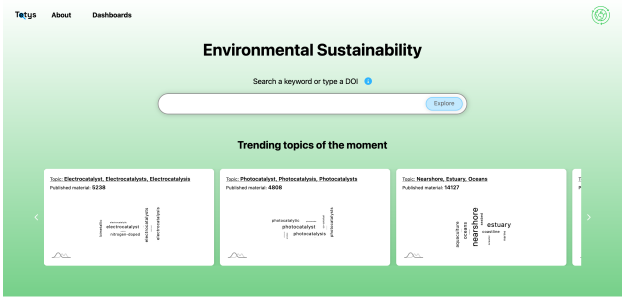
Scenario (1). Suppose that we have processed the five datasets on SDGs through the Topic Extraction Process and that we have completed the related Project Cards with the appropriate paths


Tetys
About
Dashboards
Tetys
Making data and trends of academic research publications accessibles to everyone.

Choose the area you want to investigate

Basic Human Needs and Well-being
Economic Development and Employment
Environment Sustainability
Equality and Social Inclusion
Global Partnership and Peace

Figure 2: Dashboard selector page with five available dashboards, one per *Project* covering a specific SDGs macro-area.


Tetys
About
Dashboards
Environmental Sustainability
Search a keyword or type a DOI
Explore
Trending topics of the moment
Topic: Electrocatalyst, Electrocatalysts, Electrocatalysis
Published material: 5238

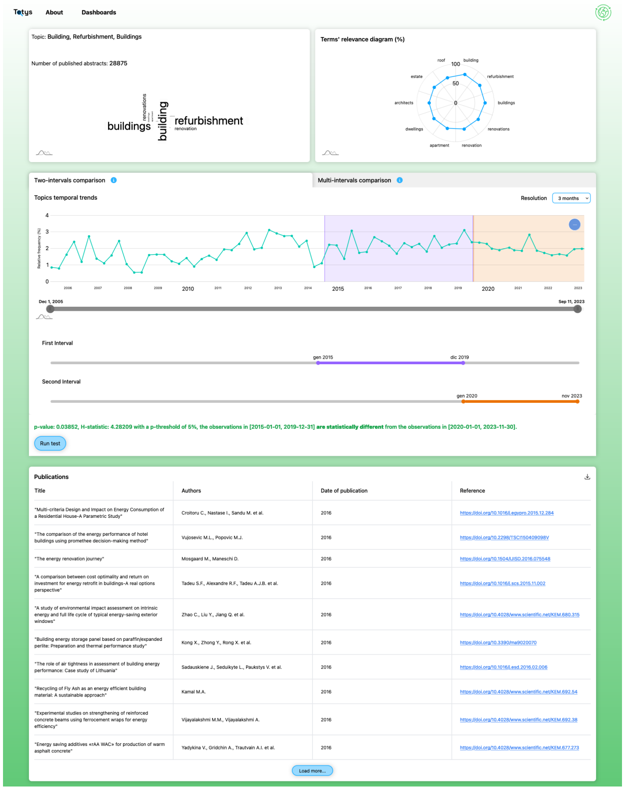
Topic: Photocatalyst, Photocatalysis, Photocatalysts
Published material: 4808

Topic: Nearshore, Estuary, Oceans
Published material: 14127

Figure 3: Dashboard home page for the "Environmental Sustainability" project.

and descriptions of the five corpora; the TETYS dashboard is ready for exploration. By accessing it at http://gmql.eu/tetys, we land on the Dashboard selector (Fig. 2), showing one colored button for each processed/imported project: red for "Basic Human Needs and Well-being", yellow for "Economic Development and Employment", green for "Environmental Sustainability", purple for "Equality and Social Inclusion", and blue for "Global Partnership and Peace". In the top-left corner, we find a "TETYS" button that redirects to the home page, "About" reports information on the funding, and "Dashboards" leads to a description of available projects. Suppose that we are interested in studying the evolution of research trends in the area of environmental sustainability. We can open the dedicated dashboard by clicking the green button, which takes us to that project's Dashboard home page (Fig. 3). Here, we notice the carousel with trending topics of the moment, i.e., topics with a number of documents within the top-ten ranking for the last month, as included in the original dataset. For each topic, an interactive word cloud with the most relevant terms is displayed and the total number of related documents is reported. If none of the trending topics was of interest to us, we can continue our search by writing keywords to find the most relevant topics related to our search. Alternatively, we can use the DOI of a publication as a search query and find the most relevant topics for that publication.

Scenario (2). SDG 11 sets targets and defines indicators towards sustainable cities and communities [20]. Suppose we are specifically interested in comparing the intensity of sustainable building renovations-related matters in scientific publications between 2023 and ten years ago (i.e., when SDGs were defined). We can use TETYS to analyze this phenomenon: after selecting the "Environmental Sustainability" dashboard, we can type the query 'building renovations' in the search bar. We see that the first topic in the search results has a relevance of 84% with our query and, in the topic's card, the word cloud contains terms such as "buildings", "refurbishment", "apartment", and "renovations", which correspond to our interest. By clicking on the topic's title we open its corresponding Single topic page (Fig. 4) to inspect its characteristics, such as KPIs (e.g., the number of published abstracts, the growth rate, etc.), terms, and its intensity over time. A table at the bottom of the page presents an expandable list of publications concerning the topic. The resolution of the time series of the intensity can be set to '3 months' to reduce the noise in the visualization. We notice that in the five years after 2015, the intensity has a slightly increasing trend, while in the last four years, it has a constant trend; With the statistical toolbox available under the line plot, TETYS allows us to verify if the difference in these two periods is statistically significant. We run a two-interval comparison test by selecting the two aforementioned intervals and clicking on 'Run test'. At the bottom of the statistical toolbox, a message tells us that the observations for the two intervals are statistically different, with a p-value of 0.038, below the 0.05 threshold. We conclude that the introduction of SDG 11 is likely correlated with the intensity of renewable building renovations in the scientific literature.



Topic: Building, Refurbishment, Buildings

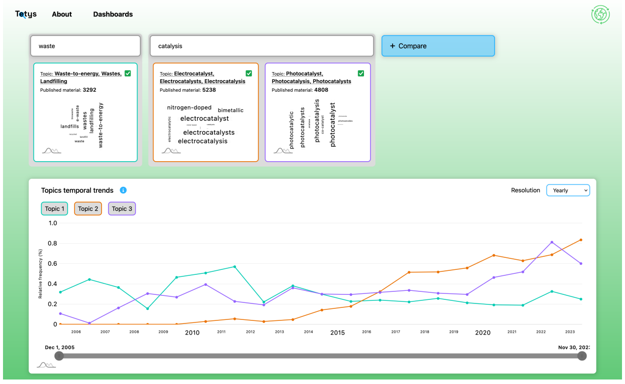
Number of published abstracts: 28875

p-value: 0.03852, H-statistic: 4.28209 with a p-threshold of 5%, the observations in [2015-01-01, 2019-12-31] **are statistically different** from the observations in [2020-01-01, 2023-11-30].

Title	Authors	Date of publication	Reference
"Multi-criteria Design and Impact on Energy Consumption of a Residential House-A Parametric Study"	Croitoru C., Nastase I., Sandu M. et al.	2016	https://doi.org/10.1016/j.egypro.2015.12.284
"The comparison of the energy performance of hotel buildings using promethee decision-making method"	Vujosevic M.L., Popovic M.J.	2016	https://doi.org/10.2298/TSCI150409098V
"The energy renovation journey"	Mosgaard M., Maneschi D.	2016	https://doi.org/10.1504/IJISD.2016.075548
"A comparison between cost optimality and return on investment for energy retrofit in buildings-A real options perspective"	Tadeu S.F., Alexandre R.F., Tadeu A.J.B. et al.	2016	https://doi.org/10.1016/j.scs.2015.11.002
"A study of environmental impact assessment on intrinsic energy and full life cycle of typical energy-saving exterior windows"	Zhao C., Liu Y., Jiang Q. et al.	2016	https://doi.org/10.4028/www.scientific.net/KEM.680.315
"Building energy storage panel based on paraffin/expanded perlite: Preparation and thermal performance study"	Kong X., Zhong Y., Rong X. et al.	2016	https://doi.org/10.3390/ma9020070
"The role of air tightness in assessment of building energy performance: Case study of Lithuania"	Sadauskiene J., Seduikyte L., Paukstys V. et al.	2016	https://doi.org/10.1016/j.esd.2016.02.006
"Recycling of Fly Ash as an energy efficient building material: A sustainable approach"	Kamal M.A.	2016	https://doi.org/10.4028/www.scientific.net/KEM.692.54
"Experimental studies on strengthening of reinforced concrete beams using ferrocement wraps for energy efficiency"	Vijayalakshmi M.M., Vijayalakshmi A.	2016	https://doi.org/10.4028/www.scientific.net/KEM.692.38
"Energy saving additives «rAA WAC» for production of warm asphalt concrete"	Yadykina V., Gridchin A., Trautvain A.I. et al.	2016	https://doi.org/10.4028/www.scientific.net/KEM.677.273

Figure 4: Single topic page for topic '*Building, Refurbishment, Buildings*' of the "Environmental Sustainability" project.

Scenario (3). Suppose we wish to assess the research trends related to SDG 12 (*Responsible Consumption and Production*). This SDG targets 'the efficient management of our shared natural resources and the way we dispose of toxic waste and pollutants' [20]. TETYS can support us in this task. We open the "Environmental Sustainability" dashboard and search for related topics. In our first attempt, we type 'waste' in the search bar and



waste

catalysis

Topic: Waste-to-energy, Wastes, Landfilling ☑
Published material: 3292

Topic: Electrocatalyst, Electrocatalysts, Electrocatalysis ☑
Published material: 5238

Topic: Photocatalyst, Photocatalysis, Photocatalysts ☑
Published material: 4808

Resolution: Yearly

Figure 5: Topic Comparison page for topics '*Waste-to-energy, Wastes, Landfilling*, '*Electrocatalyst, Electrocatalysis*', and '*Photocatalyst, Photocatalysis*'.

Group	Mental demand	Physical demand	Temporal demand	Performance	Effort	Frustration	WL
Students	46.2	18.5	30.8	32.3	42.3	28.5	33.1
Professionals	43.8	13.8	32.5	21.3	36.3	33.8	30.2
Average	45.3	16.7	31.4	28.1	40.0	30.5	32.0

Table 2: NASA-RTLX scores for the two groups and average.

select the '*Waste-to-energy, Wastes, Landfilling*' topic for comparison, by ticking the checkbox on its card. We can repeat the search with other keywords and select up to five topics for comparison; to do so, we click 'Compare' and type in the new search bar the keyword 'catalysis', a common chemical technology employed in a wide range of waste and pollutants management activities [4]. From the result, we select two topics - '*Electrocatalysis*' and '*Photocatalysis*' - and include them in the comparison. The Topics temporal trends box shows the line plot of their intensity over time (Fig. 5). We observe that the intensity of the '*Waste-to-energy, Wastes, Landfilling*' topic is quite homogeneous over time, while '*Electrocatalysis*' and '*Photocatalysis*' show a noticeable increment in the last five years, denoting an increased interest in these two topics in the scientific community.

4 EVALUATION WITH USERS

We conducted a preliminary small-scale task-based evaluation with users to understand how effectively they could interact with a TETYS dashboard. For this evaluation, we loaded the "Basic Human Need and Well-being" model (the first extracted for this demonstration, see Table 1). The user sample comprised 13 students and 8 professionals, participating on a voluntary base, both online and in presence. After a brief introduction to the project and the tool, we asked the participants to autonomously perform tasks related to finding insights on viral vaccines. Then, we asked them to respond to an anonymous questionnaire to rate their experience with the tool. To evaluate the workload required in the interaction we used the NASA Raw Task Load Index [8]; we asked participants to rate from 1 to 10 each of six dimensions (see Table 2). The overall Workload (WL) average across the two groups was ∼32 (on a 0-100 scale), slightly above the 'Medium' level. We bring out the low value for *Temporal Demand*, which validates our efforts in making a platform that provides users with insights on a corpus of documents in remarkably short time. To evaluate the usability, we employed the System Usability Scale [3]; on a set of 10 questions, participants were asked to answer according to a 5-point Likert scale. By averaging the scores on Usability and Learnability we achieved a 68.9 score on the Students group and 71.3 on the Professionals, slightly over the state-of-art recognized average (i.e., 68).

5 CONCLUSIONS

We demonstrated how TETYS can be applied to big corpora to extract insights, profile content dynamics, and map topics to documents, in a completely unsupervised fashion and leveraging fast data management and processing. We used input datasets that differ broadly in terms of semantics; our platform responded effectively to this diversity, indicating flexibility in being easily applied to different domains; moreover, the user study has shown promising results in usability. As an open-source platform, TETYS is an indispensable tool in democratizing content profiling, empowering a wide community to analyze and understand large textual datasets without barriers.

ACKNOWLEDGMENTS

This work is supported by TEThYS, a beneficiary of the EU-funded NGI Search project (Sub-grant Agreement SEARCH OC2_18).

REFERENCES

[1] amCharts. 2024. amCharts 5. https://www.amcharts.com/.
[2] Anna Bernasconi et al. 2024. TETYS: Towards the Next-Generation Open-Source Web Topic Explorer. In *CEUR PROCEEDINGS*, Vol. 3692. CEUR-WS, 26–33.
[3] John Brooke. 1996. SUS: A quick and dirty usability scale. *Usability Evaluation in Industry* (1996).
[4] Christian Candia-Onfray et al. 2021. An updated review of metal–organic framework materials in photo (electro) catalytic applications: From CO2 reduction to wastewater treatments. *Current Opinion in Electrochemistry* 26 (2021), 100669.
[5] Chart.js. 2024. Chart.js. https://www.chartjs.org/.
[6] Wenwen Dou et al. 2011. Paralleltopics: A probabilistic approach to exploring document collections. In *2011 IEEE Conf. on visual analytics science and technology (VAST)*. IEEE, 231–240.
[7] Apache Foundation. 2024. Apache Parquet. https://parquet.apache.org/.
[8] Mattias Georgsson. 2019. NASA RTLX as a novel assessment for determining cognitive load and user acceptance of expert and user-based evaluation methods exemplified through a mHealth diabetes self-management application evaluation. In *pHealth 2019*. IOS Press, 185–190.
[9] Matthew Laurence William Graham et al. 2023. TweetVi: A Tweet Visualisation Dashboard for Automatic Topic Classification and Sentiment Analysis. In *Int. Conf. on Asian Digital Libraries*. Springer, 159–166.
[10] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).
[11] Hugging Face. 2024. Hugging Face platform. https://huggingface.co/.
[12] Francesco Invernici et al. 2024. Capturing research literature attitude towards Sustainable Development Goals: an LLM-based topic modeling approach. *arXiv preprint arXiv:2411.02943* (2024).
[13] Francesco Invernici, Anna Bernasconi, and Stefano Ceri. 2024. Exploring the evolution of research topics during the COVID-19 pandemic. *Expert Systems with Applications* 252 (2024), 124028.
[14] Nahyun Kwon et al. 2023. Weedle: Composable Dashboard for Data-Centric NLP in Computational Notebooks. In *Companion Proceedings of the ACM Web Conf. 2023*. 132–135.
[15] Rui Meng et al. 2024. SFR-Embedding-2: Advanced Text Embedding with Multi-stage Training. https://huggingface.co/Salesforce/SFR-Embedding-2_R.
[16] Kawa Nazemi and Dirk Burkhardt. 2019. Visual analytics for analyzing technological trends from text. In *2019 23rd Int. Conf. information visualisation (IV)*. IEEE, 191–200.
[17] Mark Raasveldt and Hannes Mühleisen. 2019. Duckdb: an embeddable analytical database. In *Proceedings of the 2019 Int. Conf. on Management of Data*. 1981–1984.
[18] Suzanna Sia et al. 2020. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too!. In *Proceedings of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. 1728–1736.
[19] Tiangolo. 2024. FastAPI. https://fastapi.tiangolo.com/.
[20] UNDP. 2024. Sustainable Development Goals. https://www.undp.org/sustainable-development-goals.