

TETYS: Towards the Next-Generation Open-Source Web Topic Explorer

Anna Bernasconi^{1,*}, Francesco Invernici¹ and Stefano Ceri¹

¹Department of Electronics, Information and Bioengineering – Politecnico di Milano, Milan, Italy

Abstract

Users who search the web for specialized content typically lack knowledge of the precise topology of the dataset upon which the search is performed. Funded by European Union, TETYS is a beneficiary of the Next Generation Internet (NGI) Search Initiative; it proposes to build the next-generation open-source Web topic explorer. Our architecture inspects big textual corpora; it is composed of 1) a pipeline for ingesting huge data corpora, extracting highly relevant topics, clustered along orthogonal dimensions; and 2) an interactive dashboard, supporting topic visualization as word clouds and exploration of temporal series, with easy-to-drive statistical testing. The first prototype, CORToViz, explores the COVID-19 dataset (COVID-19 / SARS-CoV-2 virus research abstracts). Many different domains will be explored using TETYS (e.g., climate change and controversial debates on social media).

Keywords

Big Data Analytics, Scientific Literature, Natural Language Processing, Topic Modeling, Time Series

Project Information Project Name: Topics Evolution That You See (TETYS); Funding Agency: NGI Search (that received funding from the European Union’s Horizon Europe research and innovation programme under the grant agreement 101069364 and is framed under Next Generation Internet Initiative); URL: <https://annabernasconi.faculty.polimi.it/project-tetys/>; Running Period: 01.09.2023 – 31.08.2024; Team: Anna Bernasconi (Team Leader, Politecnico di Milano); Francesco Invernici (Politecnico di Milano); Stefano Ceri (Politecnico di Milano).

1. Project overview

Big text corpora require a very large amount of time to be parsed, processed, and summarized. In general, once a researcher first approaches a novel domain, she cannot grasp a general view of the research content that has been produced in a time-efficient manner.


TETYS (Topics Evolution That You See) proposes to build a powerful and generic architecture for inspecting the relevant topics of a big textual corpus, with an associated dashboard for one-click visualization of topic trends, modeled as time series; the dashboard also supports easy-to-drive statistical testing. TETYS is composed of two parts: a pipeline for ingesting huge data corpora, built upon state-of-the-art technologies (including large language models), and extracting from them highly relevant topics, clustered along orthogonal dimensions; and an


Research Projects Exhibition @CAiSE’24, June 3-7, 2024, Limassol, Cyprus

*Corresponding author.

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

ORCID  0000-0001-8016-5750 (A. Bernasconi); 0000-0001-6062-5174 (F. Invernici); 0000-0003-0671-2415 (S. Ceri)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

interactive dashboard, activated after a data preparation stage, supporting topic visualization as word clouds and their exploration through user-friendly interaction.

The TETYS concept is already fully demonstrated in the CORToViz prototype, which explores the CORD-19 dataset [1] collected during the pandemic, focused on COVID-19 and the SARS-CoV-2 virus (see Section 2). An unbounded number of interesting domains can be explored using the TETYS approach, including climate change and controversial debates on social media.

TETYS is a beneficiary of the 2nd Open Call of the NGI Search, a cascade funding project designed to help applicants (academic researchers, hi-tech startups, and SMEs) to adopt and develop *open source* digital innovation in the domain of searching and discovering data or other resources on the internet. Starting our path within the program in September 2023 with a Technology Readiness Level (TRL) of 3, NGI Search is supporting us to develop the testbed into a solid architecture, reaching TRL 5.

1.1. Scope

As a first testbed, we focus on research documents' datasets, considering all the text reviewed and published in edited proceedings and journals. By consolidating our first demonstrator, we aim at making the pipeline applicable to any corpus of Web textual documents and then validating the dashboard experience with solid and broad user studies. Subsequent TETYS dashboards will be used in the context of user search on widely adopted platforms, whose data can be freely accessed and processed for purposes of visualization. We build completely stateless applications, without saving any client session data or their search details on back-end databases. All results will be shared under open licenses; TETYS will contribute to fostering the diffusion of knowledge through the Open Science paradigm.

1.2. Work Plan (spanning 12 months)

WP1.1 Prototype refinement; Verification of the main technology employed [M1-3]

WP1.2 Deployment of cross-platform technology and test on different platforms/browsers [M4-6]

WP2 User-centred validation using the testbed implementation (A/B testing, Demand Validation Tests, Moderated Usability Study, etc.) [M1-6]

WP3.1 Technology consolidation based on feedback; Deployment of the advanced prototype [M7-9]

WP3.2 Selection of other test cases and experimentation, porting the prototype to other domains [M10-12]

WP4 Business plan preparation [M7-12]

Milestone 1 (end of M9): definition of user validation results (detailed report) and consolidation of the initial testbed (manual and documentation)

Milestone 2 (end of M12): production plan (redaction of documents) and multi-domain demonstrator, showing that TETYS can be quickly reapplied on unexplored domains (demo pilot)

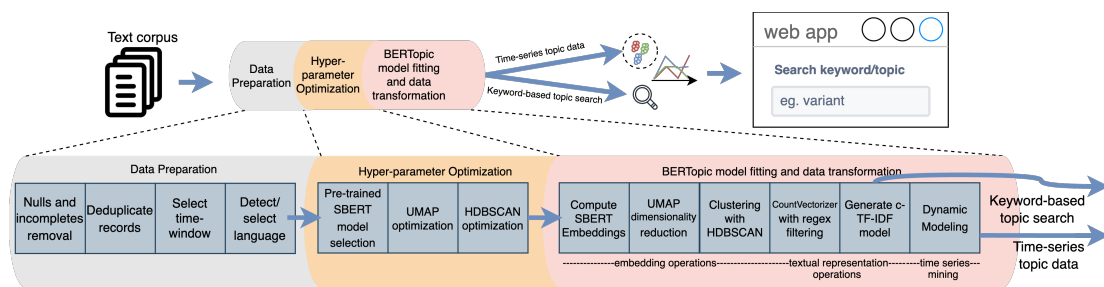


Figure 1: Steps of the CORToViz prototypal architecture

2. First prototype

CORD-19 has enabled many text mining approaches [2], leading to remarkable results [3], building for instance knowledge graphs for research acceleration [4, 5] and drug repurposing [3], resource annotation services [6, 7], claim verification systems [8], and purpose-specific language models [9]. Since May 5th, 2023, the pandemic is no longer considered a public health emergency by the Worlds Health Organization [10]; then, we may finally consider it as a concluded phenomenon and therefore analyze its history as a whole.

In this direction, the first TETYS prototype, named CORToViz (CORD-19 Topic Visualizer, online at <http://gmql.eu/cortoviz>) aims to show how the big literature corpus CORD-19 can be successfully exploited to gather a comprehensive overview of the pandemic, tracing the trends that have characterized its scientific literature narrative. Our first results are described in a journal publication [11], where the software architecture shown in Figure 1 is described, comprising modules for data preparation, hyper-parameter optimization, fitting of the topic model, with final data transformation for enabling time series visualization and keyword search functionalities. CORToViz offers insight into the various topics that have defined the global COVID-19 crisis, their interactions, and temporal dynamics. Furthermore, we demonstrate the advantages of employing a statistical approach for dynamic topic modeling based on results from deep learning-based language models. This prototype utilizes the most comprehensive dataset available to date about research on COVID-19 and SARS-CoV-2.

The code of the prototypal architecture of CORToViz is available on Docker (<https://hub.docker.com/r/frinve/cortoviz/>) and the code is open source on GitHub (<https://github.com/FrInve/TETYS/>), under license BSD 3-clause that permits distribution, changes, and commercial/private use. CORToViz addressed the community of COVID-19 researchers; TETYS will be applicable to any corpus of Web documents, reaching the wider community of Web users who explore textual content.

3. Excellence

TETYS contributes to addressing the **need** of users who search the web for specialized content. They typically use buzzwords and start searching with just one of them, then exploiting preliminary results to fuel following search iterations, thereby enriching the query set. This

process is required by the obvious lack of knowledge of the precise topology of the dataset upon which the search is performed. More importantly, it generally happens that users do not know anything about the temporal context in which specific words have been used and which topics were more trending at what time. Can a data-driven statistical/visualization-based strategy be introduced to make Web users 1) *more aware of the semantic content of the analyzed dataset* and 2) *understand the temporal evolution of the topics in the domain they are exploring*? Word clouds and topic modeling have been around for some time for representing documents of big corpora, but no full-stack approaches have been thus far proposed for working on large-scale evolving domains.

TETYS has the **ambition** to make research information more accessible to common Web users who access very technical and specialistic contributions, primarily textual data with semantics. At present times, topic modeling and – more in general techniques of machine learning-based natural language processing – are proposed in considerably customized cases and are applied successfully in specific and circumscribed contexts. With TETYS, we propose a completely general full-stack process that is virtually applicable to any corpus of medium-sized textual documents, using any topic model of choice, and a time-series visualizer. TETYS aims to address the two problems expressed above, providing a tangible way for Web users to 1) *explore semantic contents of big document corpora*; 2) *appreciate the temporal evolution of topics*. More importantly, it allows building dashboards for achieving (1) and (2) very quickly thanks to a lightweight technology stack applicable to any Web textual document dataset.

Regarding competitors, several tools and platforms employ topic modeling on top of existing datasets. These have already been used in business-oriented settings for their ease of use and effective outputs. For instance, MonkeyLearn implements several algorithms that can be customized without coding; MarketMuse helps with content optimization and automates content audits; DataScienceCentral offers tutorials and resources on various angles of data science, including topic modeling; and Gensim and Spacy are examples of popular NLP libraries for topic modeling.

BERTopic [12] is a recent Python library to create dense and interpretable topics; it is exploited within TETYS, as a core component for topic extraction. In terms of proposing an **innovative solution**, the novelty of TETYS stands in bringing a *one-click solution to apply this data science technique on virtually any possible document corpus*, allowing lightweight analytics at the service of Web searchers. Here, BERTopic – or any other topic model – can be embedded within TETYS.

Specifically within COVID-19-related matters, other works have previously focused on topic analysis: some analyzed the early stages of the pandemic [13, 14], others analyzed the broader field of coronaviruses [15], focused on topic distribution by country [16] or on the delineation and impact in scientometric terms of the early COVID-19 [17]. The approach conducted in the CORToViz prototype is broader, as it applies to the entire pandemic history without choosing a specific field of investigation *a priori*.

4. Impact

In terms of **scientific impact**, TETYS will contribute to fostering the diffusion of knowledge through the Open Science paradigm. TETYS will enable improved access to summarized and digested content both on domain-specific Web content (e.g., reviews on water-scoping machines) and more domain-general ones (e.g., climate change reports). Similarly, it will be applied to both highly technical texts (e.g., scientific research abstracts) and general texts (e.g., book reviews). The flexibility of the approach makes it applicable to solving needs in very diverse domains and markets. TETYS will be able to address the very diverse needs of stakeholders requiring a one-click stack that provides immediate high-level analytics on the topics of the text that are relevant to that field. This can potentially be exploited to quickly grasp trends in -for instance- e-commerce reviews, events feedback tweets, public engagement threads, or any other business in which observing temporal trends is crucial.

TETYS will bring an immediate positive **impact on the communication strategy** in multiple fields. We are building a second prototype on climate change-related research abstracts (and social media text, in parallel), so as to study the evolution in the interest in the topic from the research community. Studies of this kind can immediately stimulate public debate and awareness of environmental changes. TETYS will also help improve private/public services to meet relevant environmental policies or goals; indeed, environmental policy or planning decisions can be informed by the evidence collected through TETYS. For instance, this principle can be applied in public engagement in public works or urban planning (e.g., to quickly scan emails from citizens and grasp the trend in their proposals and interests). We will contribute to increasing levels of engagement of members of the public with research, and corresponding levels of confidence in public science dialogue.

TETYS can be configured as an add-on to several search engines or social media platforms. It will not save any information on the user side but only employ data from the platform that the user – eventually with a personal login – already has access to. Quick elaboration of topics contained in the text documents on the platform, resulting in powerful immediate visualization, will be an **added value to platforms**, increasing the engagement of users and prospectively bringing more interest toward them, thus economic value. Prospectively, we envision a TETYS browser extension that improves the search of known engines (e.g., Google Chrome has ChatGPT, Google Scholar, InvisibleHand, or Google Similar Pages add-ons). Notably, no extensions that analyze topics and their evolution currently exist for common browsers, making the approach of TETYS highly promising and impactful. TETYS will only work on selected domains that require a pre-processing of the Web content to allow for a quick reaction to user inputs.

5. Project status

The CORToViz prototype has been described in [11] and shared at <http://gmql.eu/cortoviz/>; we are expanding the scope of the documents' dataset encompassing major research publication editors' data (namely, Scopus and Springer). We are preparing the first dashboards regarding research macro areas (e.g., regarding climate change-related topics or inequalities/inclusion in

research).

In parallel, we are designing a completely novel front-end experience targeted to specific personas with research and communication backgrounds, while conducting market landscaping to understand key players in our (niche) industry.

The project has been presented at the SFSCON conference (South Tyrol Free Software Conference in Bozen, 2023) gathering the interest of a wide audience of open-source software practitioners; we have also been interviewed in the context of many Community events with NGI Search partners.

The one-year project duration is a short period; it is been dedicated to guaranteeing that the idea stands on solid technological and business bases, while the development of the core contribution will continue longer in our team, leveraging on the obtained funding.

6. Relevance to CAiSE

The TETYS project connects with the ‘*Artificial Intelligence and Machine Learning*’ CAiSE area of interest. Specifically, our approach relies on a meticulous selection of modern technologies, including large language models. This leads to developing an architecture tailored for clustering articles across orthogonal dimensions and employing extraction techniques for temporal topic mining. In this sense, we strengthen our connection with the ‘*Big Data architectures*’ CAiSE topic. Our focus is not on proposing a new model per se but on engineering a pipeline for identifying topics in a big corpus of text documents and exploring their temporal trend with an interactive interface. The components of the pipeline are carefully crafted and extended from the ones of an existing topic modeling framework, BERTopic. Additionally, we also benefit from having a statistical approach for dynamic topic modeling built on the results of deep learning-based language models – this feature contributes in the area of ‘*Big Data, Data Science and Analytics*’, being able to handle large sizes of text in short times. Approximately, the whole pipeline, ingesting 1 million abstracts, performing filtering, preparation, learning topic representation, classifying abstracts, building time series, and visualizing the results in the final dashboard, requires less than 20 hours.

7. Outlook

CORToViz is the first research demonstrator showcasing the TETYS approach. TETYS will consolidate this demonstrator, making the pipeline applicable to any corpus of Web textual documents, and validating the dashboard experience with solid and broad user studies. TETYS brings Web topic modeling from its current restricted research scale to the next level, making it applicable on a pilot scale on several other domains of interest for Web users.

Acknowledgements.



Funded by the
European Union

TETYS research is a beneficiary of the NGI Search 2nd Open Call. Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them. Funded within the framework of the NGI Search project under grant agreement No 101069364.

References

- [1] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. M. Raymond, D. S. Weld, O. Etzioni, S. Kohlmeier, CORD-19: The COVID-19 open research dataset, in: K. Verspoor, K. B. Cohen, M. Dredze, E. Ferrara, J. May, R. Munro, C. Paris, B. Wallace (Eds.), Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Association for Computational Linguistics, Online, 2020.
- [2] L. L. Wang, K. Lo, Text mining approaches for dealing with the rapidly expanding literature on COVID-19, *Briefings in Bioinformatics* 22 (2021) 781–799.
- [3] Q. Wang, M. Li, X. Wang, N. Parulian, G. Han, J. Ma, J. Tu, Y. Lin, R. H. Zhang, W. Liu, A. Chauhan, Y. Guan, B. Li, R. Li, X. Song, Y. Fung, H. Ji, J. Han, S.-F. Chang, J. Pustejovsky, J. Rah, D. Liem, A. Elsayed, M. Palmer, C. Voss, C. Schneider, B. Onyshkevych, COVID-19 literature knowledge graph construction and drug repurposing report generation, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, Association for Computational Linguistics, Online, 2021, pp. 66–77.
- [4] E. Logette, C. Lorin, C. Favreau, E. Oshurko, J. S. Coggan, F. Casalegno, M. F. Sy, C. Monney, M. Bertschy, E. Delattre, et al., A machine-generated view of the role of blood glucose levels in the severity of COVID-19, *Frontiers in Public Health* 9 (2021) 695139.
- [5] C. Wise, M. R. Calvo, P. Bhatia, V. Ioannidis, G. Karypus, G. Price, X. Song, R. Brand, N. Kulkarni, COVID-19 knowledge graph: Accelerating information retrieval and discovery for scientific literature, in: Proceedings of Knowledgeable NLP: the First Workshop on Integrating Structured Knowledge and Neural Networks for NLP, 2020.
- [6] T.-H. K. Huang, C.-Y. Huang, C.-K. C. Ding, Y.-C. Hsu, C. L. Giles, CODA-19: Using a non-expert crowd to annotate research aspects on 10,000+ abstracts in the COVID-19 open research dataset, in: K. Verspoor, K. B. Cohen, M. Dredze, E. Ferrara, J. May, R. Munro, C. Paris, B. Wallace (Eds.), Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Association for Computational Linguistics, Online, 2020.
- [7] G. Serna García, R. Al Khalaf, F. Invernici, S. Ceri, A. Bernasconi, CoVEffect: interactive system for mining the effects of SARS-CoV-2 mutations and variants based on deep learning, *GigaScience* 12 (2023) giad036.
- [8] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or Fiction: Verifying Scientific Claims, in: Proceedings of the 2020 Conference on Empirical Methods

in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7534–7550.

- [9] D. Korn, T. Bobrowski, M. Li, Y. Kebede, P. Wang, P. Owen, G. Vaidya, E. Muratov, R. Chirkova, C. Bizon, A. Tropsha, COVID-KOP: Integrating emerging COVID-19 data with the ROBOKOP database, *Bioinformatics (Oxford, England)* 37 (2021) 586–587.
- [10] United Nations News, WHO Chief Declares End to COVID-19 as a Global Health Emergency, <https://news.un.org/en/story/2023/05/1136367>, 2023. Last accessed: April 8th, 2024.
- [11] F. Invernici, A. Bernasconi, S. Ceri, Exploring the evolution of research topics during the COVID-19 pandemic, *Expert Systems with Applications* 252 (2024) 124028.
- [12] M. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure, *arXiv:2203.05794* (2022). URL: <https://doi.org/10.48550/arXiv.2203.05794>.
- [13] Y. Zhang, X. Cai, C. V. Fry, M. Wu, C. S. Wagner, Topic evolution, disruption and resilience in early COVID-19 research, *Scientometrics* 126 (2021) 4225–4253.
- [14] B. X. Tran, G. H. Ha, L. H. Nguyen, G. T. Vu, M. T. Hoang, H. T. Le, C. A. Latkin, C. S. Ho, R. C. Ho, Studies of Novel Coronavirus Disease 19 (COVID-19) Pandemic: A Global Analysis of Literature, *International Journal of Environmental Research and Public Health* 17 (2020) 4095.
- [15] A. Pourhatami, M. Kaviyani-Charati, B. Kargar, H. Baziyad, M. Kargar, C. Olmeda-Gómez, Mapping the intellectual structure of the coronavirus field (2000–2020): A co-word analysis, *Scientometrics* 126 (2021) 6625–6657.
- [16] P. Berchiolla, S. Urru, V. Sciannameo, The effect of COVID-19 on scientific publishing in Italy, *Epidemiologia & Prevenzione* 45 (2021) 449–451.
- [17] G. Colavizza, R. Costas, V. A. Traag, N. J. van Eck, T. van Leeuwen, L. Waltman, A Scientometric Overview of COVID-19, *PLOS ONE* 16 (2021) e0244839.