

MISSIONE 4
ISTRUZIONE
RICERCA



FOSSR

Fostering Open Science in Social Science Research
Innovative tools and services to investigate economic and societal change

TEXT MINING IN AND FOR POLICY MAKING WITHIN FOSSR

FOSSR Policy Brief Series

Issue #2 | October 2024



Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



Consiglio Nazionale
delle Ricerche

OUTLINE

- 1 Introduction and questions
- 2 Topic addressed
- 3 How FOSSR can address:
Text Mining in the STI context
- 4 Policy evidence

Presentation

This policy brief highlights the advantages of using **Text Mining (TM)** to enhance the analysis of public policies. TM enables the **efficient analysis of large volumes of policy documents**, helping to identify **trends, key issues and the potential impacts of policy decisions**.

The FOSSR project will play a crucial role in evaluating and analyzing policies, including in the context of Science, Technology, and Innovation (STI). By integrating innovative tools, data, services and methods, **FOSSR aims to enhance research practices and policy evaluation**. Applying advanced analytical techniques such as machine learning, network science, and knowledge representation through the construction of ontologies, text mining within the FOSSR project will help explore and uncover hidden relationships within large textual corpora.

In this policy brief, we demonstrate the **potential of Text Mining in the STI context** by addressing the following thematic questions: 1) can text mining help evaluate the alignment of **R&D funding policies** with **Sustainable Development Goals (SDGs)**? 2) can text mining **track trends over time in environmental and societal themes** within research topics?

Stakeholders are encouraged to explore Text Mining as a powerful tool to enhance evaluation processes across various sectors. By adopting TM techniques, institutions can **gain deeper insights in their evaluations**, leading to more **data-driven decision-making** and improved outcomes.

1 Introduction and questions

In the digital transformation era, the rapid expansion of information, especially **online textual data**, requires **innovative approaches** for effective analysis. **Text Mining** has emerged as an important tool, allowing for the **systematic examination** of large **datasets** to inform policy decisions.

Recognizing the importance of these methods, institutions like the European Commission have created specialized resources such as the **Text Mining & Analysis Competence Centre** (European Commission, 2016). This Centre acts as a hub for text mining efforts within the Commission, addressing contemporary challenges and supporting more informed and timely **decision-making processes**.

This policy brief addresses the advantages of this new approach of facilitating more **effective analysis in public policies**, especially in the context of **Science, Technology and Innovations studies (STI)**.

Text Mining involves extracting valuable information from text using **machine learning, data mining, and natural language processing (NLP)**. Unlike traditional methods like questionnaires and manual content analysis, which are labor-intensive and susceptible to bias, **Text Mining allows for the efficient and accurate analysis of large document sets** (Wencker, 2019).

In recent years, the literature has increasingly highlighted the **innovative contributions of Text Mining to policy studies**. For instance, Ngai and Lee (2016) conducted a systematic literature review on the application of Text Mining in policy contexts, utilizing the framework established by Jann and Wegrich (2007), which outlines the stages of the policy-making cycle: (i) agenda setting, (ii) policy formulation and decision-making, (iii) implementation and (iv) evaluation.

Their review emphasizes the **growing adoption of Text Mining tools**, particularly in the **agenda-setting stage**, where these tools are increasingly used to support the identification and prioritization of policy issues.

Berryhill et al. (2019) further highlight the potential of **AI tools** to enhance the entire policy-making process, with **machine learning** significantly improving public sector efficiency and **aiding government decision-making**. Similarly, Höchtl et al. (2016) argue that big data analysis accelerates report generation and strengthens evidence-based policy-making. Rubinstein et al. (2016) demonstrate how Text Mining can shed light on overlooked areas and complement traditional research methods throughout the policy cycle.

Numerous case studies have shown that **Text Mining is a powerful tool across various policy areas**, providing deep insights and supporting data-driven decision-making. In the context of **Science, Technology, and Innovation (STI) studies**,

Text Mining allows us to explore several **key questions**, such as: i) What are the **emerging technologies** and the most pressing **social challenges**? ii) how are social challenges linked to **technological developments**? iii) how well **R&D funding policies** align with the **Sustainable Development Goals (SDGs)**? iv) Can Text Mining **monitor trends** over time in environmental and societal themes within research topics?

This policy brief will show the potential of this methodological approach in the field of STI policies, presenting **two examples of Text Mining applications**:

1. The **first example** concerns the **extraction** through Text Mining techniques of **information about gender equality** in a set of documents related to **research funding programs**. This extraction has been used to **map the SDG orientation** of research programs across various **Research Funding Organizations (RFOs)** in Europe.

2. The **second example** illustrates the use of Text Mining techniques on a **dataset containing titles and abstracts of PhD theses** published in Italy from 2008 to 2022. This approach will allow us to observe the **orientation and evolution of PhD research topics** over time on themes of collective interest, such as climate change.

2. Topic addressed

Text Mining is increasingly recognized as a valuable tool in the **field of public policies**. By systematically analyzing large volumes of textual data, such as reports, feedback and public documents, Text Mining enables policymakers to uncover patterns, identify challenges, and understand the factors influencing the success or failure of policies. This approach not only enhances the **transparency and accountability of public programs** but also supports evidence-based decision-making by transforming qualitative insights into actionable information. As a result, Text Mining plays a crucial role in improving the effectiveness and responsiveness of policy initiatives. Studies on R&D financing and on research themes orientation could shed light on how policies are changing, and help institutions effectively prioritize their economic, social and environmental goals. Using tools, data and methods provided by FOSSR, we present two examples of how Text Mining can support research policy. We will apply this technique to:

a) **extracting information** from research funding documents to **map gender equality (SDG 5)** orientation across European RFOs;

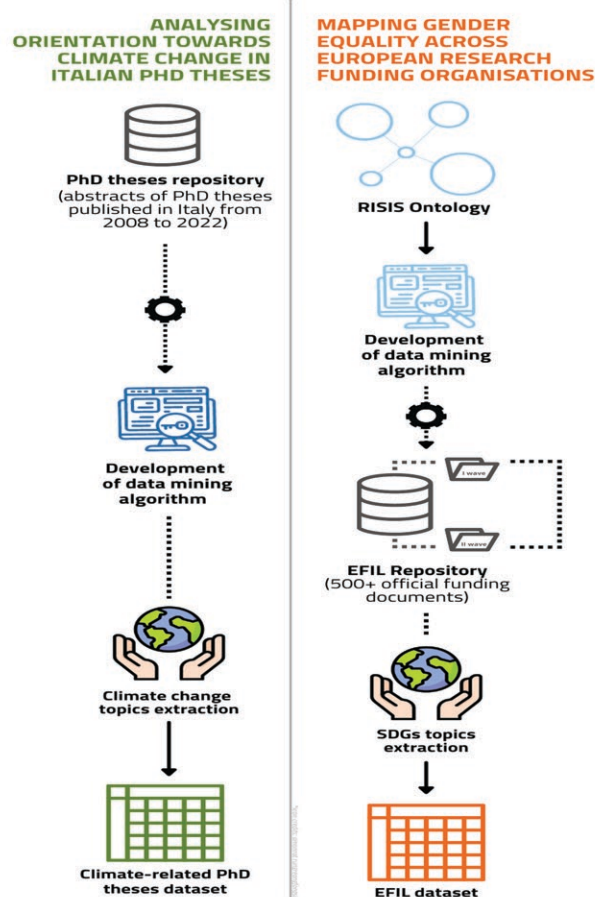
b) **analysing** the titles and abstracts of **Italian PhD theses** to track the research topics from 2008 to 2022, with a focus on the theme of climate change.

3. How FOSSR can address: Text Mining in STI context

In this paragraph, we use FOSSR data and services to illustrate **how Text Mining can aid research policy**, focusing on two examples: mapping the orientation toward gender equality (SDG 5) in European research funding and analyzing Italian PhD theses from 2008 to 2022 to track research topics related to climate change.

The **first example** is presented using the **European dataset of public R&D funding instruments within FOSSR**. **EFIL Dataset** aims at enabling users to investigate public R&D funding in Europe at the level of project funding instruments and Research Funding Organizations (RFO), addressing questions related to policy design and policy implementation (Reale et al., 2023). **EFIL store a repository of official documents** pertaining to funding instruments retrieved from RFO websites. Based on the official documents, an automated process (ontologies and statistical analysis) was used to generate **Sustainable Development Goals (SDGs) descriptors** in database (Zinilli et al., 2023). In particular, the selection of the SDGs is based on **RISIS2 ontology**.

Figure 1. SDGs Text Mining workflow



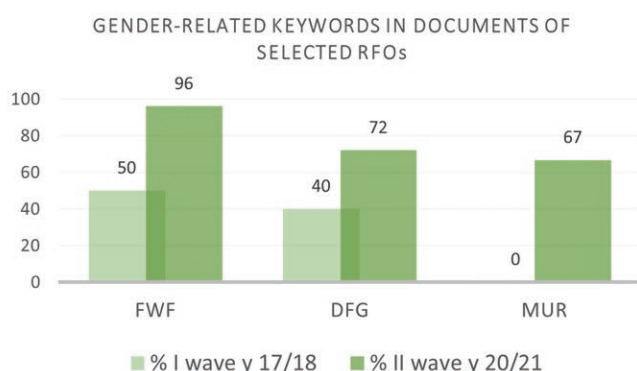
Understanding the focus and trends in academic research, particularly on **critical issues like climate change**, is crucial from a policy perspective for several reasons. First, it helps policymakers assess whether current research efforts are aligned with national and international priorities. Second, it enables the **identification of research gaps** that may require targeted funding or policy support to ensure that all important aspects of climate change are being adequately addressed. Lastly, by monitoring the evolution of research themes, **policymakers can anticipate emerging trends and challenges**, allowing them to develop forward-looking strategies that support long-term sustainability and innovation in addressing climate-related issues.

Case 1: SDGs in research funding instruments using EFIL dataset: a focus on gender equality goal

The analysis of SDG mapping enhances our understanding of how sustainability targets influence project funding instruments and aids European countries and Research Funding Organizations (RFOs) in better integrating SDGs into their research policy agendas. By applying Text Mining to over **500 extensive official funding documents stored in the EFIL repository**, we can track shifts over time in RFO orientations toward specific SDG goals. This automated process, particularly focused on SDG 5 on gender equality, provides **crucial insights into how RFOs align their strategies with sustainability objectives**, enabling more precise and impactful policy development.

The analysis of over 700 funding instruments from 55 European RFOs across ten countries, as monitored in EFIL, reveals a **17% increase in attention to the "gender dimension" between 2017–18 and 2020–21**. Austria (FWF) leads in promoting gender equality through competitive funding programs, followed by Germany (DFG), with Italy (MUR) showing significant improvement in the inclusion of gender considerations during the second wave.

Figure 2. Comparison of selected RFOs



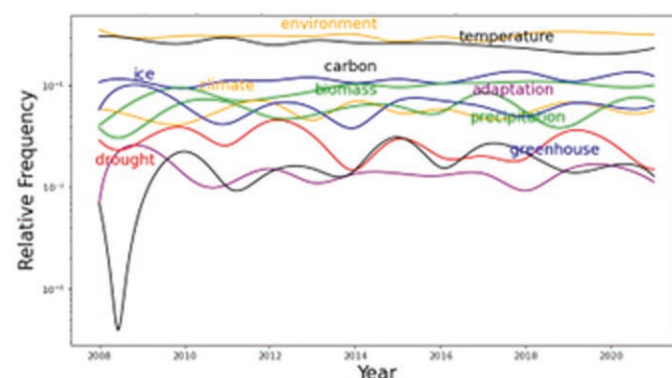
Text Mining plays a crucial role in uncovering these trends by systematically analyzing large volumes of policy documents, enabling the identification of subtle shifts in focus, such as the **growing emphasis on gender equality**. This approach allows policymakers and RFOs to track the effectiveness of their strategies over time and make data-driven decisions to enhance the inclusivity and impact of their funding programs.

By using **Text Mining to monitor and compare the integration of gender dimensions** across different RFOs and countries, institutions can not only ensure more equitable opportunities in research funding but also align their strategic planning with broader societal goals. This method empowers **RFOs to prioritize gender balance** in leadership positions, funding allocation, and policy development, ultimately contributing to a more diverse and inclusive research environment. The insights gained through Text Mining support the continuous improvement of research policies, ensuring they are responsive to evolving priorities and capable of fostering a more equitable and innovative scientific community.

Case 2: Text Mining applied on Italian PhD thesis

In this case, Text Mining is applied to a dataset containing **abstracts of PhD theses published in Italy from 2008 to 2022**, with a specific focus on climate change. By systematically analyzing this large body of academic work, we can observe how the orientation and emphasis of **PhD research on climate-related themes** have evolved over time.

Figure 3. Frequency of top key climate-related keywords (2008–2022)



Terms like **"environment"** and **"temperature"** remain **consistently high in usage**, reflecting their sustained importance in climate research. In contrast, terms such as **"drought"** show significant fluctuation, with a notable spike in the early years,

possibly linked to specific climate events. Words like "adaptation", "precipitation", and "greenhouse" show moderate but steady trends, indicating growing or sustained interest in these areas. The presence and variability of terms like "**carbon**", "**ice**", and "**biomass**" highlight ongoing attention to **specific climate issues** and their adaptation in research over time.

This analysis is particularly important for policymakers as it sheds light on **how** emerging scholars are contributing to the field and whether their **research aligns with national and global environmental priorities**. Additionally, understanding the **trends in PhD research** helps policymakers identify **potential gaps in knowledge**, allowing them to craft policies that not only support innovation but also promote a research setting that is responsive to the **most urgent environmental challenges**. This, in turn, contributes to the development of a well-prepared and knowledgeable workforce capable of driving forward climate policy and sustainability initiatives.

The following table outlines the **main research topics identified in climate-related PhD theses**, reflecting the diverse areas of focus in academic research from 2008 to 2022:

Table 1. Research topics in climate change in PhD theses

Main Research Topic
Climate Change and Human Impact
International Cooperation and Climate
Climate Change and Policy
Pollution and Ecosystems
Public Health and Safety in Crises
Industry, Pollution, and Environmental Impact
Global Warming
Climate Change: hoax or Reality
Meat Consumption, Agriculture, and Environmental Impact

This table reveals the **key areas of PhD research focused on the climate crisis**, highlighting critical intersections between international cooperation, policy, and environmental impacts. The emphasis on **topics such as global warming, pollution, and human impact** underscores the **urgent need for comprehensive policy frameworks** that address these interconnected issues. Research on public health and safety during and after extreme events further emphasizes the **necessity of integrating climate resilience into health and safety policies**.

The exploration of controversial topics, such as **climate change skepticism** and the **environmental impact of agriculture**, reflects the scientific interest of young researchers in **understanding the causes and effects of public perception** and acceptance of

climate policies. This focus is crucial for developing strategies that can **bridge the gap between scientific evidence and societal attitudes**, ultimately fostering more effective climate action.

4. Policy evidence

Stakeholders are encouraged to consider **Text Mining as a powerful tool** for enhancing evaluation processes across various sectors. By employing Text Mining techniques, **policy makers can gain deeper insights, enabling more informed decision-making and improved outcomes**. For example, analysing the research funding programs of EU countries and their alignment with the Sustainable Development Goals (SDGs) can have significant policy implications. Such studies can aid in **efficiently prioritizing the SDG agenda**, helping to develop policies and programs that are better aligned with the SDGs and **fostering international collaboration**. Policymakers can use these evaluations to **allocate resources** more effectively **to research programs** with the greatest **potential impact on sustainable development**, thus maximizing the use of limited resources and focusing efforts on the most pressing global challenges.

Through FOSSR services, policy makers can base their **strategies on concrete, timely evidence**, enhancing their ability to respond to complex and continuously evolving challenges like climate change or gender inequality. Furthermore, Text Mining allows for ongoing **monitoring of policy effectiveness**, promoting a more dynamic and **adaptive decision-making process** compared to traditional approaches.

TM not only increases transparency in policy-making but also **helps policymakers detect weak signals-emerging trends** or concerns that may not be immediately visible through other methods. For example, Text Mining can identify subtle changes in the language used in policy documents or scientific outputs, **signalling a potential shift in priorities**. Additionally, it enables the **exploration of interconnected policy areas**, revealing how issues like climate change and social justice are addressed in an **integrated manner across research and global debates**.

References

Albert, T., Moehle, M. G., & Meyer, S. (2015). Technology maturity assessment based on blog analysis. *Technological Forecasting and Social Change*, 92, 196-209.

European Commission: Competence Centre on Text Mining and Analysis.
https://knowledge4policy.ec.europa.eu/text-mining_en (2016).

Höchtel, J., Parycek, P., & Schöllhammer, R. (2016). Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce*, 26(1-2), 147-169.

Jann, W., & Wegrich, K. (2007). Theories of the policy cycle. In F. Fischer, G. Miller, & M. S. Sidney (Eds.), *Handbook of public policy analysis: Theory, politics, and methods*. CRC Press.
<https://doi.org/10.4324/9781315093192>

Ngai, E. W., & Lee, P. T. Y. (2016). A review of the literature on applications of text mining in policy making.

Reale, E., Spinello, A. O., Varinetti, E., & Zinilli, A. (2023). Documentation of RISIS datasets: EFIL (2.0). Zenodo.
<https://doi.org/10.5281/zenodo.7938134> (Citing a website)

Rubinstein, M., Meyer, E., Schroeder, R., Poel, M., Treperman, J., van Barneveld, J., ... & Svetachova, M. (2016). Ten use cases of innovative data-driven approaches for policymaking at EU level. Report prepared in the context of: Data for Policy: A study of big data and other innovative data-driven approaches for evidence-informed policymaking.

Wencker, T. (2019), Text Mining in Evaluation, DEval Policy Brief 1/2020, Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval), Bonn.

Zinilli A., Reale E., Spinello A.O., Varinetti E. (2023). Diversification vs specialization from the perspective of research programmes: A complexity approach. Atlanta Conference On Science and Innovation Policy. MAY 24-26, 2023. ATLANTA, GEORGIA, USA GEORGIA TECH GLOBAL LEARNING CENTER

FOSSR - Fostering Open Science in Social Science Research aims to become an Italian Open Science Cloud, along the lines of the European Open Science Cloud project, in which to integrate innovative services developed by the project for data collection, data curation and Fairness, and data analysis on economic and societal change. FOSSR fosters the building of an integrated knowledge sharing platform, a single point of access to all the tools and services made available by the Italian nodes of social science infrastructures: CESSDA, SHARE and RISIS adopt the common theme of the development of Open Science in the Italian context with the goal of creating a framework of tools and services for the social science scholar community.

FOSSR wants to promote toward multiple audiences, a widespread knowledge and awareness of the data and methodologies employed in empirical social science, fostering the growth of a broad societal environment favourable to further thriving of social science research in Italy, providing easy, open, streamlined access to social science data through innovative interfaces. The integration of this pool of resources shall concretely contribute to the realization of open science for scholars in social sciences, going with an important program of scientific training on methods and instruments for social science research based on FAIR empirical data.

FOSSR Policy Brief Series aim at communicate key findings from the FOSSR thematic network to a non-specialized audience with a strong emphasis on the demonstration of usage cases of FOSSR resources. The series can accomplish two goals: improving the use of data for evidence-based policymaking and assisting the stakeholders in making informed decisions.

Copyright FOSSR Project 2021

AUTHORS OF THE CURRENT ISSUE:

Antonio Zinilli, CNR-IRCrES

Emanuela Varinetti, CNR-IRCrES

EDITORIAL BOARD:

Emanuela Reale, CNR-IRCrES

Alessandra Maria Stilo, CNR-IRCrES

Serena Fabrizio, CNR-IRCrES

Rita Giuffredi, CNR-IRCrES

Mario Ciampi, CNR-ICAR

Andrea Nuzzolese, CNR- ISTC

Mario Paolucci, CNR-IRPPS

Antonio Zinilli, CNR-IRCrES

GRAPHIC DESIGN:

Serena Fabrizio, CNR-IRCrES

Rita Giuffredi, CNR-IRCrES

FOSSR WEBSITE

