Catalysts for progress? Mapping policy insights from energy research

Brian Boyle, Yen-Chieh Liao, Sarah King, Robin Rauner, and Stefan Müller
Energy Research and Social Science

This document outlines all R scripts used to collect data, conduct analysis, and produce the tables and figures in the main paper and online Supporting Information. Due to Scopus' data sharing policies, the texts of abstracts cannot be shared publicly. Please contact us if you would like to reproduce scripts that require abstract text.

The required R package versions are specified at the beginning of each script. If you encounter issues running the code, it may be due to package updates. The authors successfully executed all scripts on 8 February 2025. To replicate the analysis, you can install the specified R version or use the *groundhog* package.

We recommend creating an RProject in the directory with all data and scripts. Alternatively, you can use the *here* package or set the working directory with the `setwd()` function. It is advisable to download files in their original format from the Dataverse (select "Download" -> "Original File Format").

00_reproduce_analysis.R

This script runs all the following scripts in the correct order, ensuring the required packages (with specified CRAN versions) are installed. If you install the *rmarkdown* package, it will also generate html log files for each script, which are included in the Dataverse folder.

01_retrieve_abstracts.R

Code to call Elsevier Scopus API for top 100 journals covering energy as a subject area (done in November 2024). Keeps only journal articles with DOIs, titles, abstracts[1], affiliations, and authors.

Input:

[1]Due to Scopus' data sharing policy, the text of abstracts has been removed from publicly-available datasets.

- `energy_journals_2023_scimago.csv`: list of journals used to create API calls
- `scopus_journal_functions_cursor.R`: custom functions for Scopus data collection
- `article_freq_energy_journals_2010-2023.csv`: list of journals and number of articles
- `asjc_journal_codes.rds`: journal IDs and associated subject area codes

Output (not available on Dataverse):

- `journal_year`: folder containing scraped journals
- `scopus_energy_journals_1-50_2010-2023_combined_dontshare.rds`: top 50 journals
- `scopus_energy_journals_51-100_2010-2023_combined_dontshare.rds`: journals ranked 51-100
- `data_scopus_1_100_2010_2023_dontshare.rds`: top 100 energy journals

Output (available on Dataverse):

- `data_scopus_1_100_2010_2023_public.rds`: top 100 journals (without abstract text)

02_validate_dictionaries.R

Fine-tunes and validates a custom dictionary using F1 score.

Input: (not available on Dataverse)

- `data_abstracts_annotated_ground_truth_dontshare.csv`: stratified sample of 400 abstracts, double-annotated for general policy relevance
- `data_abstracts_annotated_held_out_dontshare.csv`: stratified sample of 200 abstracts, hand-annotated for general policy relevance

Note: We have made versions of the annotated data without abstract text available on Dataverse. These contain a unique identifier, `article_scopus_id`, that can be used to locate the article.

Input: (available on Dataverse)

- `dictionary_functions.R`: custom functions for dictionary validation
- `policy_terms.csv`: terms identified through qualitative review
- `keyness_terms_n100.csv`: terms identified through keyness analysis

Output:

- Table A2 (`tab_A2.tex` and `tab_A2.html`)
- Table A3 (`tab_A3.tex` and `tab_A3.hmtl`)
- Table A4 (`tab_A4.tex` and `tab_A4.html`)

03_classify_abstracts.R

Filters corpus using energy and net zero keywords before applying the custom policy dictionary to identify policy-relevant abstracts.

Input:

- `data_scopus_1_100_2010_2023_dontshare.rds`: corpus of abstracts from top 100 journals covering energy (with abstract text - not available on Dataverse)

Output:

- `data_scopus_1_100_2010_2023_classified.rds`: binary prediction of policy-relevant abstracts (without abstract text)
- `data_scopus_1_100_2010_2023_classified_dontshare.rds`: binary prediction of policy-relevant abstracts (with abstract text - not available on Dataverse)

04_analysis.R

Input:

- `function_theme_base.R`: custom ggplot2 scheme
- `data_scopus_1_100_2010_2023_classified.rds`: binary prediction of policy-relevant abstracts (without abstract text)
- `data_scopus_1_100_2010_2023_classified_dontshare.rds`: binary prediction of policy-relevant abstracts (with abstract text - not available on Dataverse)
- `journal_aims.csv`: hand-annotated variable for whether "policy/policies" is mentioned in journals' aims and scope

Output:

- Table 1 (`tab_01.tex` and `tab_01.html`)
- Figure 1 (`fig_01.png`, `fig_01.pdf`, and `fig_01.eps`)
- Figure 2 (`fig_02.png`, `fig_02.pdf`, and `fig_02.eps`)
- Figure 3 (`fig_03.png`, `fig_03.pdf`, and `fig_03.eps`)
- Table A1 (`tab_A1.tex` and `tab_A1.html`)
- Figure A1 (`fig_A1.png`, `fig_A1.pdf`, and `fig_A1.eps`)
- Figure A2 (`fig_A2.png`, `fig_A2.pdf`, and `fig_A2.eps`)
- Figure A3 (`fig_A3.png`, `fig_A3.pdf`, and `fig_A3.eps`)
- Figure A4 (`fig_A4.png`, `fig_A4.pdf`, and `fig_A4.eps`)
- Table A5 (`tab_A5.tex` and `tab_A5.html`)
- Figure A5 (`fig_A5.png`, `fig_A5.pdf`, and `fig_A5.eps`)

05_stm_topic_model.R

Conducts model diagnostics and applies a structural topic model (STM) with 10 topics to abstracts classified as policy relevant. Texts are preprocessed by removing numbers, punctuation, symbols, URLs, tokens less than 3 characters in length, common English stopwords, and stopwords specific to the academic abstracts. The 200 most frequent bigram collocations are also included as multi-word expressions. Rare tokens that appear less than 5 times are removed before conversion to a document-feature matrix.

Input:

- `function_theme_base.R`: custom ggplot2 scheme
- `data_scopus_1_100_2010_2023_classified_dontshare.rds`: binary prediction of policy-relevant abstracts (with abstract text - not available on Dataverse)

Output:

- Figure A6 (`fig_A6.png`, `fig_A6.pdf`, and `fig_A6.eps`)
- Table 2 (`tab_02.tex` and `tab_02.html`)
- `topdocs_expanded.csv`: top 25 documents for each topic
- Figure 4 (`fig_04.png`, `fig_04.pdf`, and `fig_04.eps`)

06_check_stats.R

Calculates all statistics mentioned in the manuscript in order of appearance.

Input: (not available on Dataverse)

- `data_scopus_1_100_2010_2023_dontshare.rds`: full collected corpus (before filtering for energy and net zero keywords)
- `data_scopus_1_100_2010_2023_classified_dontshare.rds`: binary prediction of policy-relevant abstracts
- `data_fulltexts_annotated_dontshare.csv`: sample of 50 full texts, hand-annotated for policy relevance
- `data_abstracts_annotated_ground_truth_dontshare.csv`: stratified sample of 400 abstracts, double-annotated for general policy relevance
- `data_abstracts_annotated_held_out_dontshare.csv`: stratified sample of 200 abstracts, hand-annotated for general policy relevance

Input: (available on Dataverse)

- `journal_aims.csv`: hand-annotated variable for whether "policy/policies" is mentioned in journals' aims and scope