The Evolution of Global Cleavages: A Historical Analysis of Territorial and Functional World Alignments Based on Automated Text Analysis, 1843−2020

Comparative Political Studies

Daniele Caramani, Siyana Gurova, and Tobias Widmann

2024-06-11

Readme File

Replication Material

This document describes the data and R script necessary to reproduce all plots and tables included in the paper and in the Appendix and Online Appendix. In the beginning of the replication script, the necessary packages, the versions they were used in, and the necessary data files are listed. We use the `groundhogR` package to load all required packages in the correct version.

We recommend that you set the working directory (using *setwd()*) to the folder downloaded from the CPS dataverse, which contains all necessary files.

To replicate the analyses of the manuscript and the appendices, simply run the script "001_replication_script.R", which will automatically save all figures and tables in the folders "figures" and "tables" respectively.

Descriptions of R scripts

- **001_replication_script.R**: This R script replicates all tables and graphs included in the paper. All tables and figures will be automatically saved in the respective folder ("figures" or "tables"). All necessary files are provided and will be loaded in automatically.

Descriptions of data files

- **model.Rdata**: Main model trained to predict the polarity score of the Economist article.
- **model_colony.Rdata**: Additional model trained to predict polarity scores of Economist articles, including additional seed words.
- **sample_human_coding.xlsx**: A sample of 300 articles hand-coded by a human coder which is being used to calculate model performance.
- **Information on Economist data**: This article is based on a dataset of The Economist articles between 1843 and 2020. The articles from The Economist were downloaded from the Gale Digital Scholar Lab, a digital archive that contains a vast array of primary sources including all issues of The Economist from the beginning of its existence in 1843 to 2020. Articles from The Economist's archive can be legally downloaded, mined and analysed without restriction. See https://www.gale.com/intl/primary-sources/digital-scholar-lab. The subscription was provided by the University of Zurich. Data and code are available for internal replication. Coding files are openly accessible, but due to the Gale Digital Scholar Lab's policy, we are not allowed to make the data publicly available. In Gale Digital Scholar Lab, articles are tagged and can be selected by their "Section" and/or "Type". The category "Section" allows users to select from multiple categories of topics: business, sports, news, opinions and editorials, people, preliminary matter, and advertisements. The option "Type" lists various types of material: both substantial material, such as articles, editorials, financial reports, letters, obituaries and reviews, and graphic material, such as front matter, cartoons, tables of content and advertisements. Articles with the following categories have been downloaded: "News" and "Opinions and editorials" in the tag "Section", and "Article" and "Editorial" in

the tag “Type”. Excluding items like advertisements, front matter and letters, and focussing on actual articles avoids introducing noise in the data. In addition to the articles’ text, we also downloaded the metadata associated with each item (title, date of publication, content type, publication type, identification number, etc.), which were then matched with the text items. This process produced a dataset of 316,749 textual items read into the open-source software R. By comparing the metadata files with the downloaded articles, we noticed that approximately 1,600 articles were missing for the year 1962. For another set of around 400 articles between 1903 and 1906, the date of publication was missing. Gale Digital Scholar Lab assisted us in fixing both issues so that these 2,000 items could be included in the dataset. Editorials are tagged incorrectly before 1920 making it impossible to identify them. We identified various patterns of errors in the textual data, such as missing letters and duplicates, and corrected them in the pre-processing phase. We dismissed articles with missing metadata and only considered articles longer than 150 words. Shorter articles were removed from the dataset because they typically include graphs or tables that introduce unnecessary noise into the analysis. These cleaning procedures reduced the total number of articles in the final dataset to 308,329.