Transparency materials for:
Media Consolidation and News Content Quality

November 26, 2024

Marcel Garz
Mart Ots

Summary of availability

All data are publicly available, except for the original newspaper content, which needs to be directly accessed through Retriever Mediearkivet due to copyright restrictions.

Details on each data source

- Newspaper content can be accessed from https://www.retrievergroup.com/sv/product-mediearkivet
- Newspaper circulation data ("dagstidningsförteckning") can be requested from the Swedish Media Authority at https://mediemyndigheten.se/om-oss/kontakt/
- Data on operational subsidies ("driftsstöd") can be obtained from the Swedish Media Authority at https://mediemyndigheten.se/stod-till-medier/beviljade-stod/

Software requirements

R 4.3.2, Python 3.1x, Stata 17

Inventory

- The folder "article predictions" includes:
    - The fine-tuned deep-learning models (one for each classifier, labeled "fine_tuned_model_[name of classifier]")

    - Juypiter notebooks used to implement the finetuning, get performance metrics, and get the machine predictions (one for each classifier, labeled “classification [name of classifier].ipynb”)
    - Machine predictions stored as CSV files (for computational reasons two files per classifier, labeled “predictions_[name of classifier]_1.csv” and “predictions_[name of classifier]_2.csv”)
- The folder “newspapers and owners” includes:
    - Subfolder “mprt raw data”, which includes PDF files on newspaper circulation and operational subsidies published by the Swedish Media Authority and its predecessor, the Swedish Press and Broadcasting Authority
    - “circulation data.dta”: the circulation data extracted by the authors from the original PDFs and stored as an intermediate Stata dataset
    - “consolidation measures.dta”: intermediate Stata dataset including indices of ownership concentration (compiled by the authors via “prepare ownership and circulation.do”)
    - “newspaper month data”: intermediate Stata dataset that combines information on newspaper ownership, circulation, operational subsidies, and mergers (compiled by the authors via “prepare ownership and circulation.do”)
    - “operating aid.xlsx”: information on operational subsidies, extracted by the authors from the original PDFs and stored as an intermediate Excel file
    - “outlets owners areas editions.xlsx”: overview over newspapers, availability of content data, entry, exit, and ownership changes (all manually compiled by the authors from media companies’ websites and press releases)
    - “prepare ownership and circulation.do”: Stata do-file used to compile “consolidation measures.dta” and “newspaper month data.dta”
- The folder “prepare training data” includes:
    - Subfolder “coded batches”, which contains content annotations stored as Excel files [removed for copyright reasons]
    - “coding instructions.pdf”: summary of coding instructions used by the annotators
    - Processed data training data (one CSV file per classifier, labeled “[name of classifier]_train.csv” [removed for copyright reasons]
    - “all_ids.dta”: intermediate Stata dataset containing IDs and content of annotated articles [removed for copyright reasons]
    - “prepare training data.do”: Stata do-file used to combine individual annotations and assemble training data CSV files
- The folder raw articles includes:
    - Subfolder “sets”, which includes the raw content data (TXT files) extracted from Retriever Mediearkivet [removed for copyright reasons]

        - “all_articles_for_pred_1.csv” and “all_articles_for_pred_2.csv”: intermediate CSV files including IDs and content of all articles [removed for copyright reasons]
        - “clean corpus and create batches.do”: Stata do-file used for cleaning of raw article data and creation of Excel files for annotation
        - “cluster1.dta”, “cluster2.dta”, “cluster3.dta”, “cluster4.dta”, and “cluster5.dta”: intermediate Stata datasets including processed article content [removed for copyright reasons]
        - “create corpus.R”: R script used to process the raw article content
- “article level quality data.dta”: Stata dataset including quality data on individual news articles (see below codebook)
- “conduct main analysis.do”: Stata do-file used to create tables and figures in the paper
- “create analysis datasets”: Stata do-file used to create “article level quality data.dta” and “outlet month quality data.dta”
- “outlet month quality data.dta”: Stata dataset including newspaper-month level data for main analysis

Steps to compile analysis datasets

1) Download raw article content from Retriever Mediearkiviet and save as TXT files in “raw articles/sets/”
2) Run “raw articles/create corpus.R”
3) Run “raw articles/clean corpus and create batches.do” part “draw sample and create batches for human coders”
4) Obtain content annotations by human coders
5) Run “newspapers and owners/prepare ownership and circulation.do”
6) Run “prepare training data/prepare training data.do”
7) Run “raw articles/clean corpus and create batches.do” part “get articles for prediction”
8) Run “article predictions/classification [name of classifier].ipynb”
9) Run “create analysis datasets.do”
10) Run “conduct main analysis.do”

Codebook for main analysis dataset I (“article level quality data.dta”)

- id: article identifier
- predicted: 1 if quality variables are based on machine predictions, 0 if based on human annotations
- type (article type):
    - LABEL_0 = journalistic content

        - LABEL_1 = non-journalistic content
- topic (topic relevance):
    - LABEL_0 = politics
    - LABEL_1 = economics/business
    - LABEL_2 = arts/culture
    - LABEL_3 = sports
    - LABEL_4 = human interest/other
- level (actor relevance):
    - LABEL_0 = macro
    - LABEL_1 = meso
    - LABEL_2 = micro (functional)
    - LABEL_3 = micro (other)
- context (thematic orientation):
    - LABEL_0 = thematic reporting
    - LABEL_1 = episodic reporting
- format (interpretative performance):
    - LABEL_0 = news flash
    - LABEL_1 = news report
    - LABEL_2 = list, agency report
    - LABEL_3 = portrait, interview, review
    - LABEL_4 = reportage
    - LABEL_5 = opinion article
- style (objectivity):
    - LABEL_1 = cognitive-normative reporting
    - LABEL_1 = moral-emotional reporting
- geo (geographical reference):
    - LABEL_1 = local/regional
    - LABEL_1 = national/bilateral
    - LABEL_2 = other country
    - LABEL_3 = international/multilateral

Codebook for main analysis dataset II ("outlet month quality data.dta"): see embedded Stata labels