Social Media and Political Agenda Setting

Description of Replication Material, Datasets, and R Scripts

Fabrizio Gilardi Theresa Gessler Maël Kubli Stefan Müller

In this file, we describe all datasets and R scripts required to rerun and reproduce the analyses reported in the paper and Supporting Information. In order to have the data, scripts, and outputs saved in the correct folders, which is required to run the scripts successfully, we recommend to download all files as a single ZIP file (click the "Access Dataset" button in the right-hand corner and select "Original Format ZIP"). If you only need selected files, make sure to download data in the original file format. Please do not hesitate contacting the authors if you have any questions.

Datasets

The repository for replicating the empirical analyses contains the following datasets. The following files (in alphabetic order) are used in the R scripts listed below. All other files in the repository, which are not listed here are intermediate products of the analysis and are there to save time if only some steps of the analysis should be reproduced.

- *dl_1st_best_de*: Best deep learning model for German tweets
- *dl_1st_best_fr*: Best deep learning model for French tweets
- *dl_2nd_best_de*: Second best deep learning model for German tweets
- *dl_2nd_best_fr*: Second best deep learning model for French tweets
- *dl_01_de*: Best deep learning model for German newspaper articles and press releases
- *dl_01_fr*: Best deep learning model for French newspaper articles and press releases
- *dl_02_de*: Second best deep learning model for German newspaper articles and press releases
- *dl_02_fr*: Second best deep learning model for French newspaper articles and press releases
- *gbm_1st_best_de*: Best gradient boosting model for German Tweets and press releases
- *gbm_01_de*: Best gradient boosting model for German newspaper articles and press releases
- *gbm_01_fr*: Best gradient boosting model for French newspaper articles and press releases
- *h2o_3.24.0.4.tar.gz*: Installation Container for H2O on R
- *main_data_for_paper_all_topics_18_19.RDS*: Data for Papers VAR analysis
- *main_data_for_paper_all_topics_18_19_rt.RDS*: Data for Appendix VAR analysis with retweets.
- *main_data_for_paper_all_topics_sep_party_18_19.RDS*: Data for Appendix VAR analysis with disaggregated parties.
- *named_entities_swiss_politics_de.RDS*: Named Entities in Swiss politics (politicians etc.)
- *named_entities_swiss_politics_fr.RDS*: Named Entities in Swiss politics (politicians etc.)

- *non_political_words_swiss_politics_de.RDS*: Keyword list
- *non_political_words_swiss_politics_fr.RDS*: Keyword list
- *onetime-structural-shock-irfs-results_all_big.RDS*: Onetime structural shocks of item response function for the VAR with disaggregated parties.
- *onetime-structural-shock-irfs-results_all_small_diff_parallel_sim.RDS*: Onetime structural shocks of item response function for the main VAR with extended simulations
- *onetime-structural-shock-irfs-results_all_small_no_pr.RDS*: Onetime structural shocks of item response function for the VAR without press releases.
- *onetime-structural-shock-irfs-results_all_small_rt.RDS*: Onetime structural shocks of item response function for the VAR with retweets.
- *onetime-structural-shock-irfs-results_all_small.RDS*: Onetime structural shocks of item response function for the main VAR of the paper.
- *political_words_swiss_politics_de.RDS*: Keyword list
- *political_words_swiss_politics_fr.RDS*: Keyword list
- *pressreleases_2018-2019.RDS*: Data with classified press releases (no text)
- *Pressreleases_to_classify.RDS*: Data of press releases
- *smd_minified_2018-2019.RDS*: Data with classified newspaper articles (no text)
- *train_test_data_2019_smd_600_de.csv*: Test & Training data newspapers German
- *train_test_data_2019_smd_600_fr.csv*: Test & Training data newspapers French
- *train_test_data_twitter_400_de.csv*: Test & Training data tweets German
- *train_test_data_twitter_200_fr.csv*: Test & Training data tweets French
- *Tweets_2018_2019_curated_sentiment_class.RDS*: Data with classified tweets (no text)
- *var_irfs_main_all_big.RDS*: VARs IRFs with disaggregated parties
- *var_irfs_main_all_small_no_pr.RDS*: VARs IRFs without press releases
- *var_irfs_main_all_small_rt.RDS*: VARs IRFs with retweets
- *var_irfs_main_all_small.RDS*: Main VARs IRFs
- *var_model_all_big.RDS*: VAR model with disaggregated parties
- *var_model_all_small_no_pr.RDS*: VAR model without press releases
- *var_model_all_small_rt.RDS*: VAR model with retweets
- *var_model_all_small.RDS*: Main VAR model
- *W2V_2019_SMD*_Model_600_DE: Word embedding for German newspaper articles and press releases
- *W2V_2019_SMD*_Model_600_FR: Word embedding for French newspaper articles and press releases
- *W2V_2019_SMD_Model_600_DE_ClassVectors.csv*: Word vector for German newspaper articles and press releases
- *W2V_2019_SMD_Model_600_FR_ClassVectors.csv*: Word vector for French newspaper articles and press releases
- *W2V_Twitter*_Model_400_DE: Word embedding for German tweets
- *W2V_Twitter*_Model_200_FR: Word embedding for French tweets
- *W2V_Twitter_Model_400_DE_ClassVectors.csv*: Word vector for German tweets
- *W2V_Twitter_Model_200_FR_ClassVectors.csv*: Word vector for French tweets

R Scripts

The repository contains the following R scripts. If you want to reproduce the entire analysis, please load and execute the scripts in the following order:

- *00-01-install_h2o_3_24_0_4.R*: script to install the h2o package with the correct version number which was used for training all models. To run this script you will need the following file:
    - *h2o_3.24.0.4.tar.gz*
- *01-01-classify_smd_data_01.R*: script to classify all articles from newspapers in German. This script cannot run since the input data set cannot be added to the replication files. We are unable to provide the newspaper articles for replication, due to a binding data protection agreement with the Swiss Media Database. To run this script (if you are in possession of the Newspaper Articles Dataset) you will need the following files:
    - *W2V_2019_SMD_Model_600_DE*
    - *dl_01_de*
    - *dl_02_de*
    - *gbm_01_de*
    - *named_entities_swiss_politics_de.RDS*
    - *political_words_swiss_politics_de.RDS*
    - *non_political_words_swiss_politics_de.RDS*
    - *SMD_TO_CLASSIFY.RDS* (Missing Dataset)
- *01-01-classify_smd_data_02.R*: script to classify all articles from newspapers in French. This script cannot run since the input data set cannot be added to the replication files. We are unable to provide the newspaper articles for replication, due to a binding data protection agreement with the Swiss Media Database. To run this script (if you are in possession of the Newspaper Articles Dataset) you will need the following files:
    - *W2V_2019_SMD_Model_600_FR*
    - *dl_01_fr*
    - *dl_02_fr*
    - *gbm_01_fr*
    - *named_entities_swiss_politics_fr.RDS*
    - *political_words_swiss_politics_fr.RDS*
    - *non_political_words_swiss_politics_fr.RDS*
    - *SMD_TO_CLASSIFY.RDS* (Missing Dataset)
- *01-01-classify_smd_data_03.R*: script to merge the classified newspaper articles in German and French. To run this script you will need to run the scripts *01-01-classifiy_smd_data_01.R* and *01-01-classifiy_smd_data_02.R* first and have the file:
    - *SMD_TO_CLASSIFY.RDS (Missing Dataset)*
- *01-02-classify_press_data_01.R*: script to classify all articles form press releases in German. To run this script you will need the following files:
    - *W2V_2019_SMD_Model_600_DE*
    - *dl_01_de*
    - *dl_02_de*
    - *gbm_01_de*
    - *named_entities_swiss_politics_de.RDS*
    - *political_words_swiss_politics_de.RDS*
    - *non_political_words_swiss_politics_de.RDS*
    - *Pressreleases_TO_CLASSIFY.RDS*
- *01-02-classify_press_data_02.R*: script to classify all articles form press releases in French. To run this script you will need the following files:
    - *W2V_2019_SMD_Model_600_FR*

    - *dl_01_fr*
    - *dl_02_fr*
    - *gbm_01_fr*
    - *named_entities_swiss_politics_fr.RDS*
    - *political_words_swiss_politics_fr.RDS*
    - *non_political_words_swiss_politics_fr.RDS*
    - *Pressreleases_TO_CLASSIFY.RDS*
- *01-02-classify_ press_data_03.R*: script to merge the classified press releases in German and French. To run this script you will need to run the scripts *01-02-classifiy_press_data_01.R* and *01-02-classifiy_press_data_02.R* first and have the file:
    - *Pressreleases_TO_CLASSIFY.RDS*
- *01-03-classify_twitter_data_01.R*: script to classify all tweets in German. This script cannot run since the input data set cannot be added to the replication files. We are unable to provide the tweets for replication, due to the Developer Agreement we have with Twitter. To run this script if you are in possession of the Tweets you will need the following files:
    - *W2V_Twitter_Model_200_DE*
    - *dl_1st_best_de*
    - *dl_2nd_best_de*
    - *gbm_1st_best_de*
    - *Tweets_2018_2019_curated_sentiment.rds (Missing Dataset)*
- *01-03-classify_twitter_data_02.R*: script to classify all tweets in French. This script cannot run since the input data set cannot be added to the replication files. We are unable to provide the tweets for replication, due to the Developer Agreement we have with Twitter. To run this script if you are in possession of the Tweets you will need the following files:
    - *W2V_Twitter_Model_400_FR*
    - *dl_1st_best_fr*
    - *dl_2nd_best_fr*
    - *Tweets_2018_2019_curated_sentiment.rds (Missing Dataset)*
- *01-03-classify_ twitter_data_03.R*: script to merge the classified tweets in German and French with each other. To run this script you will need to run the scripts *01-03-classifiy_tweets_data_01.R* and *01-03-classifiy_tweets_data_02.R* first and have the file:
    - *Tweets_2018_2019_curated_sentiment.rds (Missing Dataset)*
- *02-01-prep_data_main_paper.R*: This script produces the data for the VAR analysis. To run the script, you need the following files:
    - *pressreleases_2018-2019.RDS*
    - *smd_minified_2018-2019.RDS*
    - *Tweets_2018_2019_curated_sentiment_class.RDS*
- *03-01-generate_paper_tables.R*: This script produces all tables of the paper except for the metrics table of the classifiers used in the appendix. To run this script, you need the following files:
    - *pressreleases_2018-2019.RDS*
    - *smd_minified_2018-2019.RDS*
    - *Tweets_2018_2019_curated_sentiment_class.RDS*
- *04-01-var.R*: This script contains the vector autoregression model used in the paper. To run this script, you need the following files:
    - *main_data_for_paper_all_topics_18_19.RDS*

- *05-01-figure-2.R*: This script produces Figure 2 (Figure 1 is a flow chart that does not require any code). To run this script, you need the following files:
    - *ggplot_theme_ddl.R*
    - *pressreleases_2018-2019.RDS*
    - *smd_minified_2018-2019.RDS*
    - *Tweets_2018_2019_curated_sentiment_class.RDS*
- *05-02-figure-3.R*: This script produces Figure 3. To run this script, you need the following files:
    - *ggplot_theme_ddl.R*
    - *onetime-structural-shock-irfs-results_all_small.RDS*
- *05-03-figure_4.R*: This script produces Figure 4. To run this script, you need the following files:
    - *ggplot_theme_ddl.R*
    - *var_model_all_small.RDS*
- *0A-01-01-prep_data_appendix_retweets.R*: This script produces the data for the VAR analysis with retweets. To run the script, you need the following files:
    - *pressreleases_2018-2019.RDS*
    - *smd_minified_2018-2019.RDS*
    - *Tweets_2018_2019_curated_sentiment_class.RDS*
- *0A-01-01-prep_data_appendix_split_parties.R*: This script produces the data for the VAR analysis with the parties disaggregated in the appendix. To run the script, you need the following files:
    - *pressreleases_2018-2019.RDS*
    - *smd_minified_2018-2019.RDS*
    - *Tweets_2018_2019_curated_sentiment_class.RDS*
- *0A-02-01-var_retweet.R*: This script contains the vector autoregression model used in the appendix of the paper for the analysis with retweets included. To run this script, you need the following files:
    - *main_data_for_paper_all_topics_18_19_rt.RDS*
- *0A-02-02-var_no_press.R*: This script contains the vector autoregression model used in the appendix of the paper for the analysis without press releases. To run this script, you need the following files:
    - *main_data_for_paper_all_topics_18_19.RDS*
- *0A-02-03-var_parties_split.R*: This script contains the vector autoregression model used in the appendix of the paper for the analysis with disaggregated parties. To run this script, you need the following files:
    - *main_data_for_paper_all_topics_sep_party_18_19.RDS*
- *0A-03-02-figure_A1.R*: This script produces figure number A1 of the appendix. To run this script, you need the following files:
    - *ggplot_theme_ddl.R*
    - *onetime-structural-shock-irfs-results_all_small_rt.RDS*
- *0A-03-02-figure_A2.R*: This script produces figure number A2 of the appendix. To run this script, you need the following files:
    - *ggplot_theme_ddl.R*
    - *onetime-structural-shock-irfs-results_all_small.RDS*
- *0A-03-03-figure_A3_A4.R*: This script produces figure number A3 and A4 of the appendix. To run this script, you need the following files:
    - *ggplot_theme_ddl.R*
    - *onetime-structural-shock-irfs-results_all_small.RDS*

- *0A-03-04-figure_A5.R*: This script produces figure number A5 of the appendix. To run this script, you need the following files:
    - *ggplot_theme_ddl.R*
    - *onetime-structural-shock-irfs-results_all_small.RDS*
- *0A-03-05-figure_A6_A7.R*: This script produces figure number A5 of the appendix. To run this script, you need the following files:
    - *ggplot_theme_ddl.R*
    - *onetime-structural-shock-irfs-results_all_big.RDS*
- *0A-04-01-class_perf_text_de.R*: This script outputs the performance metrics used in the appendix for the performance metrics table A1. To run this script, you need the following files:
    - *W2V_2019_SMD_Model_600_DE*
    - *W2V_2019_SMD_Model_600_DE_ClassVectors.csv*
    - *train_test_data_2019_smd_600_de.csv*
    - *dl_01_de*
    - *dl_02_de*
    - *gbm_01_de*
- *0A-04-02-class_perf_text_fr.R*: This script outputs the performance metrics used in the appendix for the performance metrics table A1. To run this script, you need the following files:
    - *W2V_2019_SMD_Model_600_FR*
    - *W2V_2019_SMD_Model_600_FR_ClassVectors.csv*
    - *train_test_data_2019_smd_600_fr.csv*
    - *dl_01_fr*
    - *dl_02_fr*
    - *gbm_01_fr*
- *0A-04-03-class_perf_twitter_de.R*: This script outputs the performance metrics used in the appendix for the performance metrics table A2. To run this script, you need the following files:
    - *W2V_Twitter_Model_200_DE*
    - *W2V_Twitter_Model_200_DE_ClassVectors.csv*
    - *train_test_data_twitter_200_de.csv*
    - *dl_1st_best_de*
    - *dl_2nd_best_de*
    - *gbm_1st_best_de*
- *0A-04-04-class_perf_twitter_fr.R*: This script outputs the performance metrics used in the appendix for the performance metrics table A2. To run this script, you need the following files:
    - *W2V_Twitter_Model_200_fr*
    - *W2V_Twitter_Model_200_fr_ClassVectors.csv*
    - *train_test_data_twitter_200_fr.csv*
    - *dl_1st_best_fr*
    - *dl_2nd_best_fr*

The beginning of each R script lists the packages that need to be loaded and the package versions that have been used when running the scripts.
The entire analysis has been successfully executed using the following platform and package versions:

```
R version 4.0.2 (2020-06-22)
Platform: x86_64-apple-darwin17.0 (64-bit)
```

```
Running under: macOS  10.16

Matrix products: default
LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] compiler  parallel  stats     graphics  grDevices utils     datasets  methods
[9] base

other attached packages:
 [1] stargazer_5.2.2    cowplot_1.1.1      magrittr_2.0.1     scales_1.1.1
 [5] graphlayouts_0.7.1 lubridate_1.7.9.2  forcats_0.5.1      purrr_0.3.4
 [9] tibble_3.0.5       tidyverse_1.3.0    stringr_1.4.0      pbmcapply_1.5.0
[13] tm_0.7-8           NLP_0.2-1          Hmisc_4.5-0        Formula_1.2-4
[17] survival_3.2-7     h2o_3.24.0.4       caret_6.0-86       lattice_0.20-41
[21] doParallel_1.0.16  iterators_1.0.13   foreach_1.5.1      data.table_1.14.0
[25] readr_1.4.0        jsonlite_1.7.2     rjson_0.2.20       showtext_0.9-2
[29] showtextdb_3.0     sysfonts_0.8.3     forecast_8.13      ggrepel_0.9.1
[33] ggforce_0.3.3      ggplotify_0.0.5    ggraph_2.0.5       ggplot2_3.3.3
[37] rio_0.5.16         boot_1.3-27        vars_1.5-3         lmtest_0.9-38
[41] urca_1.3-0         strucchange_1.5-2  sandwich_3.0-0     zoo_1.8-8
[45] MASS_7.3-53        dplyr_1.0.5        tidyr_1.1.2

loaded via a namespace (and not attached):
 [1] readxl_1.3.1         backports_1.2.1      plyr_1.8.6           igraph_1.2.6
 [5] splines_4.0.2        digest_0.6.27        htmltools_0.5.1.1    viridis_0.5.1
 [9] checkmate_2.0.0      cluster_2.1.1        openxlsx_4.2.3       recipes_0.1.15
[13] modelr_0.1.8         gower_0.2.2          xts_0.12.1           tseries_0.10-48
[17] jpeg_0.1-8.1         colorspace_2.0-0     rvest_0.3.6          haven_2.3.1
[21] xfun_0.20            crayon_1.3.4         RCurl_1.98-1.2       glue_1.4.2
[25] polyclip_1.10-0      gtable_0.3.0         ipred_0.9-9          quantmod_0.4.18
[29] DBI_1.1.1            Rcpp_1.0.6           viridisLite_0.3.0    htmlTable_2.1.0
[33] gridGraphics_0.5-1   foreign_0.8-81       stats4_4.0.2         lava_1.6.8.1
[37] prodlim_2019.11.13   htmlwidgets_1.5.3    httr_1.4.2
RColorBrewer_1.1-2
[41] ellipsis_0.3.1       pkgconfig_2.0.3      farver_2.0.3         nnet_7.3-14
[45] dbplyr_2.1.0         tidyselect_1.1.0     rlang_0.4.10         reshape2_1.4.4
[49] munsell_0.5.0        cellranger_1.1.0     tools_4.0.2          cli_2.3.1
[53] generics_0.1.0       broom_0.7.3          fs_1.5.0
ModelMetrics_1.2.2.2
[57] knitr_1.30           tidygraph_1.2.0      zip_2.1.1            nlme_3.1-151
[61] slam_0.1-48          xml2_1.3.2           rstudioapi_0.13      curl_4.3
[65] png_0.1-7            reprex_0.3.0         tweenr_1.0.1         stringi_1.5.3
[69] Matrix_1.3-2         vctrs_0.3.6          pillar_1.4.7         lifecycle_1.0.0
[73] BiocManager_1.30.10  bitops_1.0-6         R6_2.5.0
latticeExtra_0.6-29
[77] gridExtra_2.3        codetools_0.2-18     assertthat_0.2.1     withr_2.4.0
[81] fracdiff_1.5-1       hms_1.0.0            quadprog_1.5-8       grid_4.0.2
[85] rpart_4.1-15         timeDate_3043.102    class_7.3-17         rvcheck_0.1.8
[89] TTR_0.24.2           pROC_1.17.0.1        base64enc_0.1-3
```