Replication Files for "A Male Hostility Spiral? Polarized Communication among Political Elites on Social Media"

Albert Wendsjö

September 2025

This document describes the replication data and code for the *Journal of Politics* paper "A Male Hostility Spiral? Polarized Communication among Political Elites on Social Media", by Albert Wendsjö, Hanna Bäck and Andrej Kokkonen.

Overview

The data and code are available at `https://dataverse.harvard.edu/dataverse/jop`, and for full replication the scripts should be run in the following order.

1. Retrieve and hydrate tweet IDs from *The Twitter Parliamentarian Database*.
   (a) In: `all_tweet_ids.csv` (can be accessed from `https://figshare.com/articles/dataset/The_Twitter_Parliamentarian_Database/10120685`).
   (b) Can for example be hydrated with using *hydrator*, available at: `https://github.com/DocNow/hydrator`. Requires access to the Twitter API.
   (c) Out: `all_tweets.csv`
2. Preprocess dataset and extract mentions, direct replies (will be stored in `data1.csv`) and full name references (will be stored in `data2.csv`).
   (a) In: `all_tweets.csv`
   (b) Script: `preprocess.R`
   (c) Out: `data1.csv` and `data2.csv`
3. Merge on user level and party level variables.
   (a) In: `data1.csv`, `data2.csv`, `user_data.xlsx`, `party_data.xlsx`, `parlgov.csv` (can be accessed from: `https://www.parlgov.org/data-info/`), `whogov.xlsx` (can be accessed from: `https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/YTRCQE`), and `manifestodata.csv` (can be accessed from: `https://manifesto-project.wzb.eu/datasets`).

(b) Script: `merging_variables.R`

(c) Out: `data1.csv` and `data2.csv`

4. Translate texts to english, sample texts for human annotation, train classifier and predict sentiment, stance and emotions in tweets (our trained stance classifier can also be accessed via Huggingface at: `https://huggingface.co/awendsjo/twitter_stance`). Requires

   (a) In: `data1.csv` and `data2.csv`

   (b) Script: `annotation_sample.R` and `train_and_predict.py`

   (c) Out: `annotated_sample.csv`, `data1.csv` and `data2.csv`

5. Providing descriptive statistics.

   (a) In: `data1.csv`, `country_data.csv`

   (b) Script: `descriptives.R` and `text_validation.R`

   (c) Results will be saved in `./tables_and_figures/`.

6. Full empirical analysis and fitting regression models.

   (a) In: `data1.csv`, `data2.csv`, `country_data.csv`

   (b) Script: `regressions.R`

   (c) Results will be saved in `./tables_and_figures/`.

7. Using the main regression model to initiate simulation.

   (a) In: `data1.csv`.

   (b) Script: `simulation.R`

   (c) Results will be saved in `./tables_and_figures/`.

Tweet level metadata

Due to Twitter's (now X) terms of service, we cannot share tweet's text and corresponding metadata. Therefore full replication requires step 1 and the creation of `all_tweets.csv`. We share the finalized datasets created in step 4, but with tweet level meta data removed. We also share user level data (`userinfo.xlsx`) but with metadata from twitter removed. This means tweet level data such as raw text, likes, retweets, exact dates, is removed, same goes for the number of followers for each user in user level data. We share the predicted variables, party level variables and dates aggregated to the month. Using this data, most of the main results in step 5, 6 and 7 can directly be replicated without hydrating the tweet ids.

Software

The R scripts was run using R version 4.3.2. The Python scripts was run in Colab, with A100 GPU.