Readme for the replication package of

Virality: What Makes Narratives Go Viral, and Does it Matter?

Overview

This package contains the replication materials for "**Virality: What Makes Narratives Go Viral, and Does it Matter?**" by Kai Gehring and Matteo Grigoletto. It includes all data and code needed to reproduce the tables and figures in the paper. Data are provided in standard formats, and code is available for Stata, Python, and R. Due to restrictions on the redistribution of content obtained via the Twitter API, the Python pipeline used for tweet annotation cannot be executed in full. Instead, we provide anonymized final datasets that permit full replication of the tables and figures using the accompanying Stata and R scripts. The full replication completes in approximately 2 hours.

Data Availability and Provenance Statements

Statement about Rights

- ☑ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
- ☑ I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package.

Summary of Availability

- ☐ All data **are** publicly available.
- ☑ Some data **cannot be made** publicly available.
- ☐ **No data can be made** publicly available.

Details on each Data Source

Data.Name	Data.Files	Location
Observational Data		
"All Tweets US"	df_x_pred_full_small_ANON.dta	input/data/dta/
"Auxiliary variables"	df_x_auxiliary_vars_ANON.dta	input/data/dta/
"Authors information"	df_x_profiles_ANON.dta	input/data/dta/
Experimental Data		
"Hero-Hero Experiment"	experiment_hh.xlsx	input/rawdata/experiment/
"Hero-Hero-Villain experiment"	experiment_hhv.xlsx	input/rawdata/experiment/
"Villain-Villain-Hero experiment"	experiment_vvh.xlsx	input/rawdata/experiment/
"Hero-Hero follow up"	follow_up_hh.xlsx	input/rawdata/experiment/
"Hero-Hero-Villain follow up"	follow_up_hhv.xlsx	input/rawdata/experiment/
"Villain-Villain-Hero follow up"	follow_up_vvh.xlsx	input/rawdata/experiment/
"Hero-Hero open questions"	experiment_hh_openquestions.dta	input/data/dta/
"Hero-Hero open questions follow up"	follow_up_hh_openquestions_fu.dta	input/data/dta/
"Hero-Hero-Villain open questions"	experiment_hhv_openquestions.dta	input/data/dta/
"Hero-Hero-Villain open questions follow up"	follow_up_hhv_openquestions_fu.dta	input/data/dta/
"Villain-Villain-Hero open questions"	experiment_vvh_openquestions.dta	input/data/dta/
"Villain-Villain-Hero open questions follow up"	follow_up_vvh_openquestions_fu.dta	input/data/dta/
"Demographics Hero-Hero"	demo_hh.csv	input/rawdata/experiment/
"Demographics Hero-Hero-Villain"	demo_hhv.csv	input/rawdata/experiment/
"Demographics Villain-Villain-Hero"	demo_vvh.csv	input/rawdata/experiment/
"Auxiliary variables"	df_x_auxiliary_vars_experiment.dta	input/data/dta/
Appendix Data		
"MTurk annotation data"	df_validation.csv	input/data/dta/
"Newspaper snippets"	df_newspaper_predictions.dta	input/data/dta/

Data.Name	Data.Files	Location
"TV transcripts"	`climate_segments_classified_round2.csv`	input/data/dta/

All datasets are anonymized to comply with Twitter's privacy policy: tweet text is omitted, and tweet/user IDs are replaced with randomly generated identifiers so they cannot be used to rehydrate the original content.

A more detailed list of every dataset and their source is provided below:

Observational data

a) Processed tweets data and related information. This dataset share a series of information regarding the tweets extracted using the historical Twitter APIv2 Each tweet is labeled according to the drama triangle using OpenAI API. This version is anonimized to comply with Twitter distribution policies. (`df_x_pred_full_small_ANON.dta`)

b) Dataset with sentiment analysis of tweets. Provides a series of information regarding tone, emotions, and various statistics of each tweet. This version is anonimized to comply with Twitter distribution policies. (`df_x_auxiliary_vars_ANON.dta`)

c) Dataset with profile information of twitter users in our dataset. This version is anonimized to comply with Twitter distribution policies. (`df_x_profiles_ANON.dta`)

Experimental data

a) Survey data from the pre-registered experiments. We conducted three separate studies, each with a representative sample of the U.S. population based on age, ethnicity, gender, and political affiliation. We provide the raw experiment responses. (`experiment_hh.xlsx`, `experiment_hhv.xlsx`, `experiment_vvh.xlsx`)

b) Survey data from the follow-up experiments. We provide the raw follow up experiment responses. (`follow_up_hh.xlsx`, `follow_up_hhv.xlsx`, `follow_up_vvh.xlsx`)

c) Open ended responses for the three experiments and their follow ups.

d) Demographic information regarding the experiment participants. Three datasets, one for each experiment. (`demo_hh.csv`, `demo_hhv.csv`, `demo_vvh.csv`)

Appendix data

a) Data from MTurk classification of random tweets. We hired workers from Amazon Mechanical Turk to perform human classification of tweets. For privacy reasons, we directly provide final processed dataset with MTurk workers and GPT annotations. (`df_validation.csv`)

b) Data from newspaper sources. A dataset with newspaper snippets labeled using the OpenAI pipeline. It contains the original segments enriched with both relevance and narrative role data. (`df_newspaper_predictions.dta`)

c) Data from TV transcripts. A dataset with TV transcripts snippets labeled using the OpenAI pipeline. It contains the original segments enriched with both relevance and narrative role data. (`climate_segments__classified_round2.csv`)

Computational requirements

Software Requirements

The replication code for the figures and tables of the paper was written in Stata and R. The code was last run on a blank environment in September 2025 with the following software:

- Stata (code was last run with version 18)
  - `estout` (as of September 2025)
  - `egenmore` (as of September 2025)
  - `coefplot` (as of September 2025)
  - `heatplot` (as of September 2025)
  - `palettes` (as of September 2025)
  - `colrspace` (as of September 2025)
  - `reghdfe` (as of September 2025)
  - `ppmlhdfe` (as of September 2025)
  - `grc1leg` (as of September 2025)
  - `sankey` (as of September 2025)
  - the program "`0_data_prep_final.do`" and "`1-analysis_final.do`" will install locally all dependencies to run each of the scripts.
- R 4.4.3
  Before running the scripts, open the R project in RStudio and run:
  `install.packages("renv")`
  `renv::restore()`
  In the directory `/input/R/` you find renv.lock and .Rprofile. To use these, follow the above code chunk.
- Python (we include the requirements to run the python codes, although we cannot share the data to replicate the data processing). Each python script need to be compiled into a specific environment. The requirements files are:
  - "`acn_data_manage.yml`" for the data preparation.
  - "`acn_geo_requiremnts.txt`" for handling geospatial data.
  - "`acn_data_analysis.yml`" for the annotation pipeline and for the data analysis.
  - "`acn_news.yml`" for the newspaper analysis.

Memory, Runtime, Storage Requirements

Summary Approximate time needed to reproduce the analyses on a standard (CURRENT YEAR) desktop machine:

☐ <10 minutes
☐ 10-60 minutes
☐ 1-2 hours
☑ 2-8 hours

☐ 8-24 hours
☐ 1-3 days
☐ 3-14 days
☐ > 14 days

Details The code was last run on a 4-core Intel Core i7-1365U laptop with Windows 11 (64-bit), 32 GB RAM, and 300 GB of free disk space..

Description of Programs and Code that Can Be Replicated

Stata `.do` files:

- `0_data_prep_final.do`: This script takes as input the intermediate datasets described above and generates the final datasets used to produce the main tables and figures in the paper.
- `1_analysis_final.do`: This is the main analysis script. It generates all tables and figures in the main text, except for those related to the validation analysis in Appendix B.

R scripts:

- `04_data_analysis.R`: This script compares GPT and human-coded classifications of political narratives on Twitter using the drama triangle framework. It reproduces the figures and statistics presented in Appendix B.

Description of Programs and Code that Cannot Be Replicated

R scripts:

- `01_mturk_qualification_test.R`: Processes the MTurk qualification test results to identify workers suitable for the climate change classification tasks. The data necessary to run this script are not included in the replication package.
- `02_tweets_samples.R`: Generates the random sample of tweets assigned to MTurk workers. The raw Twitter data required to run this script cannot be shared due to Twitter's terms of service.
- `03_mturk_character_role_classification.R`: Processes survey responses from MTurk workers related to climate character identification and role classification in social media texts. The input files required to run this script are not publicly available.

Python scripts

- `01_tweet_extraction.py`: Queries the Twitter Historical API v2 to extract tweets from random days and Saturdays between 2010 and 2021, based on predefined keywords. Outputs raw JSON files with tweet data.
- `02_data_prep.py`: Cleans the raw tweet JSON files, flattens metadata, removes duplicates, and generates features such as word count and mentions. Outputs fully processed daily and aggregate tweet datasets.
- `03_pgeocode_usa.py`: Geocodes user-reported tweet locations using the Nominatim API and maps them to U.S. states using shapefiles. Filters and labels tweets by geographic origin.

- `04_gpt_annotation_tweets.py`: Sends geolocated tweets to the OpenAI API in batches for classification. Assigns topic relevance and character roles using a two-stage prompt system.
- `05_predictions_prep.py`: Merges GPT predictions with tweet data and derives final binary indicators for analysis. Prepares cleaned datasets for downstream statistical analysis and export to Stata.
- `06_auxiliary_vars.py`: Computes textual features from tweets (e.g., mentions, hashtags, caps, emojis, sentiment, entities, text complexity). Outputs enriched .dta datasets for econometric modeling.
- `07_gpt_annotation_profiles.py`: Sends user profile descriptions to GPT for annotation on topics such as political leaning, religiosity, education, and parenthood. Outputs labeled CSV and .dta files.
- `08_experiment_classification.py`: Uses GPT to classify open-ended memory recall answers from experimental surveys. Labels responses with structured narrative categories. Outputs Stata .dta datasets.

- `a_clean_newspaper_articles.py`: Parses raw HTML files downloaded from Factiva and extracts metadata and full text for each article.
- `b_clean_newspaper_sentence_restriction.py`: Splits each article into 3-sentence snippets and retains only those containing climate-related keywords.
- `c_classify_newspaper_snippets.py`: Sends snippets to OpenAI's GPT-4o model to classify relevance and assign narrative roles.
- `d_prep_newspaper_predictions.py`: Merges predictions back to snippets and prepares the final dataset in Stata format for analysis.
- `I_get_clean_tv_transcripts.py`: Downloads and cleans TV news transcripts from the GDELT API for MSNBC and Fox News (2010–2020).
- `II_classify_climate_segments_full_pipeline.py`: Uses OpenAI's GPT-4o-mini to first assess relevance and then assign narrative roles for each segment.

- `functions.py` is the python code that contains all the functions necessary to run the analysis.

Instructions to Replicators

- Organize properly the folder structure. The correct structure is the one automatically downloaded with this replication package.
- Edit `0_data_prep_final.do` to adjust the default path. Run this script to generate the final datasets for the figures and tables generation.
- Edit `1_analysis_final.do` to adjust the default path. Run this code once to generate all the figures except Appendix B.
- Edit `04_data_analysis.R` to adjust the default path. Run this code to generate figures and data in Appendix B.

List of tables and programs

The provided code reproduces:

☑ All numbers provided in text in the paper
☑ All tables and figures in the paper

Figures and Tables in the Paper

Figure/Table #	Program
Figure 2	`1_analysis_final.do`
Figure 3	`1_analysis_final.do`
Figure 4	`1_analysis_final.do`
Figure 5	`1_analysis_final.do`
Figure 6	`1_analysis_final.do`
Figure 7	`1_analysis_final.do`
Figure 8	`1_analysis_final.do`
Figure 9	`1_analysis_final.do`
Figure 10	`1_analysis_final.do`
Figure 11	`1_analysis_final.do`
Figure 12	`1_analysis_final.do`
Table 2	`1_analysis_final.do`
Table 3	`1_analysis_final.do`
Table 4	`1_analysis_final.do`

Figures and Tables in the Appendix

Figure/Table #	Program
Figure B.1	`04_data_analysis.R`
Figure B.2	`04_data_analysis.R`
Figure B.3	`04_data_analysis.R`
Table C.3	`1_analysis_final.do`
Table C.4	`1_analysis_final.do`
Table C.5	`1_analysis_final.do`
Table C.6	`1_analysis_final.do`
Figure C.2	`1_analysis_final.do`
Figure D.1	`1_analysis_final.do`
Figure D.2	`1_analysis_final.do`
Figure E.1	`1_analysis_final.do`
Figure E.2	`1_analysis_final.do`
Figure E.3	`1_analysis_final.do`
Figure E.4	`1_analysis_final.do`
Figure E.5	`1_analysis_final.do`
Figure E.6	`1_analysis_final.do`
Figure E.7	`1_analysis_final.do`
Figure E.8	`1_analysis_final.do`
Figure E.9	`1_analysis_final.do`
Figure E.10	`1_analysis_final.do`
Figure E.11	`1_analysis_final.do`
Figure E.12	`1_analysis_final.do`
Figure E.13	`1_analysis_final.do`
Table E.1	`1_analysis_final.do`
Table E.2	`1_analysis_final.do`
Table E.3	`1_analysis_final.do`

Figure/Table #	Program
Figure F.1	`1_analysis_final.do`
Figure F.2	`1_analysis_final.do`
Table F.1	`1_analysis_final.do`
Table F.2	`1_analysis_final.do`
Table F.3	`1_analysis_final.do`
Table F.4	`1_analysis_final.do`
Table F.5	`1_analysis_final.do`
Table G.1	`1_analysis_final.do`
Table G.2	`1_analysis_final.do`
Table G.3	`1_analysis_final.do`
Table G.4	`1_analysis_final.do`
Table G.5	`1_analysis_final.do`
Table G.6	`1_analysis_final.do`
Table G.7	`1_analysis_final.do`
Table G.8	`1_analysis_final.do`
Table G.9	`1_analysis_final.do`
Table G.10	`1_analysis_final.do`
Table G.11	`1_analysis_final.do`
Table G.12	`1_analysis_final.do`
Table G.13	`1_analysis_final.do`
Table G.14	`1_analysis_final.do`
Table H.1	`1_analysis_final.do`
Table H.2	`1_analysis_final.do`