

README file for replication package for:

PREHEATING PROSOCIAL BEHAVIOR

Casey J. Wichman¹ and Nathan W. Chan²

Economic Journal

April 2023

Summary:

The primary data source for this project is user timelines scraped from Twitter. We have supplied the code for scraping although since scraping the data, Twitter's API has changed and tweets may have been removed, changed, or are now located too far back in a users' history to obtain via Twitter's API. As a result, re-doing the initial scraping would not produce the same data set used in this paper. Thus, we have supplied a sanitized, static data set on which our analysis is based and can be archived publicly.

Software requirements:

Scraping Twitter data, sentiment analysis of tweets, and some data manipulation is done in R (version 4.1.0), while the rest of data assembly, analysis, and creation of output is conducted in Stata (version 14). Any personal identifying information of Twitter users (e.g., names or Twitter handles) has been purged from the static data set. We have also removed all tweet text from our data set after performing sentiment analysis.

- All Stata packages and dependencies can be loaded in the “/script/_MASTER.do” file; to install all required packages, simply uncomment lines 10-20 of this .do file. The required packages are: gtools, eclplot, estout, listtab, reghdfe, ftools, unique, parmest.
- All necessary R libraries are loaded within the R scripts.
- This code was last run in April 2023 on an **8-core iMac with 64 GB of RAM running MacOS version 10.15.7.**
- Running the replication files from start to finish would take approximately 3 days.

Data availability statement:

Raw data is publicly available via Twitter. Using Twitter's v1 “full search” API (users can register for a Twitter developer account here: <https://developer.twitter.com/en/docs/twitter-api>), we first search for users who posted “#iloveWikipedia” between July 1, 2018 and December 31, 2018. Then, we scraped the full tweet history for users who tweeted “#iloveWikipedia” using Twttier's v1 “get statuses” API. Twitter has since changed its API interface.

Supplementary experimental data were obtained through a Qualtrics survey among Twitter users who had previously tweeted “#iloveWikipedia” or “I just donated”. Users were solicited via direct messages on Twitter and would be eligible to earn \$50 for themselves and/or a charity of

¹ Wichman: School of Economics, Georgia Institute of Technology; wichman@gatech.edu.

² Chan: Department of Resource Economics, University of Massachusetts Amherst; nchan@umass.edu.

choice for their participation, with 10% of completed surveys drawn for actual payment. Prior to data collection, ethical approval was obtained from the Georgia Institute of Technology Institutional Review Board (Protocol H21411). Full details of the experiment and instructions are available in this replication package and Online Appendix B for the paper. In total, \$700 was paid to collect experimental data. Based on the respondents' decisions in the charitable dictator game in the survey, respondents received \$105 in Amazon gift cards and the remaining \$595 was sent to charity

Replication files:

- In the “preheating_replication/script/r_scraping/” folder, there are a series of scraping scripts that were used to generate the primary data. These scripts cannot be run to reproduce the exact data set used in our analysis because, e.g., users may have deleted their Twitter accounts or the available tweet history may be truncated due to Twitter's API limits. Moreover, Twitter has changed its API since we originally scraped these data. Therefore, our replication begins from a static data set of this scraped output.
- The only difference in the scraped output is that Twitter usernames, user IDs, and tweet IDs have been anonymized and several unnecessary fields have been dropped.
- The static, anonymized raw data sets are placed in the restricted “4 Confidential data not for publication” folder because we cannot publicly archive tweets. We perform sentiment analysis on the text of these tweets and then sanitize the twitter data (i.e., removing any identifying tweet information) to produce a static data set and all analysis follows from this data set.
 - These .csv files include the sentiment scores of (a) #ilovewikipedia tweets, which constitute our initial scrape of Twitter posts that include the #ilovewikipedia hashtag, and (b) a set of user timelines based on the users identified in (a) who posted #ilovewikipedia within our sample period.
 - We can supply interested researchers with the original “unsanitized” tweet text that we cannot archive publicly.

Description of files:

- “/script/_MASTER.do” executes all data processing and generates all regression results, tables figures.
 - All packages needed to run the replication files are documented and loaded in this script.
 - Approximate run times for each script are provided in this script as well.
- Files in “/script/r_scraping/” scrapes Twitter to generate raw data set in R.
- Files in “/script/r_sentiment_analysis/” processes raw Twitter data and generates sentiment for tweets in R.
- The .do files in “/script/” import and process the data in Stata, generate all regression results and tables/figures for the final manuscript. The “/script/_MASTER.do” file includes notes on which results each of these files produce.

File	Description, output, run time.
<i>R files</i>	
/script/r_sentiment_analysis/createUserInfo.R	Generates “../4 Confidential data not for publication/ processed/user.csv” (twitter user information); (<2 min)
/script/r_sentiment_analysis/sentiment_analysis_JulSep.R	Performs sentiment analysis on Jul-Sep data. Generates “../4 Confidential data not for publication/ processed/tweets sentiment JulSep.csv”. (~4 hours)
/script/r_sentiment_analysis/sentiment_analysis_Oct.R	Performs sentiment analysis on Oct data. Generates “../4 Confidential data not for publication/ processed/tweets sentiment Oct.csv”. (~5 hours)
/script/r_sentiment_analysis/sentiment_analysis_Nov.R	Performs sentiment analysis on Nov data. Generates “../4 Confidential data not for publication/ processed/tweets sentiment Nov.csv”. (~10 hours)
/script/r_sentiment_analysis/sentiment_analysis_Dec.R	Performs sentiment analysis on Dec data. Generates “../4 Confidential data not for publication/ processed/tweets sentiment Dec.csv”. (~14 hours)
/script/r_sentiment_analysis/sentiment_combine.R	Combines/processes results from sentiment analysis above. Generates “../4 Confidential data not for publication/ processed/tweets sentiment.csv”. (<2 min)
<i>Stata files</i>	
/script/sanitizeTwitterData.do	Removes tweet text, creates Table A.1, and generates two files: “rawdata_anon/tweets_sentiment_sanitized.csv” “rawdata_anon/tweets_sentiment_justwikipedia_anon_sanitized.csv” (<5 min)
<i>Publicly available data begins here.</i>	
/script/makeRegData.do	generates “data/regData.dta” for the following files (<1 min)
/script/makeRegData_justwikipedia.do	generates “data/regData_justwikipedia.dta” for sample comparison (<1 min)
/script/makeSummaryStats.do	Creates Tables 1, A.1, A.2, & A.3 (<5 min)
/script/makeRegressions.do	Estimates all primary regressions (~20 min)
/script/makeRegTables.do	Creates Tables A.4, A.5, & A.6 (<1 min)
/script/makeRegFigures.do	Creates Figures 1, A.4, & A.5 (<5 min)
/script/makeNoTweets.do	Creates Figure A.2 (<1 min)
/script/makePolyRegs.do	Creates Figure A.3 (<5 min)
/script/makeFalsificationTest1.do	Creates Figure A.8 (<5 min)
/script/makeFalsificationTest2.do	Creates Figure A.9 (<5 min)
/script/makeRandomizationInference.do	Creates Figures 2, A.6, & A.7 (~36 hours)
/script/makeOnlineExperiment.do	Creates Figure 3 & Tables A.1-A.4 (<2 min)

Dataset list:

- All raw data sets are provided in “/rawdata_anon/” – these data sets have been anonymized to remove identifying information for Twitter users.
 - **tweets_sentiment_sanitized.csv** – a clean data set of combined twitter user’s timelines after performing sentiment analysis on the tweets and removing sensitive twitter information (i.e., tweet texts) that we cannot post publicly. This file is produced from the raw data by running the scripts in “/script/r_sentiment_analysis/” on the confidential data and then sanitizing the data using “/script/sanitizeTwitterData.do”.
 - **tweets_sentiment_justwikipedia_anon_sanitized.csv** – a file that contains data on the timelines of (anonymized) Twitter users who posted “Wikipedia” on randomly selected days within our sample, for comparison.
 - **onlineexperiment.csv** – anonymized Qualtrics survey output for online experiment.
 - **onlineexperiment.dta** – same as above, but. dta.

Description of output:

- All processed data is saved in “/data/”
- Regression results are saved in “/estimates/” and results from the randomization inference bootstrapping process are saved in “/estimates/BOOTSTRAP/”
- Intermediate graphs in Stata format (.gph) are saved in “/figures/”
- Final results for the manuscript and appendix are saved in “/FINAL_RESULTS/”. Tables are saved as .tex files and figures are saved as .pdf files.

Instructions for replicators:

1. Open the “/script/_MASTER.do” file and change the directory.
2. Run the contents of “/script/_MASTER.do” in the order listed.
 - a. Each of the files within “/script/r_sentiment_analysis/” require changing the directory to your local filepath.
 - b. While the “/script/r_sentiment_analysis/” files can be run via shell commands within the _MASTER.do file, it is recommended to run these individually.