nature portfolio

Corresponding author(s): DBPR NCOMMS-22-44025-T

Last updated by author(s): Nov 21, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ ☒ A description of all covariates tested

☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Our study utilized an open access dataset built by researchers from George Washington University who collected ~ 40 million tweets related to climate change. The dataset contains tweets posted in a two-year period, from Sept 2017 to May 2019. We used Hydrator, a desktop application, to retrieve raw twitter data based on tweet IDs.
Data analysis	Our tweets classifier was built upon OpenAI GPT-2, a large transformer-based language model. Our model was built upon the Huggingface Transformers library and implemented in PyTorch. Co-retweeted network analysis, correlation analysis, and time series analysis were performed in R software and visualized in Gephi. We employed the Latent Dirichlet Allocation (LDA) algorithm to automatically extract the main topics. The model was implemented in Python's gensim package along with the Java-based package Mallet to accelerate data processing. Code used to replicate the results is available at: https://github.com/jianxuny/Climate-Denial-on-Twitter

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The primary data used in this analysis are publicly available at: https://tweetsets.library.gwu.edu/datasets

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender

Our sample consists of ~7 million tweets posted by 1.3 million Twitter users across the U.S. Our secondary analysis does not distinguish based on gender or sex of Twitter users.

Population characteristics

Our sample consists of 1.3 million Twitter users residing in the USA based on their location as reported in their user profile. Fig. S1 depicts the spatial allocation of our sample. In terms of spatial variation, most tweets are from users residing in urban areas and along the coasts. By calculating tweets volume per county, we find over 50% of counties have more than 100 tweets and over 75% of counties have at least 30 tweets. To account for spatial variations, we use tweets count per county as a weight for our calculations. To test the representativeness of our dataset, we calculated the total number of users and tweets at the county and state levels and its correlation with the population of the respective jurisdictional level. As shown in Fig. S2a, the total number of tweets highly correlates with the population at the state level (R = 0.89, p-value < 0.001).

Recruitment

The primary data used in this analysis are publicly available and include tweets posted online. Our analysis did not involve interaction with our sample.

Ethics oversight

The study was approved by the University of Michigan board of ethics (IRB).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences □ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

This study uses Twitter data from 1.3 million users residing in the U.S to (i) estimate the prevalence of climate change denialism at the state and county level, (ii) identify characteristics of climate change deniers, (iii) understand how social media promulgates climate change denialism including the key influencers, and (iv) determine how world events are leveraged to promulgate climate change attitudes

To answer these questions, we used a Deep Learning text recognition model to classify 7.4 million geocoded tweets, collected between September 2017 to May 2019, containing keywords related to climate change. We classified these tweets about climate change into ‘for’ (belief) and ‘against’ (denial). Our analysis resulted in a profile of climate change deniers at the county level, insight into the networks of social media figures influential in promoting climate change denial, and knowledge of how these influencers use current events to foster this denial.

Research sample

Our sample consists of ~7 million tweets posted by 1.3 million Twitter users across the U.S.

Sampling strategy

The initial ~40 million climate change related tweets database include tweets from users around the world. We included only tweets from users residing in the U.S in our analysis. To extract tweets located in the U.S., we developed a rule based on the geo-attributes in the raw data. We extracted the self-reported location information in an account profile. A large proportion of users (> 73%) provided the location information in our dataset. To standardize the addresses and improve the geocoding process, we first transformed all the user locations to lower case and removed the URL links, emojis, punctuation marks, and other non-ASCII characters. Next, we extracted all the unique user locations (~ 640,000 “clean” addresses) and standardized all the U.S. state and city abbreviations. As a final step, we manually inspected and removed national level and obviously fake user locations. To reduce the incidence of non-human accounts in our sample, we removed users who tweeted more than 20 times a day.

Data collection

Our study utilized an open access dataset built by researchers from George Washington University who collected ~ 40 million climate

Data collection

change related tweets, via the Twitter Stream API.1 The dataset contains tweets posted in a two-year period, from Sept 2017 to May 2019. We used Hydrator, a desktop application to retrieve raw twitter data based on tweet IDs. In Nov 2020, we successfully retrieved ~27.3 million raw tweets. The following keywords were set as the filters, which contain popular hashtags from both climate change believers and deniers.
#climatechange, #climatechangeisreal, #actonclimate, #globalwarming, #climatechangehoax, #climatedeniers, #climatechangeisfalse, #globalwarminghoax, #climatechangenotreal, climate change, global warming, climate hoax

We then extracted the necessary attributes both for each tweet and user who posted the tweet. Tweet attributes include the full text, author ID, time of creation, geo location, tweet type (original tweet, retweet, quote, reply), and cumulative number of retweets and likes. User attributes include the username, self-defined user location, and number of followers.

Timing and spatial scale

Our sample covers a two-year period from September 2017 to May 2019 and includes Twitter users residing in the contiguous U.S

Data exclusions

We excluded tweets posted from users residing outside of the U.S. We also excluded users with no location in their profile or with national level information or obviously fake addresses. To reduce the incidence of non-human accounts in our sample, we removed users who tweeted more than 20 times a day. Finally, we excluded model predictions with low confidence (CI < 0.75) from any subsequent analyses.

Reproducibility

Our analysis is based on publicly available data (Twitter, census, CDC, Presidential election results, FEMA, Prism, EIA) and our code is available at https://github.com/jianxuny/Climate-Denial-on-Twitter. The models we use are open source based on pre-trained libraries and the analysis was conducted solely on open source software (Python, R, QGIS).

Randomization

N/A

Blinding

N/A

Did the study involve field work? □ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging