


Datasheet for 2020 Cooperative Election Study*

The survey dataset used for ‘Fake News vs Fox News’




April 18, 2024

This datasheet is the extract of the questions from Gebru et al. (2021). And it was put together with the help of 

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
 - The dataset was created to study how Americans view Congress during elections, their voting behavior and experiences, and how these vary across different political and social contexts. The goal was to measure the distribution of voters’ preferences across states (Schaffner, Ansolabehere, and Luks 2021). For this study, this dataset satisfies all needs, and there was no specific gap that needed to be filled.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
 - This dataset is created by 60 research teams and organizations participating in the Cooperative Election Study (CES) initiative.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
 - The creation of the dataset was funded by various sources, including the National Science Foundation which provided support for all even-year surveys from 2010 onward (Schaffner, Ansolabehere, and Luks 2021).
4. *Any other comments?*
 - No further comments.

Composition

*Code and data are available at: 

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
 - Each instance is 1 individual in the US who responded the CES2020 survey.
2. *How many instances are there in total (of each type, if appropriate)?*
 - There are 61,000 instances in total.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
 - The dataset is a sample of all American adults, drawn using a sample matching methodology. The larger set would be the entire population of American adults. The sample is representative of the larger set in terms of demographic characteristics, as the matching process aimed to ensure similarity between the sample and the target population (Schaffner, Ansolabehere, and Luks 2021).
4. *What data does each instance consist of? “Raw” data (for example, unprocessed text or images) or features? In either case, please provide a description.*
 - Each instance consists of their demographic, income, education, political views, voting behaviours, and other relevant features.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
 - No label or target association found.
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
 - No, there is no missing information related to my research.
7. *Are relationships between individual instances made explicit (for example, users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
 - No explicit relationships available.
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

- No recommended data splits available.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
- There are no errors, sources of noise, or redundancies in the dataset.
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
- The dataset is self-contained, will exist and remain constant over time.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
- No
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
- No
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
- Yes, this dataset identify sub-population traits including but not limited to age, gender, race, religion, political opinions, past/current voting choices, income, etc.
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
- No, the questions are generally not specific to the individual identification level.
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
- There are extensive questions over survey respondents' race, religion, political opinions, past/current voting choices, income, etc.

16. *Any other comments?*

- No further comments.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

- All data were acquired through survey.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

- Survey respondents respond to these questions through means including smartphones, laptops, and tablets,

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

- The dataset is a sample drawn using YouGov’s matched random sample methodology. The sampling strategy involves selecting a random sample from the target population and then matching each member of the target sample to one or more matching members from a pool of opt-in respondents using proximity matching (Schaffner, Ansolabehere, and Luks 2021).

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

- The individuals involved in the data collection process include American adults who participated in the survey. Compensation details, such as whether respondents were compensated and the amount, are not provided in the information provided.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

- Data was collected twice. The pre-election questionnaires were distributed from September 29 to November 2, 2020; the post-election questionnaires from November 8 to December 14, 2020. This timeframe match the creation timeframe of the data associated with the instances.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
 - No
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
 - The data was collected from Harvard Dataverse <https://doi.org/10.7910/DVN/E9N6PH>. I downloaded through library `dataverse` (Kuriwaki, Beasley, and Leeper 2023) from R Programming Language (R Core Team 2023).
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
 - Yes, survey respondents answered the surveys willingly.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
 - Yes, survey respondents answered the surveys willingly knowing that their data will be used.
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
 - No
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
 - No
12. *Any other comments?*
 - No further comments.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

- I only selected the columns I need and removed all missing observations from the columns. The columns are “votereg,” “CC20_410,” “cc20_300a,” “cc20_300c,” “cc20_300b_1,” “cc20_300b_2,” “cc20_300b_3,” “cc20_300b_4,” “cc20_300b_5,” “cc20_300b_6,” “cc20_300b_7,” “cc20_300b_8,” and “CC20_433a.”
2. *Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.*
 - The “raw” data was saved; it can be access through <https://doi.org/10.7910/DVN/E9N6PH>
 3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
 - The software used is R Programming Language R Core Team (2023).
 4. *Any other comments?*
 - No further comments.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
 - Yes, this dataset was made public by the researchers; therefore it has been widely used for many purposes.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
 - No
3. *What (other) tasks could the dataset be used for?*
 - More election prediction could be performed using the dataset.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
 - As this dataset was collected in 2020, using it to predict future elections might not be appropriate.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

- There are no tasks for which the dataset should not be used.
6. *Any other comments?*
- No further comments.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
 - The dataset is available via the URL link from Harvard Dataverse. Any company, institution, or organization could access it through the link.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
 - It is available via the URL link <https://doi.org/10.7910/DVN/E9N6PH>.
3. *When will the dataset be distributed?*
 - Data release 1 happened in March 26, 2021.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
 - No
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
 - No
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
 - No
7. *Any other comments?*
 - No further comments.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

- Harvard Dataverse is responsible for supporting the dataset.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
 - There is a contact us button in the Harvard Dataverse website where user could message Dataverse.
 3. *Is there an erratum? If so, please provide a link or other access point.*
 - No
 4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
 - No, the final version has been released on the website, and it would not be updated.
 5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
 - No
 6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
 - Older versions of the dataset is still available via Harvard Dataset.
 7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
 - No, this dataset was put together by 60 research teams each with their own Principal Investigators. It is designed to sample American adults during the 2020 election. Therefore it can not be extended.
 8. *Any other comments?*
 - No further comments.

References

- Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. “Datasheets for Datasets.” *Communications of the ACM* 64 (12): 86–92.
- Kuriwaki, Shiro, Will Beasley, and Thomas J. Leeper. 2023. *Dataverse: R Client for Dataverse 4+ Repositories*.
- R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Schaffner, Brian, Stephen Ansolabehere, and Sam Luks. 2021. “Cooperative Election Study Common Content, 2020.” Harvard Dataverse. <https://doi.org/10.7910/DVN/E9N6PH>.