Datasheet for 'A dataset'*

Dong Jun Yoon Yuean Wang Yang Zhou

March 16, 2024

The CES 2022 dataset, collected through an online survey by YouGov, encompasses 60,000 instances representing U.S. adult citizens' electoral behaviors, political opinions, and demographic characteristics. The instances are detailed and categorized by election experiences, contextual data, demographic information (e.g., gender, race, education), political affiliations, and opinions on various policy issues. The dataset employs a matched random sampling methodology, with its representativeness validated through a two-stage weighting process involving entropy balancing and iterative proportional fitting. Preprocessing included sample matching, weighting, and vote validation, ensuring data quality and representativenes in order to avoid representative bias.

Extract of the questions from Gebru et al. (2021).

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 Cooperative Election Study (CES) aims to study how Americans perceive and hold their representatives accountable during elections. Besides, the study also interested in understanding their voting behavior and how their behavior be affected by political geography and social context. They study the group by capturing mass data across most of legislative constituencies. CES aims to fill the gap of understanding the American electoral process by analyzing congressional elections and national surveys.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

    - The Cooperative Election Study. It form 62 research teams and organizations.

*Code and data are available at: https://github.com/yangzhoucoco/Political-support-in-the-United-States.git.

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.*

    - CES is supported by the National Science Foundation (NSF). Award #2148907 is the associated grant it has. The grantors involve numerous teams and collaborations.

4. *Any other comments?*

    - Based on real data, the guide provide a deep research of understanding American electoral behavior and different impacts of social contexts on voting aspect. It might provide implication for analyzing the trend of American elections.

Composition

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.*

    - The instances in CES 2022 dataset represent individual respondents attend in the survey, which designed to record various information related to electoral behaviour, political perspective and demographic characteristics. The The types of instances include election experience, contextual data, and demographic information.

2. *How many instances are there in total (of each type, if appropriate)?*

    - CES involves 60,000 cases. Those cases are based on the survey of adult pariticipants in 2022 fall.

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).*

    - This dataset does not contain all possible instances. By contrast, it is more focus on drawn the sample from a larger set of U.S. adult citizens.
      The sampling method used by YouGov in the CES involved a matched random sampling method. This method begins with a count of the target population, which for general population studies includes all adults and is usually counted through a comprehensive survey such as the American Community Survey. CES then draws a random sample from that target population, called the target sample. The representativeness was validated by weighting procedures. This procedures aims to address any remaining imbalances among the matched sample. The sample is weighted to match the distribution of the frame across multiple demographic and political

dimensions. The weighting process consists of two stages: entropy balancing, followed by iterative proportional fitting (tilting) of several key variables and their interactions.

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.*

   - It consists demographic information, political opinions and behaviors, electoral experiences, voter validation an device and participation information. In terms of political opinions, it includes respondents' political affiliations, voting president and their positions on various policy issues and political figures.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

   - No.

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.*

   - Vote file miss matching might be one of missing information from individual instances. The CES relies on matching respondent to outside voter files. There might be some chance that name, address, or other details are mismatch with respondent, leading to mismatch between voting status and respondent.

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.*

   - CES 2022 dataset does not buil up a explicitly model relationship between individual instances. Each instance in the CES dataset represents a single respondent's survey answers, capturing their demographic details, political views, voting behavior, and other relevant information.

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.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

   - Random measurement error might be one of potential errors. For example, some respondents provides wrong answer, which might lead to a number of respondents being incorrectly categorized.

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.*

    - No, it does not link to external source. This dataset is mainly rely on survey collection of individual participants.Voter file matching might be an external consideration. Data for the CES 2022 study, including vote validation for all respondents, is archived and available at the Harvard University Dataverse. The restriction might be it does not include a specific restriction related to the TargetSmart database or other external resources used for voter file matching and verification are not detailed.

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.*

    - The CES 2022 dataset consists primarily of structured survey data collected from respondents about their political views, behaviors, and demographic information. Such data is generally not subject to legal privileges such as doctor-patient confidentiality or confidentiality of legal communications.

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.*

    - In terms of Racial and Sexual Agreement, it might be offensive for the respondent, especially when the question design in a way that might present bias.

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.*

    - There are lots of sub-population in the dataset. 1)Gender: The dataset categorizes respondents into "Man", "Woman", "Non-binary", and "Other", allowing for gender-based analyses. 2)Race: The dataset categorizes respondents into "White", "Black", "Hispanic", "Asian", "Native American", "Middle Eastern", "Two or more races", and "Other" 3)Education: Respondents' highest education levels are detailed, ranging from no high school diploma ("No HS") to postgraduate degrees 4)Political Affiliation: The dataset records respondents' party affiliation, including "Democrat", "Republican", and "Independent/Other". 5)Age: The dataset provide age range (e.g., 18-29, 30-44, 45-64, 65 and over).

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, all the dataset is anomymize.

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.*

    - Yes, it includes race, sexual orientation, religious, political opinion, transgender status, health data and so on. For example, respondents are asked to answer their sexual orientation, with options including heterosexual/straight, lesbian/gay, bisexual, and others.

16. *Any other comments?*

    - No

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.*

    - The data associated with each instance in the CES 2022 was primarily acquired through survey responses reported directly by the subjects. The survey was conducted online by YouGov and involved a methodology known as sample matching.The target population is first counted and then a representative sample is selected from a non-randomly selected group of respondents.

- Regarding the data segments concerning voter validation, the process involved aligning individual entries with the TargetSmart database, which contains records of U.S. registered voters. This alignment aimed to confirm the accuracy of survey respondents' claims about their voter registration and their actual voting activity in both the primary and general elections of 2022.

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?*

- The CES 2022 data was collected through an online survey conducted by YouGov. They employed "sample matching" methodology, which fit with online access panels.This method identify available respondents who are very similar in terms of measured characteristics to each member of the target sample. The total distance is calculated as the weighted sum of these individual distances on all attributes, and the weights can be adjusted for variables that are important to the study. This approach allows matched samples to closely represent the characteristics of the target population.

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 CES 2022 dataset's sampling strategy employed YouGov's matched random sample methodology.

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 CES involved 60 parties and research teams. The dataset does not provide specific details on how individuals be compensated.

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.*

    - Collected over two time frame: the first period is the pre-election data collection occurred from September 29 to November 8, 2022,and the second period is post-election data was gathered from November 10 to December 15, 2022.The timeframe is relevant to survey participants about their opinions, voting intention in 2022 election.

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.*

    - The dataset does not provide explicit details regarding any ethical review processes.

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 for the 2022 Cooperative Election Study (CES) was collected directly from individuals through a survey conducted over the internet by YouGov.

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, because it is conducteddirectly from individuals through a survey conducted by YouGov over the internet. Cooperative Election Study (CES) or provide the exact language used in the notification.

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.*

    - The individuals participating in the 2022 Cooperative Election Study (CES) did provide their consent for data collection.The respondent will be asked if they agree to attend in the survey by inquired "Consent to participate: Do you agree to participate in the study?" The exact language used to request consent, is straightforward.

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).*

    - The dataset does not provide specific information about whether individuals were given a mechanism to revoke their consent after agreeing to participate in the 2022 Cooperative Election Study (CES) or for certain uses of their data.

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.*

    - The provided CES Guide 2022 document does not specifically describe conducting a data protection impact analysis. It emphasizes on on the methodology, sample matching, data collection,

12. *Any other comments?*

    - No

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.*

    - Yes, CES involves some data preprocessing, cleaning, and labeling to make sure the quality of the data. The descriptions have been included in the sampling methodology, matching procedures, and weighting processes. CES done 3 things, which are sampling and sample matching, weighting, and vote validation.To be more specific, in order to correct for any residual imbalances between the matched sample

and the target population, the CES data was weight through 2 steps. First, using entropy balancing to match the distributions of the frame on key demographic variables (age, gender, education, race, etc.), and then using iterative proportional fitting ("raking") on additional variables (like voter registration status and 2020 Presidential vote choice) to further refine the representativeness of the sample.

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 CES Guide 2022 does not explicitly mention whether the "raw" data save in addition to the preprocessed/cleaned/labeled data.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - No.

4. *Any other comments?*

    - No.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

    - The CES Guide 2022 does not provide specific examples of tasks for which the dataset has already been used.

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.*

    - The CES Guide 2022 does not explicitly mention a repository that links to paper.

3. *What (other) tasks could the dataset be used for?*

    - There are multiple task that the dataset can be used for. For example, to analyze voting behavior and study the impact of socioeconomic factor on political outcome.

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?*

- Yes.The methodology section details efforts to create a representative sample and address potential biases through sample matching, weighting, and vote validation in order to avoid representation bias.The customer can ensure all research complies with applicable laws, ethical standards, and best practices for data privacy and protection.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

   - No.

6. *Any other comments?*

   - No.

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.*

   - Yes, it will. It indicated that the data for the Cooperative Election Study (CES) 2022 is available at the Harvard University Dataverse. This can be seem as the dataset is accessible to third parties outside of the institutions directly involved in the creation of the dataset, such as Harvard University and YouGov.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

   - The Cooperative Election Study (CES) 2022 will be distributed through the Harvard University Dataverse. This distribution method ensures that the dataset is accessible for academic and research purposes. The DOI for the Cooperative Election Study Common Content 2022 is https://doi.org/10.7910/DVN/PR4L8P,

3. *When will the dataset be distributed?*

   - Dataset is distributed in two period: first, on March 20, 2023. It included the CES 2022 Common Content, which is the data collected that is common across all participating research teams.The second period is on September 8, 2023, it included vote validation for all respondents, enhancing the dataset's utility for electoral research by verifying the voting behavior reported by the survey participants.

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.*

    - TBD

7. *Any other comments?*

    - No.

Maintenance

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

    - The dataset for the CES 2022 is maintained and hosted by the Harvard University Dataverse.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - People can directly contact The Harvard University Dataverse platform.

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)?*

    - The guide does not provide detail information about update.

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.*

    - The CES Guide 2022 does not provide specific information regarding limits on the retention of data associated with the instances.

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.*

- The guide does not discuss about the policies regarding the support, hosting, or maintenance of older versions of the 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.*

    - The guide does not include a explicit mechanism for others to extend to the dataset.

8. *Any other comments?*

    - No

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.