Datasheet for 'A dataset'*

My subtitle if needed

First author Another author

Invalid Date

First sentence. Second sentence. Third sentence. Fourth sentence.

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 made the dataset was created to analyze the number of U.S. adult citizens who vote in U.S. elections and whether their background affects the probability of voting. Select each state as a sample, and then select matches from the middle for further investigation. In this way, the data obtained by combining multiple variables will be more accurate.
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 build this dataset and it was made of 60 groups.
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 2022 Cooperative Election Study was supported by the National Science Foundation Award # 2148907.
4. *Any other comments?*
    - The data given by CES is general, and new data needs to be combined to form a more accurate result.

Composition

*Code and data are available at: LINK.

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 CES 2022 data represents the electoral behavior of voters when they cast their ballots for candidates, with examples showing that voters' choice lines differ by gender and education.

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

    - CES has 60,000 cases. It divided the samples of these cases themselves into different variables, which he then screened and cleaned.

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 does not include all instances, the data is too large, so the CES calculation uses each state to collect data. Voter status is also restricted, for example, to choosing respondents only among adult U.S. citizens. Each state's calculations are better at collecting local data and matching it. A lot of voter data may be lost, the data will not be matched, so the instance will not be able to match other samples.

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.*

    - Instance included voter information, political party, election status, equipment participating in the election. These are all variables derived from CES raw data.

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

    - Yes. For example, in the CES gender's label, his label is gender4.

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.*

    - CES only records the data and does not show the corresponding trend, it is easy to analyze a single variable and cannot be combined togethe

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 does not give relationships between individual instances, they are more like separate categories and then show data, for example, gender4 in only gender differentiation without adding people and other variables.

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.*

    - The data recorded by the respondents was lost and could not be matched to the data in the next part of the sample.

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.*

    - CES data was collected from individuals on the web or through YouGov's 1000 data survey. The data is kept at the Harvard Dataverse, where it is accessible to all.

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.*

    - CES did not have data that could be classified as confidential because it did not involve excessive privacy exposure, and he asked respondents more about their status at the time of the election.

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.*

    - Data sets about gender issues and sexual orientation may offend respondents.

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 a lot of sub-populations in the data. For example, based on gender, there will not be a single gender, but a variety of genders for respondents to choose from. Secondly, there are 5 education levels. In terms of age, most of the respondents were adults, so the lowest age was from 18 and then all the way up to 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, he did not specify a particular person, the data is more to gather each person's data into different categories.

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.*

    - Dataset Indicates sensitive data. The Dataset covers race, ethnic origin, religious beliefs, political views, party affiliation, sexual orientation, health data, and crime.

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.*

    - They collect information from YouGov, which collects a random sample of 1,000 people, and from individual respondents, who collect information using reported data, for example, on the web. After an initial screening, the missing data will not be matched. Therefore, collection is carried out after matching corresponding samples to improve the accuracy of data.

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?*

    - CES data is collected online through the web. Sample matching is used to balance the whole data.

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 dataset uses the web to collect large amounts of data and to enter secondary data updates with smaller samples.

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)?*

    - CES has more than 60 teams conducting research with around 60,000 respondents, but does not show how many are paid.

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.*

    - The first questionnaire was in the field from September 29 to November 8; the post-election wave was from November 10 to December 15.

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,there is no ethical review processes in the text.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - CES data was collected from the web, and the website collected data from the respondents.

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, the individuals was indicated on the website that the data was collected.

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, the Respondent personally knows that their data will be collected and used, because the respondent's voter registration status mentions whether the respondent is active or not. And in each of the following options there is an absence or other state.

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).*

- There is no claim of a mechanism for withdrawing consent in the future or for certain purposes.

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.*

    - CES only provides data on a single variable and no more data analysis.

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.*

    - Four will be selected first for distribution, followed by voteberg, educ, gender4 and the voters. And then it was cleaned it up a little bit, filter voteberg out of 1. We say votereg == 1, presvote20post %in% c(1, 2)) votereg equals 1 which means the respondent can vote. Filter out the ones that aren't votereg and aren't 1 or only look at people who register to vote. presvote20post %in% c(1, 2) Means 2020 President Vote Post Election Who did you vote for in the election for President in 2020 presvote20post. However, it only focus on Biden or Trump, so we are second. It only select these four columns presvote20post votebreg, educ and gender4 in the raw dataset for variable study.

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.*

    - CES Guide 2022 does not have any other raw data except for the data recorded on voter turnout

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?*

    - I think CES needs to create more data and put it out there.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - CES does not give any tasks, it just provides data reference.
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.*
    - CES does not give a repository of any papers or systems for the dataset, which is mostly obtained directly through the Web.
3. *What (other) tasks could the dataset be used for?*
    - Data sets can be used to analyze the influence of gender, age, race and so on on politics.
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?*
    - The composition of the data or other manipulation can affect the unfair treatment of individuals or groups, for example, if a group mostly supports Biden, it is likely that they will suffer discrimination from fans of Trump or other candidates. So we increase the balance by weighting, which allows for more accurate data with less bias.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - No, data sets can do a lot of things.
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.*
    - CES data sets are sent to Harvard University's Dataverse, and the corresponding information is available.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

- The dataset is available for study and study on the website in the form of doi:10.7910/DVN/PR4L8P.

3. *When will the dataset be distributed?*

   - The author has released four Dataset versions. The Dataset version first existed on 2023-03-20, this one was followed by 2023-03-21 and 4-08, the authors added citation Metadata and Files2. The final version was voted validation by TsrgetSmart to increase the authenticity and accuracy of the data.

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.*

   - TBD

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.*

   - TBD

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?*

   - TBD

Maintenance

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

   - Zheng(Zeb) Yang

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

   - zhengzeb.yang@mail.utoronto.ca

3. *Is there an erratum? If so, please provide a link or other access point.*

   - Not yet.

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)?*

- Yes, we will update the dataset in a timely manner according to the requirements listed in the replies received in the mailbox from the dataset consumers.

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.*

   - TBD

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.*

   - TBD

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.*

   - TBD

8. *Any other comments?*

   - TBD

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.