Datasheet for 'CES 2020 Election data'*

Thu Dong

CES 2020 Election data is provided by Harvard Dataverse. The content of the data are the informations of American representative who participated in the election 2020. In general, the data set includes variables regarding voters backgrounds, economics condition, party affiliation, and related factors.

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 data set was created to study how Americans voted their electoral experiences, and how their behavior and experiences vary with political geography and social context. The data was constructed to provide more demographic-related information for the 2020 presidential election.

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

    - The 2020 CES involved 60 teams, yielding a Common Content sample of 61,000 cases. Each research team purchased a 1,000-person national sample survey, conducted by YouGov of Redwood City, CA.

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.*

    - Cooperative Election Study (CES),a national online survey conducted before and after USA presidential election and mid-term, is funded by educational institutions including Harvard University.

Composition

*Code and data are available at:https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/E9N6PH.

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 2020 election data is comprised of one type of instance, specifically, American representatives of the population. Each instance includes information about demographics, social economics, political beliefs, and their choice of candidate in the 2020 election.

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 data set contains a sample of instances of all Americans who voted in the 2020 presidential election. The sample is representative of the population because the samples are drawn evenly from all 50 states.

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 demographic details(gender, race, multiracial, etc), socioeconomic status (household income, stock ownership, etc), and political preferences(party affiliation, choice of candidate, etc). Each instance consists of discrete numerical variables that represent categorical variables which is specified in the Data Guide

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

    - There were no targets or labels associated with any instances.

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.* -There is some information missing from the individual instances. The missing information is either because of the failure to register for voting or unprovided by survey participants.

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.*

    - The instances are independent of each other

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.*

    - Since the data was large, the model was run using only a sample of the data set

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

    - There aren't many 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.*

    - Since the website provides data for ongoing studies, there are links to other Harvard websites to provide answers to frequently asked questions. This website is not restricted to any users.

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 data provided by CES is publicly available

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.*

    - The dataset does not contain data that if viewed might be offensive, insulting, threatening, or might otherwise cause anxiety

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.*

- Since the data set provides demographic information, there are groups of subpopulations such as gender, race, etc. These subpopulations are identified through provided questions that are included in the survey.

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.*

    - It is not possible to identify individuals from the data set as the answers to all survey questions are categorical.

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.*

    - The data set provides information regarding race, political beliefs, religious beliefs, household income, etc which could be considered sensitive information.

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 is reported by subjects and it is unclear whether or not the information is verified. However, looking at the outcome of the 2020 election, the information can be somewhat verified as the data set is a sample of a whole population.

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?*

    - Each of the 60 teams purchased 1,000-person surveys. All cases were selected through the Internet and YouGov constructed matched random samples for this study.

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 sampling method uses YouGov's matched random sample methodology. Sample matching is a methodology for the selection of "representative" samples from non-randomly selected pools of respondents.

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 data collection process was led by researchers in 39 universities in the US.

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 pre-election questionnaire was conducted from September 29 to November 2; the post-election survey was in the field from November 8 to December 14.

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.*

    - There was no information provided regarding the ethical review process conducted

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 3rd party sources, specifically 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.*

    - Individuals know about the data collection as the survey explicitly asks whether they want to participate in the study before any other questions.

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.*

    - Individuals consent to the collection of data by consenting to their participation in the study. The survey explicitly asks whether they want to participate in the study before any other questions.

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.*

- There was no preprocessing/cleaning/labeling of the data.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
   - A similar version of the data set but in a different year was 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.*
   - https://www.tandfonline.com/doi/abs/10.1080/03623319.2020.1809901?casa_token=Efu-21iYgfoAAAAA:W4T6qZ5B3XY8il3y5CNp12MpdWiBchQsfs9WtymNOc6haNz0_5bgCKNuqtX3Y9Dc9M2jRT1fhHR This is the link to an article that use a similar data set constructed by CCES.
3. *What (other) tasks could the dataset be used for?*
   - Similar studies could be conducted using the 2020 CES presidential election data.
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?*
   - There are no signs of any harm in using the data set as it is conducted to only understand voters' behavior.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
   - Ensure to use the data set with the correct year since CES is an ongoing study that provides information about American's voting behaviors.

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 data set is publicly provided so there is no specifies in any third parties that the data will be distributed to

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - As the study is still ongoing, the teams from universities will maintain the dataset

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

    - For any information needed, data users can contact Havard Dataverse directly or Brian Schaffner from Tufts University (brian.schaffner@tufts.edu)

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

    - There is no erratum in the dataset

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)?*

    - There will be no update on the 2020 data,

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.*

    - There are no limits to the retention of data

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.*

    - Since the study started in 2006 and is ongoing, the previous version of data throughout the years is on Havard Dataverse.

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.