Datasheet for 'An Analysis of Voter Factors in the 2020 US General Election'*

Yuchen Chen

March 17, 2024

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 enable analysis of presidential candidate's team build effective policies to voters, potentially increasing their support rate and appeal. The main goal was to really look into how things like whether someone is a man or woman, their race, how old they are, and where they live can change who they decide to vote for. But this dataset is only for 2020, which is during the epidemic period, which is a relatively special historical background. Therefore, the results of this study have certain time limitations.

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

   - The dataset was created by a team including Brian Schaffner, Stephen Ansolabehere, and Sam Luks. It was a part of the CCES Dataverse, which is associated with Harvard University and Tufts University. The study includes data from a survey of American adults and aims to provide a comprehensive look at voter opinions and behaviors for the 2020 US General Election.

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 Cooperative Election Study (CES) for the year 2020 was supported by the National Science Foundation (NSF) under the award number 1948863, managed by Program Officer Jan Leighley.

4. *Any other comments?*

*Code and data are available at: https://github.com/Victor1114/Political-support-in-the-United-States

- TBD

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.*
    - In studies like the 2020 Cooperative Election Study (CES), the instances represent individual survey responses from 61,000 American adults voters with many different aspects like, various regions, ages, genders, racial groups, etc.
2. *How many instances are there in total (of each type, if appropriate)?*
    - 61,000 American adults
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).*
    - It is a sample of the larger set of eligible voters in the United States but it is not all possible instances because it would be impractical to survey every single voter. The sample is represented larger population, which participants selected to reflect the demographics and geographic distribution of the nation's electorate. Representativeness is validated by comparing the sample demographics to known characteristics of the population, such as census data, and ensuring that the sample includes people from various regions, ages, genders, racial groups, etc.
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.*
    - Gender, age, race, census area, and presidential preferences of voters.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
    - **Gender**: The voter is male or female.
    - **Age**: The age of the voter at the time of the survey.
    - **Race**: The race or ethnicity the voter identifies with.
    - **Census Area**: Geographic information which state of US the voter's residence.
    - **Presidential Preferences**: The voter vote Biden or Trump.
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.*

- It's possible because in guide of dataset, the first category shows some individual instances didn't fully answer the survey which makes some individual instances will miss information.

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.*

   - There is not relationship stated in survey or dataset between survey responses, but we can analyze the dataset by analyzing the relationship different responses with different features and how their behavior in voting to made relationship explicit.

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.*

   - The paper "An Analysis of Voter Factors in the 2020 US General Election" focuses on analyzing voter behavior and preferences based on survey data, the primary goal is to understand the relationship between demographic factors and presidential preferences among voters, rather than predicting outcomes. Therefore, the paper not doing data splits. But we do some testing to make sure the dataset which we cleaned is good and completed to use.

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

   - The dataset may have some potential sources of error or noise, like response bias which the tendency of respondents to answer questions untruthfully or in a socially desirable manner or non-response bias which people do not fully respond the question.

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.*

    - This paper primarily utilizes the 2020 Collaborative Election Study (CES) dataset, which is self-contained in terms of gathering voter demographics, opinions, and presidential preferences. The dataset does not inherently link to or rely on external resources like websites or tweets for its core data.

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.*

- It is not be considered confidential, because everyone can get data from Harvard Dataverse.

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.*

    - This dataset does not contain offensive, insulting, threatening, or might otherwise cause anxiety, but political topic covered in the survey could potentially be sensitive or provoke strong emotional responses in individuals with differing views.

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.*

    - **Gender**: The voter is male or female.
    - **Age**: The age of the voter at the time of the survey.
    - **Race**: The race or ethnicity the voter identifies with.
    - **Census Area**: Geographic information which state of US the voter's residence.
    - **Presidential Preferences**: The voter vote Biden or Trump.

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 impossible because all information or survey question are general, there are not distinct information that can identify individuals.

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.*

    - Race or Ethnic Origins: The dataset includes specific questions about Asian and Hispanic heritage.
    - Religious Beliefs: The dataset includes questions about specifying different religious affiliations and levels of religious engagement.
    - Health Data: The dataset includes questions related to the COVID-19 pandemic which indirectly relates to health data.

16. *Any other comments?*

    - TBD

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 dataset was collected through survey responses reported by subjects, covering their demographic details, political opinions, and voting behavior.Some key portions of the data, were validated by matching individual records to the Catalist database.

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 dataset was collected via online surveys conducted by YouGov.

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 was created through two process. First, they picked a random group of U.S. adults with detailed demographic info from sources like the American Community Survey and voter registries. Then, for each person in this group, they found someone with similar demographics from YouGov's online panel to ensure the sample reflected the broader population accurately.

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 dataset was collected via online surveys conducted by YouGov. The rewards includes points that can be redeemed for rewards, such as gift cards or small amounts of money.

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 dataset was collected over two periods: pre-election from September 29 to November 2, and the post-election 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.*

   - The dataset and its guide doesn't specifically mention the details of 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 dataset was collected directly from individuals through online surveys conducted 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.*

    - Participants in the CES20 survey were informed about the data collection through a notice provided by YouGov , but there are no specify about how to informed in website.

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, individuals participating in the CES20 survey provided their consent for the collection and use of their data.

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 mentioned in website or materials, but usually participants are informed about their right to withdraw consent at any stage of the survey.

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 mentioned in website or materials, but usually academic studies incorporate ethical considerations and privacy protections.

12. *Any other comments?*

    - TBD

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 are no Preprocessing/Cleaning/Labeling of Data showing in website or materials , but I cleaned it when I use it to analyze. I only clean out various regions, ages, genders, racial groups and presidential preferences as variables in new dataset.
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.*
    - https://github.com/Victor1114/Political-support-in-the-United-States/tree/main/data/raw_data
    - https://dataverse.harvard.edu/file.xhtml?fileId=4949558&version=4.0
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - https://www.r-project.org/
4. *Any other comments?*
    - TBD

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - The CES20 dataset is used to analyze political behavior, voter preferences, electoral outcomes, and public opinion trends in the United States during the 2020 election cycle.
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://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/E9N6PH
3. *What (other) tasks could the dataset be used for?*
    - Studies on social identity and political polarization.
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 CES20 dataset's collection may potentially leading to bias if not carefully considered in analyses.To reduce such risks, users should critically evaluate the dataset's representativeness.

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

   - The CES20 dataset focus on political opinions and behaviors, should not be used for tasks that could lead to privacy invasions or discrimination.

6. *Any other comments?*

   - TBD

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.*

   - No, The CES20 dataset is typically made available to researchers and the public through academic repositories by Harvard Dataverse.

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

   - The CES20 dataset is often distributed by Harvard Dataverse and available for download in format CSV. It is usually assigned a DOI to ensure it can be easily cited and accessed.

3. *When will the dataset be distributed?*

   - Aug 4,2021.

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.*

   - Any people can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission.

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 any limitation.

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 any limitation.
7. *Any other comments?*
    - TBD

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - IQSS, Harvard University / Tufts University.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - On the website https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/E9N6PH, there is an option called "contact owner".
3. *Is there an erratum? If so, please provide a link or other access point.*
    - No findings about erratum in website.
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)?*
    - Will not be updated because the data effectiveness is very low since for a long time.
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 mentioned on the website,but usually institutional data policies will schedule for deletion.
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.*
    - No, the dataset is not useful since it was created for 2020 presidential election which is already past.
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.*

- They have to contact the research team first, because people is impossible to have those data if they want to ensure consistency and accuracy with Harvard Dataverse.

8. *Any other comments?*

    - TBD