Datasheet for '2020 US voter file'*

Hyuk Jang

2 April 2024

This paper looks into the dataset 2020 US voter file using the datasheet. We look into the motivation, composition, collection process, preprocessing/cleaning/labelling, uses, distribution, and maintenance.

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 dataset was made to help understand American voters better, like their backgrounds and political views. There wasn't a ready dataset available with detailed info on US voters, so this one was put together to fill that gap.
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 put together by a team at a private company that works with data, especially about politics and voters.
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 company itself paid for making this dataset; there wasn't any outside funding or grants involved.
4. *Any other comments?*
    - TBD

Composition

*Code and data are available at: https://github.com/anggimude/2020-US-Cooperative-Election-Study.git.

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 the dataset represent individual voters in the United States, each with associated demographic and political attributes.

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

    - The total number of instances in the dataset is approximately 150 million, each representing a unique voter record.

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 voter records from a larger set, namely a US voter file record from a private company. This sample was chosen to be representative of the larger set by training a model on the 2020 US Cooperative Election Study and post-stratifying it, on an individual basis, based on the voter file.

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 various features such as demographic information (age, gender, income), geographic location, and etc.

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

    - The dataset may include a target label indicating the voter's political affiliation or voting behavior.

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.*

    - Some instances may have missing data due to various reasons such as incomplete voter records or data collection errors.

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 dataset may include implicit relationships between individual voters, such as interactions within social networks or affiliations with specific political groups.

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.*

   - No

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.*

    - It is self-contained

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.*

    - Yes. US voter file record data is confidential.

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. US voter file record is data for a person with regards to their age, gender, resident, 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.*

    - Yes. As stated the dataset contains information about individual's voting record, age, gender, and etc.

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. Political information can be identified.

16. *Any other comments?*

    - NA

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 about each voter is provided from a company's voter list. Then, we used a study from 2020 about US elections to make our data better. We checked and corrected the information about each voter to make sure it was accurate.

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?*

    - We used computer programs to collect and organize the voter data from the company's list and the 2020 election study. We tested these programs to make sure they worked correctly.

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

    - We took a sample of voters from the bigger voter list. We used a method that gave each voter in the list a chance to be chosen for the sample. This way, our sample represents different kinds of voters.

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)?*

    - Our team of data experts collected the voter information. They are employees of our company and were paid according to company rules.

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.*

    - We collected the voter data around the time of the 2020 US presidential election. This matches the time when the voter information was relevant and up-to-date.

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)?*

    - We obtained the data from a third-party source, specifically a private company's voter file record.

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.*

    - NA

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.*

    - NA

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).*

    - NA

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.*

    - NA

12. *Any other comments?*

    - NA

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.*
    - No
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/anggimude/2020-US-Cooperative-Election-Study
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - R Core Team (2023)
4. *Any other comments?*
    - NA

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - The dataset has been utilized in various studies focusing on political behavior analysis, voter profiling, and electoral predictions. It can be used to understand voter preferences and the impact of various factors on election outcomes
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://github.com/anggimude/2020-US-Cooperative-Election-Study.git
3. *What (other) tasks could the dataset be used for?*
    - Predictive modeling of voter behavior in future elections
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?*
    - No

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - The dataset should not be used for purposes that could result in discrimination, unfair treatment, or privacy violations.
6. *Any other comments?*
    - NA

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.*
    - Access to the dataset may be granted under certain conditions to ensure compliance with data privacy regulations and ethical guidelines.
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 could be distributed through various means like on a secure website, access via API, or hosted on platforms like GitHub
3. *When will the dataset be distributed?*
    - April 2, 2024
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.*
    - April 2, 2024
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?*

- NA

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - Hyuk Jang
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - https://github.com/anggimude/2020-US-Cooperative-Election-Study.git
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 it will 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.*
   - If the dataset contains data related to individuals, there may be limits on the retention of this data in accordance with data protection regulations or ethical considerations.
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 may or may not continue to be supported, hosted, or maintained, depending on the policies of the entity responsible for the dataset. If older versions become obsolete, this will be communicated to dataset consumers through appropriate channels to ensure they are aware of the latest version.
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.*
   - One can fork the Github repository or download as a zipfile and work on top of it.
8. *Any other comments?*
   - NA

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.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.