Datasheet for 2020 US Cooperative Election*

Yihang Cai

March 29, 2024

The 2020 US Cooperative Election Voter File Dataset, developed by academic researchers and a non-profit dedicated to enhancing democracy, centralizes detailed voter information across the US. Containing about 250 million records, it includes anonymized voter demographics, affiliations, and voting history, compiled from public records for comprehensive representativeness. This dataset, preprocessed for uniformity and usability, supports analyses in political campaigns, academic studies, and policy-making. It's maintained under strict ethical guidelines, ensuring data integrity and privacy, making it a pivotal resource for electoral research and democratic engagement.

Extract of the questions from @gebru2021datasheets.

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 facilitate a comprehensive analysis of voter behavior, demographics, and patterns across the United States to inform political campaigns, academic research, and public policy development. It aims to fill the gap of a centralized, detailed, and accessible voter database for the entire country.
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 compiled by a consortium of academic researchers from various universities in collaboration with a non-profit organization dedicated to enhancing democratic engagement.
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.*

[*] Code and data are available at: https://github.com/peachvegetable/sta302-mini-essay-12-MRP

- The project was funded by ...

4. *Any other comments?*

    - The dataset aims to be a living document, updated regularly with new voter registrations, changes in voter information, and election outcomes.

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.*

    - Each instance represents an individual voter, including their demographic information, voting history, party affiliation, and other relevant attributes.

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

    - Approximately 250 million instances, representing the estimated number of eligible voters in the US.

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 comprehensive aggregation of all registered voters in the US, compiled from state and local election boards. Efforts were made to ensure representativeness across geographic regions, demographics, and political affiliations.

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 includes voter ID (anonymized), age, gender, race/ethnicity, party affiliation, voting frequency, and geographical location (state, county, precinct).

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

    - Each instance includes labels such as party affiliation and voting history in past elections.

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 lack complete voting history due to variations in record-keeping practices across jurisdictions.

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.*

    - Voter instances are linked to electoral events, geographic locations, and demographic groupings.

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.*

    - Not applicable; the dataset is intended for analysis and modeling rather than machine learning training/testing.

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

    - Minor inconsistencies may exist due to discrepancies in state and local record-keeping.

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.*

    - The dataset does not rely on external resources; all information 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.*

    - The dataset does not include data protected by legal privilege, such as doctor-patient confidentiality or individual's non-public 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.*

    - The dataset itself does not contain information that is offensive, insulting, threatening, or likely to cause anxiety. However, the inclusion of political affiliations and voting history may be considered sensitive by some individuals. Users are advised to handle this data responsibly and consider the implications of its use in their analyses.

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.*

    - The dataset identifies sub-populations based on available voter registration information, such as age and gender. Sub-populations are also identifiable by geographical location (state, county, precinct), political affiliation, and in some cases, race or ethnicity, depending on the information provided at registration. The distribution of these sub-populations reflects the demographic composition of the electorate and varies by region.

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.*

    - While direct identifiers like names and addresses are removed or anonymized, it is theoretically possible, though challenging, to re-identify individuals indirectly by combining the dataset with other publicly available datasets. Measures are taken to mitigate this risk, including data aggregation and limiting access to sensitive or detailed fields.

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 dataset includes data that might be considered sensitive, such as political affiliations, voting history, and in some cases, race or ethnicity. This information is collected as part of standard voter registration processes and is treated with care to ensure compliance with privacy laws and ethical standards. The dataset does not include direct personal identifiers, financial data, health data, biometric or genetic data, forms of government identification like social security numbers, or criminal history.

16. *Any other comments?*

    - Users of this dataset are encouraged to respect the privacy and confidentiality of voter information. Ethical considerations should guide the use of the dataset, especially in analyses that could impact individual voters or groups. Researchers and analysts should also be aware of the legal framework governing the use of voter data in their jurisdiction to ensure compliance.

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.

- Data were compiled from publicly available voter registration records, with additional information inferred from census and public demographic datasets.

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?*

   - Automated scripts aggregated data from public sources, with manual verification processes to ensure accuracy and consistency.

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

   - Not applicable; the dataset aims to include all registered 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)?*

   - The project was conducted by a team of data scientists, political scientists, and legal experts to ensure compliance with data privacy laws.

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.*

   - Data compilation began in January 2021 and was completed in December 2021, with ongoing updates.

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 project underwent ethical review by an institutional review board, ensuring compliance with privacy laws and ethical research standards.

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

   - Primarily from public voter registration records, with no direct data collection from individuals.

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.*

    - In the case of a private company's voter file, individuals might not be directly notified by the company about the collection of their data for this specific dataset. The data could be compiled from various sources, including publicly available information and possibly proprietary data acquired through partnerships or services. The notification would depend on the original source of the data and the policies of the entity collecting it.

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.*

    - For a dataset compiled by a private company, the consent process might differ significantly. If the dataset includes data beyond public voter records, such as consumer data or proprietary insights, the company would likely need to obtain consent from individuals for the collection and use of such data, depending on the jurisdiction and applicable privacy laws (like GDPR or CCPA). The specifics of how consent was obtained would be determined by the company's privacy policy and data collection practices.

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).*

    - A private company handling voter and additional consumer data would typically provide a mechanism for individuals to manage their consent and data preferences, in compliance with data protection regulations. This could include options to opt-out, delete personal data, or modify what information is held by the company. The details of this mechanism would be available in the company's privacy policy or through direct communication channels like customer service.

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.*

    - Given the involvement of a private company and the potential for the dataset to include proprietary or sensitive data, conducting a Data Protection Impact Analysis (DPIA) or similar assessment would be more likely and might even be required, depending on the regulatory environment. The outcomes of such an analysis would help in identifying and mitigating risks related to privacy, data protection, and potential harm to individuals. This analysis would be part of the company's internal documentation and might not be publicly accessible, though summaries or findings could be shared with stakeholders or regulatory bodies as required.

12. *Any other comments?*

    - When dealing with a voter file from a private company, additional layers of complexity are introduced due to proprietary data, commercial interests, and possibly enhanced privacy concerns. Users of this dataset must be diligent in understanding the source of the data, the legal and ethical implications of its use, and any limitations or biases that might be present. It's crucial to adhere to strict ethical standards and privacy laws, ensuring that any analysis does not compromise individual privacy or lead to unintended consequences.

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.*

    - The dataset underwent significant preprocessing to ensure uniformity and usability. This included anonymization of personal identifiers, normalization of state and county names, categorization of age groups, and imputation of missing values for certain demographic fields. Voter histories were standardized to a common format to facilitate analysis.

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 "raw" data, as initially received from various sources, was preserved separately from the processed dataset to enable validation of the preprocessing steps and support potential future uses that require access to the original data formats. Access to the raw data is restricted and subject to approval, ensuring compliance with data protection and privacy standards.

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

    - The software tools and scripts used for data preprocessing, cleaning, and labeling are available in a dedicated GitHub repository. These tools are open-source, allowing for community review, contributions, and use in similar projects.

4. *Any other comments?*

- The preprocessing stage was meticulously documented to provide transparency about the decisions made and the methodologies applied. This documentation is available alongside the dataset to assist users in understanding its structure and potential limitations.

Uses

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

   - The dataset has already been employed in a range of tasks, including academic research on voter behavior, studies on electoral trends, and predictive modeling by political campaigns to understand voter demographics and preferences.

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.*

   - A dedicated repository that catalogs studies, analyses, and systems utilizing this dataset is maintained by the data custodian. This repository serves as a resource for researchers and analysts to explore existing work and collaborate on new projects.

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

   - Beyond electoral analysis, the dataset holds potential for sociological research, studies on geographic distribution of political affiliations, and the development of models to predict voter turnout 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?*

   - Users should be cognizant of the inherent biases that may exist due to the dataset's composition and collection methodologies. For example, over-representation of certain demographics could lead to skewed analyses. Mitigating these risks involves applying appropriate statistical techniques and considering diverse data sources.

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 individual voter targeting, manipulation of voter behavior, or any form of discriminatory practice. It is intended for research and analysis that respects the privacy and autonomy of individuals.

6. *Any other comments?*

- The ethical use of this dataset is paramount. Users are encouraged to consider the societal impact of their work and engage in responsible data science practices.

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 dataset will be made available to qualified entities, including academic institutions, research organizations, and political campaigns, under strict data use agreements that outline the terms of use 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)?*

    - Access to the dataset is facilitated through a secure API, ensuring controlled and traceable data access. The dataset itself does not currently have a DOI, but each version released will be uniquely identifiable to maintain a clear version history.

3. *When will the dataset be distributed?*

    - The dataset is scheduled for an initial release, with subsequent updates planned post-election cycles or as significant new data becomes available.

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.*

    - Distributed under a specific license that balances accessibility with the need to protect the data and its subjects. The terms of use will explicitly prohibit misuse and ensure users commit to ethical standards.

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.*

    - The dataset is subject to copyright and data protection laws. Users will need to comply with these regulations, which will be clearly outlined in the terms of use.

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.*

    - Certain data within the dataset may be subject to export controls or other regulatory restrictions, particularly data that can be considered sensitive. Users will be informed of such restrictions and must agree to comply.

7. *Any other comments?*

   - None

Maintenance

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

   - The dataset will be maintained by the private company, with a dedicated team responsible for updates, quality assurance, and user support.

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

   - For inquiries or support, users can contact the data management team via a dedicated email address provided with the dataset documentation.

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

   - An erratum section will be available for reporting errors or inconsistencies in the dataset. Regular updates will be communicated through an official channel, such as a dataset newsletter or a dedicated section on the company's 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)?*

   - he dataset will be updated periodically to reflect new data or corrections. Users will be notified of such updates through the chosen communication channel.

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.*

   - In compliance with data protection regulations, the dataset will include information on the retention period of data and the process for data deletion or anonymization as required.

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 will be archived and accessible for historical comparison and audit purposes. Changes between versions will be documented and communicated.

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.*

    - External contributions to the dataset, such as corrections or additions, will be vetted through a structured process to ensure data integrity and quality. Contributions that meet the criteria will be incorporated into future versions of the dataset.

8. *Any other comments?*

    - The maintenance and support of the dataset are designed to ensure its long-term value and relevance to users, with a strong commitment to data quality, user support, and ethical standards.