Datasheet for 'Cooperative Election Study Common Content, 2022'*

My subtitle if needed

First author Another author

December 3, 2024

The 2022 Cooperative Election Study (CES) ...

This datasheet answers a set of questions from Gebru et al. (2021) about the Cooperative Election Study Common Content (CES) 2022 Survey Dataset from Schaffner, Ansolabehere, and Shih (2023), obtained from Harvard Dataverse.

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 2022 CES survey dataset was created to understand the American electorate
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - The 2022 CES dataset was
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.*
4. The Cooperative Election Study (CES), formerly known as the Cooperative Congressional Election Study (CCES) has received funding and support from the National Science Foundation in all even years (midterm and presidential election years) since 2010 (Schaffner and Kuriwaki 2022).
   - The CES has received the following awards/grants from the National Science Foundation:

*Code and data are available at: https://github.com/taliafabs/US-Midterms-2022.git

Table 1: National Science Foundation awards that the CES has received since 2010, as indicated on the FAQ page (Schaffner and Kuriwaki 2022).

Study year	NSF Award Number
2022	2148907
2020	1948863
2018	1756447
2016	1559125
2014	1430505
2012	1225750
2010	0924191
2010 panel	1430473 and 1154420
2014 panel	1430473 and 1154420

5. *Any other comments?*
    - 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.*
    - Each instance in the dataset represents one American adult. This is the only type of instance.
2. *How many instances are there in total (of each type, if appropriate)?*
    - The 2022 CES survey dataset contains 60,000 instances in total.
    - Each instance is of the same type: an American adult.
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 2022 CES survey dataset is a sample. Its target population, or the larger set that it is meant to be representative of, is all American adults (Schaffner, Ansolabehere, and Shih 2023).
    - Weighting and respondent validation were used to make the CES survey sample nationally representative (Schaffner, Ansolabehere, and Shih 2023).

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 of the raw data consists of 707 variables.
    - The 2022 CES has five parts: sample identifiers (including state and congressional district), demographic and profile questions, pre-election questions, post-election questions, and contextual data (Schaffner, Ansolabehere, and Shih 2023).
    - The 707 variables include demographic information matched to a respondent's TargetSmart or YouGov database voter record, their TargetSmart or YouGov voter record match status, answers to questions in the pre-election wave, and answers to questions in the post-election wave.

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

    - The raw dataset does not associate a particular label or target with each instance. However, the data can be prepared and subsetted to train models to predict vote preference, partisan affiliation, stance on an issue such as abortion or the economy

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.*

    - TBD

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 are no explicit relationships between individual instances in the 2022 CES survey dataset.

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.*

    - There are no recommended data splits for training, development/validation, and testing. In fact, Schaffner, Ansolabehere, and Shih (2023) actually warns against randomly subsetting the dataset or splitting it into training, validation and testing datasets because the CES is a large survey that includes observations of small subpopulations. Splitting the data or using a subsample as training data may cause a small sub-population to be excluded and lead to a model that mislabels or miscategorizes them (Schaffner, Ansolabehere, and Shih 2023).

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

- Errors, sources of noise, and redundancies are possible in this dataset. TargetSmart voter validation records are available to determine if there is a record of an individual respondent voting in an election, but voter records only indicate that someone voted and how they voted (i.e. on election day, absentee, via mail, or early), not *who* they voted for. The only way for the 2022 CES survey to determine who a respondent voted for in the 2020 presidential election is by asking them. The variable `presvote2020post` includes each respondent's answer to the post-election wave question about 2020 presidential vote. There is no way to verify that a respondent answered this question truthfully. (**silvergut?**) found that this is subject to social desirability bias: respondents who actually voted for Trump in 2020 may say that they voted for Biden in 2020 because Trump is polarizing or because they want to say they voted for the winning candidate. Additionally, identifying non-voters is subject to error. Schaffner, Ansolabehere, and Shih (2023) outlines three ways to do this: identify both TargetSmart matched non-voters and voters with no TargetSmart record as non-voters, identify only TargetSmart matched non-voters as non-voters, or identify TargetSmart matched non-voters and unmatched voters who indicated in the post-election survey wave that they did not vote as non-voters. None of these three options is perfect; it is possible that some respondents who did not match to a TargetSmart record actually voted, but those who did not match to a TargetSmart record are more likely to be non-voters than voters (Schaffner, Ansolabehere, and Shih 2023). Using self-reporting to determine which un-matched respondents are non-voters relies on the assumption that neither voting nor abstaining is more socially-desirable, and respondents feel no incentive or social pressure to answer one way over the other (Schaffner, Ansolabehere, and Shih 2023).

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 2022 CES survey data set is self contained. It does not rely on any external resources.

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.*

    - No, the 2022 CES survey data set does not contain data that might be considered confidential.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threat-*

ening, or might otherwise cause anxiety? If so, please describe why.

- There is no reason for me to believe that the 2022 CES survey dataset contains anything that might be offensive, insulting, threatening, or 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.*

    - Yes, the dataset includes information about each respondent's age, gender, race, marital status, home state, congressional district, education, political preferences, and sexual orientation.
    - However, the dataset does not specifically identify them as subpopulations.
    - For example, those who voted for Trump in 2020, do not have a college education, and who have low trust in the government are underrepresented in the dataset because they are less likely to respond to a survey. To address this, this subpopulation was assigned a higher weight in the dataset (Schaffner, Ansolabehere, and Shih 2023).

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, it is not possible to identify individuals either directly or indirectly from the 2022 CES dataset or in combination with another dataset. The dataset only contains voter status and responses to the survey questions; there is no uniquely identifiable information provided in the dataset. Sure, we could know that a survey respondent is a 20-year-old Asian female who lives in California, voted for Biden in 2020, voted for Democrats in the 2022 mid-term election, and has specific views on abortion and the economy. However, the dataset contains no information that could be used to identify who that respondent actually is.

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, the 2022 CES survey dataset contains data that might be considered sensitive. The pre-election wave contains questions about race, gender, marital status, religion, sexual orientation, union membership, income, vote preferences, trust in government, and views on controversial issues including abortion, guns, economic policy, and foreign policy. All survey responses are anonymous.
    - However, there is no financial data, health data, government identification such as a social security number or even a name, criminal history, or any other kind of data that can actually be used to identify an individual respondent is included.

- The survey is designed to maintain the privacy and anonymity of respondents, and so that those who use the dataset will not be able to trace responses back to a natural person.

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.*

    - TBD

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?*

    - TBD

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

    - TBD

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)?*

    - TBD

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.*

    - TBD

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.*

    - TBD

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

    - I did not collect the data that I used for my project from the individuals in question directly. I obtained the CES 2022 data from Harvard dataverse.

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.*

    - TBD

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.*

    - TBD

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).*

    - TBD

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.*

    - TBD

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.*

    - TBD

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.*

- Yes. The raw data was saved, but it is not available in this github repository due to file size limitations. The raw data can be found using the following link: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910/DVN/PR4L8P

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

   - Yes. The statistical programming language R from (**citeR?**) was used to preprocess, clean, and label the raw data.

4. *Any other comments?*

   - TBD

Uses

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

   - Yes, the CES 2022 survey dataset has been used for tasks. It has been used for political science publications in the past. I used the dataset for a project about vote preference and voter turnout.

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.*

   - Recent publications that used the CES survey dataset can be found at this link: CES Recent Publications

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

   - TBD

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?*

   - TBD

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

   - This dataset should not be used to forecast U.S. presidential elections. Its purpose is to examine how Americans view their government, hold it to account during elections, how American adults voted in 2022, and how different geographic, demographic, and social factors influenced voting behavior (Schaffner, Ansolabehere,

and Shih 2023). Biden 2020 and college-educated voters are over-represented in this dataset. Weighting was used to account for groups/demographics who are less likely to respond to this survey, but this is not a representative sample of the American electorate's voting preferences or behavior. Neither Donald Trump nor Joe Biden will be on the ballot in any future presidential election, so that is another reason why this dataset should not be used to train an election forecasting model.

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.*
    - The 2022 CES survey dataset is publicly available for download via Harvard Dataverse. Any individual or researcher from any company, institution, or organization can download and use it.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - This dataset was distributed via Harvard Dataverse and it can be downloaded there. Additionally, it has a GitHub repository and a DOI.
3. *When will the dataset be distributed?*
    - The first release of the 2022 CES Common Content dataset was first distributed via Harvard dataverse on March 20, 2023. An updated version of the 2022 CES Common Content dataset with vote validation appended was distributed via Harvard Dataverse on September 8, 2023.
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.*
    - The License/Data Use Agreement can be found at this link: https://creativecommons.org/publicdomain/zero/1.0/
    - Schaffner, Ansolabehere, and Shih (2023) makes it clear that individuals using the dataset for their own research projects are required to cite it.
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.*

- As far as I am aware, no third parties have imposed IP-based or other restrictions on the data associated with the instances.
- The dataset is publicly available, so there are no licensing terms or fees required to access it.

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 export controls or regulatory restrictions apply to the dataaset or individual instances.

7. *Any other comments?*

    - TBD

Maintenance

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

    - TBD

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

    - The owner of the dataset can be contacted by clicking the "contact owner" button on its Harvard Dataverse page.

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

    - TBD

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)?*

    - TBD

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.*

    - Older versions of the dataset are available and supported via Harvard dataverse.

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.*

    - There is no method for others to extend, augment, or build directly on the dataset.

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.

Schaffner, Brian, Stephen Ansolabehere, and Marissa Shih. 2023. "Cooperative Election Study Common Content, 2022." Harvard Dataverse. https://doi.org/10.7910/DVN/PR4L8P.

Schaffner, Brian, and Shiro Kuriwaki. 2022. *Frequently Asked Questions*. Harvard University. https://cces.gov.harvard.edu/frequently-asked-questions.