Datasheet for 'Elections - Voter Statistics*

Janel Gilani

April 15, 2024

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 created to provide voter statistics for the 2022 municipal election in Toronto. It includes information on the number of eligible electors, additions and corrections to the voter's list, and the number of voters who cast ballots, broken down by ward and subdivision.
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 the City Clerk's Office on behalf of the City of Toronto.
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 creation of the dataset was funded by the City of Toronto.
4. *Any other comments?*
    - No comments.

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.*

*Code and data are available at: https://github.com/JanelGilani/toronto-voter-turnout

- The instances in the dataset represent the wards of the City of Toronto. Each instance corresponds to a specific ward within the city, providing voter statistics for the 2022 municipal election.

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

    - There are 25 instances in total, each representing a unique ward within the City of Toronto.

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 contains all possible instances of wards within the City of Toronto for the 2022 municipal election. As such, it is not a sample but rather a complete enumeration of all wards and their corresponding voter statistics for the specified election period. Therefore, the dataset is representative of the entire city and does not require validation or verification of representativeness.

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 voter statistics for a specific ward within the City of Toronto for the 2022 municipal election. The dataset includes 23 continuous variables, such as the number of eligible electors, the number of voters, and the eligible voter turnout percentage. Additionally, each instance includes a categorical feature identifying the ward represented.

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

    - Yes, the label associated with each instance is the ward ID, which represents the unique identifier for each ward within the City of Toronto.

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 no missing information from individual instances. The dataset provides comprehensive voter statistics for each ward, and all relevant features are included without any missing values.

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 relationships between individual instances.

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.

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

    - There are no 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.*

    - The dataset 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.*

    - There is no confidential data, and the dataset 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.*

    - No, the dataset does not contain data that, if viewed directly, 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.*

    - The dataset does not identify any sub-populations such as age or gender. It primarily focuses on voter statistics for different wards in the City of Toronto, without specific demographic information about individual voters. Therefore, there are no identifiable subpopulations within the dataset.

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 in any way.

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.*

    - None.

16. *Any other comments?*

    - None.

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 associated with each instance was acquired through interviews conducted with subjects from all 25 states. Subjects reported the data directly, providing information relevant to voter statistics for the respective wards in the City of Toronto.

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 mechanism used to collect the data was manual human curation. This involved individuals conducting interviews and recording the information provided by the subjects. The validation of this process likely involved quality checks during data entry and potentially cross-referencing the collected information with other sources 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)?*

    - The dataset represents all possible instances, meaning it contains data for every ward in the City of Toronto. Therefore, there was no sampling strategy involved as the dataset encompasses the entire population of interest. However, if there were sampling involved, the strategy would likely have been geographic stratification, subdividing each state into regions and systematically selecting primary sampling units with probability proportional to the size of the strata.

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 likely involved individuals from the City Clerk's Office or other relevant departments within the municipal government of Toronto. These individuals would have been responsible for conducting interviews, gathering voter statistics, and maintaining the dataset. It's reasonable to assume that they were compensated as part of their regular employment with the municipal government, although specific details regarding compensation are not provided in the dataset information.

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 data was last refreshed on February 7, 2023, as indicated in the dataset details. This timeframe corresponds to the last update or revision of the dataset. The data collection process likely occurred during the 2022 municipal election period, which took place on October 24, 2022. Therefore, the data collection timeframe aligns with the creation timeframe of the data associated with the instances.

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.*

   - Ethical review processes were not 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)?*

   - We obtained the data via the Open Data Toronto website

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.*

   - The dataset information does not specify whether individuals were notified about the data collection process. However, given that the data is publicly available and pertains to voter statistics for the 2022 municipal election, it is likely that individuals were aware of the data collection process through public records and disclosures related to the election.

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.

- The dataset information does not provide details on how consent was obtained from individuals for data collection and use. Given that the data pertains to voter statistics for the 2022 municipal election, it is likely that consent was not explicitly requested or required, as this information is typically considered public record and collected as part of the election process.

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 mechanism to revoke consent was not provided.

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.*

    - There is no mention of an analysis of the potential impact of the dataset and its use on data subjects.

12. *Any other comments?*

    - None.

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.*

    - We cleaned the data by removing any unnecessary columns and renaming the columns to make them more descriptive. We also converted the data types of certain columns to ensure consistency and ease of 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 downloaded using Open Data Toronto is available in the folder data/raw_data/raw_data.csv.

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 Software is avalaible at https://www.R-project.org/

4. *Any other comments?*

    - None.

Uses

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

    - The dataset has not been used for other tasks yet.

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.*

    - No, there is no repository that links to any papers or systems that use the dataset.

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

    - The election dataset could be used for various analyses, such as predicting voter turnout, identifying trends in voter participation, and evaluating the impact of socio-economic factors on voting behavior. Additionally, the dataset could be used to assess the effectiveness of voter registration campaigns, analyze the distribution of voters across different wards, and explore the relationship between voter turnout and demographic characteristics.

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 dataset contains voter statistics for the 2022 municipal election in Toronto and does not include any personal information about individual voters. However, dataset consumers should be aware that the data pertains to voter participation and may be subject to privacy and confidentiality considerations. To avoid potential risks or harms, consumers should use the data responsibly and ensure that any analyses or interpretations are conducted in an ethical and unbiased manner. Additionally, consumers should be mindful of the limitations of the dataset and consider the context in which the data was collected when using it for future tasks.

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

    - Not applicable.

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 dataset is openly available and being used for personal uses only.

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 is distributed using Open Data Toronto and will be available for download as a CSV file. The cleaned dataset is available on Github.

3. *When will the dataset be distributed?*

   - The raw dataset is already available for download from Open Data Toronto. The cleaned dataset will be available on Github on 18th April 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.*

   - The dataset has been released under the Open Government License - Toronto. The cleaned dataset used in this paper will be available on Github under the MIT License.

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.*

   - There are no restrictions

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 such controls or restrictions are applicable.

7. *Any other comments?*

- None.

Maintenance

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

   - Janel Gilani

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

   - Can be contacted via github

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

   - There is no erratum available currently.

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)?*

   - Currently there is no plan of updating the dataset. If there are updates in the future, it will be done through Github.

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.*

   - The dataset was made publicly available and does not contain personal information about individuals. Therefore, there are no limits on the retention of data associated with the instances. The data will be retained indefinitely for research and analysis purposes.

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.*

   - The older versions would not be hosted. Dataset consumers will be able to check whether the dataset has been updated through Github commit history.

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 mechanism for accepting contributions from other users as of now.

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of ACM* 64 (12): 86–92.