Datasheet for ‘Toronto Poll Participation’*

Ruiying Li

December 13, 2024

Extract of the questions from Gebru et al. (2021).

The dataset discussed in this datasheet and used in “Analysis factors affecting voter participation rate”(Open data Toronto 2015a),obtained from Open data Toronto.

More relevant information could be found on the”Polls Regarding Changes in a Neighborhood”(Open data Toronto 2015b) by City of Toronto.

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 the analysis of voter participation patterns in Toronto. Specifically, it aims to understand how municipal policies, such as parking regulations and business improvement initiatives, influence voter participation. The structured dataset addresses consolidated and accessible data on neighborhood polls insufficiency and provides a foundation for quantitative analysis and policymaking.
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 curated and published by the City of Toronto to the public through portal called Open Data 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 dataset is maintained as part of the City of Toronto’s open government initiative, funded by municipal resources.
4. *Any other comments?*

*Code and data are available at:https://github.com/Liruiying0414/Toronto-polls-response-rate/tree/main.

- None.

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.*

    - Instances in the raw dataset represent neighborhood poll results in Toronto, specifically focusing on voter responses to municipal policy proposals. Each instance is linked to a unique poll and includes clear details about the poll's attributes with numeric and categorical outcomes.

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

    - The dataset contains 1,296 instances in total, but the number would change as it was update monthly by 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 represents a full collection of neighborhood poll results available during the data collection period. However, it might not include polls conducted outside this timeframe or under specific exceptional conditions. The dataset's representativeness is reflected in its broad coverage of neighborhood-level polls across Toronto.

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 in the raw dataset includes the following aspects: Poll identifiers (POLL_CD, POLL_ID); Address details (ADDRESS); Application type (APPLICATION_FOR); Voter counts (POTENTIAL_VOTERS, FINAL_VOTER_COUNT); Ballots information (BALLOTS_DISTRIBUTED, BALLOTS_CAST, BALLOTS_IN_FAVOUR, BALLOTS_OPPOSED, BALLOTS_SPOILED, BALLOTS_BLANK); Pass rates (PASS_RATE, RESPONSE_RATE_MET); Poll outcome (POLL_RESULT); Relevant dates (OPEN_DATE, CLOSE_DATE, MORATORIUM_DATE); Additional metadata (BALLOTS_NEEDED_TO_PROCEED).

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

- Yes, key targets include PASS_RATE, it indicates percentage of votes in favor of a proposal, and RESPONSE_RATE_MET (whether voter participation met required thresholds).

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 have missing values in fields like ADDRESS, likely due to incomplete administrative records or anonymization for privacy. Missing data affects a small subset of records and does not have huge impact dataset usability.

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.*

   - Yes, relationships between instances are made explicit through unique poll identifiers (POLL_CD and POLL_ID). These identifiers link related voting attributes to a specific poll.

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 dataset is not pre-split, allowing users to determine appropriate splits based on their analysis needs. For example, splits could be based on poll types (APPLICATION_FOR) or geographic regions (ADDRESS).

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

   - The dataset contains redundant records, some missing value, and minor inconsistencies in metadata labels, and noise introduced by human error during data entry. These issues are minimal and were addressed during preprocessing for the cleaned 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.*

- Yes, the dataset is self-contained and provides all necessary information for analyzing poll outcomes. However, external documents and policies referenced in the dataset such as city by-laws, was provided additional context.

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 dataset contains publicly available information collected from municipal records.

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 primarily contains neutral data about municipal poll results.

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 include explicit demographic data. However, geographic information (ADDRESS) could indirectly indicate neighborhood-level sub-populations.

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. The dataset does not include individual-level information; all data is aggregated at the poll level.

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.*

    - No. The dataset is limited to non-sensitive information about neighborhood-level polling outcomes.

16. *Any other comments?*

    - The dataset serves as a foundational resource for studying voter participation and evaluating the effectiveness of municipal policies.

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 is directly observable. It was collected through municipal records of neighborhood polls conducted in 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?*

   - Data collection involved administrative records from polling events managed by the City of Toronto.

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 includes polls across various neighborhoods, aiming for broad coverage of local civic events.

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 City of Toronto governments staff and administrative teams.

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 includes polls from multiple years, recording from 2015 to recent 2024, and reflects the timeframe of their respective collection.

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, it is not applicable as the dataset contains publicly available municipal data.

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

   - No,data was obtained directly from the City of Toronto's open data portal.

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.*

- It is not applicable as the data does not include individual-level information.

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.*

    - It is not applicable as the data is aggregated and public.

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, it is not applicable.

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 formal impact analysis has been noted, but the dataset is part of public transparency efforts.

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.*

    - Yes. The cleaned dataset involved removing duplicates, handling missing values, standardizing variable names, and creating derived variables such as response_rate.

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 and used as the basis for creating the cleaned analysis dataset.

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

    - Preprocessing steps were implemented only using R script, which are documented in the project repository "scripts" folder.

4. *Any other comments?*

    - The cleaned dataset is optimized for statistical modeling and research applications, by only focusing the variables we interested to and relate to the analyze project.

Uses

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

    - Yes, it has been used to study factors affecting voter participation rates in Toronto.

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 known repositories as of now.

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

    - It could be used to study civic participation, future election forecasting, and urban policy impacts.

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?*

    - There are no known risks, given the dataset's public and aggregated nature.

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 profiling or predictions beyond its civic context.

6. *Any other comments?*

    - None.

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.*

    - Yes, it will distribute via Open Data Toronto's portal due to its public availability, anyone who go to Opendata Toronto portal could download it.

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

   - It is available for download through Open Data Toronto as CSV files.

3. *When will the dataset be distributed?*

   - It is already publicly accessible in Open Data Toronto portal.

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 data is distributed under the terms of the Open Government License - Toronto. The license writes that everyone is free to "copy, modify, publish, translate, adapt, distribute or otherwise use the Information in any medium, mode or format for any lawful purpose." and no fee requirements.

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 restrictions beyond those stated in the open data license.

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, there are no export controls or other regulations apply to the dataset and individual instances.

7. *Any other comments?*

   - None.

Maintenance

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

   - The maintenance of the dataset are supported and funded by the City of Toronto through its Open Data portal.

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

   - It could be contacted through the City of Toronto's official contact email:"opendata@toronto.ca"

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

- Erratum is not specified

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)?*

    - Yes, dataset updates occur upon the closing and certification of each poll, and are managed by the City of Toronto.

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, it is not applicable.

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,older versions of the dataset is not observable and does not maintained.

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 does not have a mechanism for extending the dataset, but potential questions, or suggestions for improvement could be sent to the Open Data Team on Twitter, GitHub, Medium, or via e-mail opendata@toronto.ca.

8. *Any other comments?*

    - None.

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.

Open data Toronto. 2015a. *Polls Conducted by the City*. City Clerk's Office. https://open.toronto.ca/dataset/polls-conducted-by-the-city/.

———. 2015b. *Polls Regarding Changes in a Neighbourhood*. City Clerk's Office. https://www.toronto.ca/city-government/planning-development/polls-regarding-changes-in-a-neighbourhood/.