Datasheet for City of Toronto, 2023, Hate Crimes Open Data. City of Toronto Open Data Portal*

Arshh Relan

2024-12-03

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 document hate crimes reported in the City of Toronto. The primary purpose is to provide a comprehensive view of hate crimes, categorized by bias type, location type, and arrest outcomes, to support analysis and policy-making. In our study, the dataset enabled us to explore the factors influencing arrest likelihood in hate crime cases.
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 and maintained by the City of Toronto's Open Data Portal team.
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's creation and maintenance were funded by the City of Toronto.
4. *Any other comments?* The dataset reflects the City of Toronto's commitment to transparency and addressing hate crimes through publicly accessible data.

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 a reported hate crime, including details about the bias type, location type, and whether an arrest was made.

[*] Code and data are available at: https://github.com/Arshh-Relan/hate_crime_in_toronto

2. *How many instances are there in total (of each type, if appropriate)?* The dataset contains 1,350 reported hate crimes.
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 only includes reported hate crimes and does not account for unreported incidents, which could introduce reporting bias.
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 structured fields such as the type of bias (e.g., Race, Religion), location type, occurrence date, and arrest status.
5. *Is there a label or target associated with each instance? If so, please provide a description.* Yes, the target variable is arrest_made (1 for Yes, 0 for No).
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.* Certain fields, such as bias type or location type, may occasionally be missing due to incomplete reporting.
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.* No, the dataset treats each hate crime as an independent event.
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 can be split temporally (e.g., by year) or by categorical variables (e.g., bias type) for analysis.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.* Potential noise may exist due to inconsistencies in data entry and underreporting of certain types of hate crimes.
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 does not rely on 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 dataset does not include any personally identifiable or confidential information.

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 includes details of hate crimes, which could be sensitive or distressing for some users.
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, subpopulations are implicitly identified through bias types (e.g., race, religion, sexual orientation).
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 is anonymized and does not include any information that could identify individuals.
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 dataset includes information on sensitive topics like bias against race, religion, and sexual orientation.
16. *Any other comments?* The dataset provides a valuable resource for understanding hate crimes but must be used responsibly, given the sensitive nature of the data.

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 was collected from police reports of hate crimes filed with the Toronto Police Service.
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 was collected manually through police records and then digitized and structured by the City of Toronto's Open Data team.
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 all hate crimes reported to the Toronto Police Service and is not a sample.
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)?* Police officers and administrative staff were involved in collecting the raw data, while the Open Data team structured and published it.
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 spans multiple years, covering hate crimes reported from the early 2000s to the most recent year of publication.

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 formal ethical review process is documented, as the dataset is derived from public records.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?* The data was collected directly by the Toronto Police Service.
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.* Individuals reporting crimes are aware that their reports will be recorded but may not be aware of the dataset's publication.
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.* Consent is implied through the act of filing a police report.
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).* There is no documented mechanism for individuals to revoke consent.
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 documented impact analysis exists.
12. *Any other comments?* The dataset is anonymized to minimize potential harm to individuals.

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, bias types were consolidated into a single column, and arrest status was converted to binary.
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 dataset provided is a cleaned version; raw police records are not publicly available.

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 was conducted using tidyverse in (R Core Team 2023), and the code is available in our repository.
4. *Any other comments?* Cleaning focused on standardizing column names and addressing missing values.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.* Yes, it has been used in this study to analyze arrest likelihood in hate crimes.
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 centralized repository exists.
3. *What (other) tasks could the dataset be used for?* It could be used for temporal trend analysis, bias prevalence studies, or spatial analyses of hate crimes.
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?* Future uses must consider the dataset's sensitivity to avoid perpetuating stereotypes or misinterpreting findings.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.* It should not be used to profile individuals or groups or for any purpose that violates human rights.
6. *Any other comments?* Proper context and ethical considerations are critical for using this dataset.

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 is publicly available through the City of Toronto's Open Data Portal.
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 as a .csv file from the Open Data Portal.
3. *When will the dataset be distributed?* The dataset is already publicly accessible.
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.* Yes, it is distributed under the City of Toronto's Open Data 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.* No restrictions are imposed beyond 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 export controls apply.
7. *Any other comments?* Users must adhere to the dataset's terms of use.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?* The City of Toronto's Open Data team.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?* Via the contact form on the Open Data Portal website.
3. *Is there an erratum? If so, please provide a link or other access point.* No erratum is provided.
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, updates are made periodically as new hate crimes are reported.
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 documented limits exist.
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 are not explicitly maintained but may be archived.
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.* Contributions are not accepted; only the City of Toronto can update the dataset.
8. *Any other comments?* Maintenance aligns with the City of Toronto's commitment to transparency and open data.

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/.