Datasheet for 'Motor Collisions dataset'*

Qingyang Feng

27 Septe0ber 2004

This study explores the rising incidence of motorcycle accidents in urban areas, particularly in Toronto, highlighting that rear-end collisions are most prevalent during peak traffic hours on dry roads. Despite improvements in traffic safety measures, the analysis reveals that many accidents still occur in favorable conditions, raising concerns about rider behavior and awareness. By identifying these trends, the research aims to inform targeted interventions and policy decisions to enhance road safety for all users.

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 analyze urban traffic accidents in Toronto, specifically examining the influence of impact types, road conditions, and time of day on accident frequency and severity.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - Qingyang Feng
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.*
    - TBD
4. *Any other comments?*
    - TBD

Composition

*Code and data are available at: https://github.com/Fqy10987/Motor-Collision.git

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.*

   - The instances represent traffic collision records, which include information about the date, time, impact type, injury severity, and road conditions.

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

   - 18,957 recorded entries

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 recorded traffic incidents in Toronto since January 1, 2006

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 relevant attributes: date, time, accident number, impact type, and road condition.

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

   - Labels include injury severity (Fatal, Major, Minor, None).

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 records may have missing data in non-essential fields, but essential fields are preserved for analysis.

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.*

   - Relationships are established through the linking of incidents to their corresponding attributes.

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.*

- TBD

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

    - Potential inaccuracies in reported data due to police data entry variability

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 relies on the Open Data Toronto platform for data access.

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.*

    - Personal identifiers have been anonymized to protect the privacy of individuals involved.

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 traffic accident data which may evoke concerns for individuals sensitive to injury or fatality statistics.

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 can identify various groups based on injury severity and collision types.

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.*

    - Individuals are not directly identifiable due to anonymization.

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.*

- Data reflects traffic incidents that can be distressing due to the nature of injuries involved.

16. *Any other comments?*

    - TBD

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 was sourced from the City of Toronto's traffic incident records.

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 collected through standardized reporting from the Toronto Police Service.

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 recorded incidents since 2006, making it comprehensive.

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)?*

    - Data was collected by city authorities and traffic management 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.*

    - Data has been collected continuously since January 1, 2006.

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)?*

- Data was obtained via the city's reporting system.

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.*

    - Data was cleaned to filter essential columns for analysis without losing significant entries.

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.*

    - Raw data is available from the Open Data Toronto portal.

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 and associated packages (tidyverse, ggplot2, etc.) were used for data analysis.

4. *Any other comments?*

    - TBD

Uses

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

    - The dataset has been used for analyzing traffic accident trends and informing policy decisions.

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.*

    - Code and data are available on GitHub:https://github.com/Fqy10987/Motor-Collision.git

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

    - The dataset can be used for further research into traffic safety measures and urban planning.

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 completeness and accuracy of data can impact analyses related to traffic safety and policy formulation.

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 any discriminatory analysis or profiling of individuals based on accident involvement.

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.*
   - Yes, the dataset will be distributed to third parties, including researchers and policymakers interested in traffic safety.
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 will be distributed via the Open Data Toronto platform and potentially through a GitHub repository.
3. *When will the dataset be distributed?*
   - TBD
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 will be shared under an open license that permits reuse and redistribution, provided proper attribution is given to the source (e.g., City of Toronto).
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 known restrictions imposed by third parties on the data at this time
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.*
   - The dataset complies with local privacy laws and regulations regarding the handling of public data.
7. *Any other comments?*
   - TBD

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - The dataset will be supported and maintained by the City of Toronto's traffic management department.

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

   - Users can contact the traffic management department via email for support

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)?*

   - Yes, the dataset will be regularly updated to correct labeling errors, add new instances, or delete instances.

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.*

   - Yes, individuals were informed at the time of collection that their data would be retained for a fixed period before being deleted.

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.*

   - Yes, older versions will continue to be supported, and users will be informed of updates.

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.*

   - Yes, users can submit contributions to the dataset via GitHub, and all contributions will be validated.

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.