Datasheet for the Toronto Fire Incidents Analysis Dataset*

Supplementary document to the paper "Investigating Fire Incidents Severity Levels in Toronto: Enhancing Resource Allocation for Fire Response"

Maggie Zhang

2024-12-11

This supplementary file accompanies the paper, "Investigating Fire Incidents Severity Levels in Toronto: Enhancing Resource Allocation for Fire Response". Following the structure outlined by (Gebru et al. 2021), this datasheet describes the motivation, composition, collection process, preprocessing/cleaning/labeling, uses, distribution, and maintenance of the cleaned analysis dataset. The dataset is available in the Repository: TFS_Toronto_Fire_Incident_data_folder, under data folder, named: tfs_analysis_data. The original data was sourced from the Open Data Toronto portal (Gelfand 2022) and is titled Fire Incidents (Fire Services 2024).

Table of contents

*Code and data are available at: https://github.com/MaggieZ111119/TFS_Toronto_Fire_Incident.

1 Acknowledgement

The original dataset, titled Fire Incidents (Fire Services 2024), was obtained from the Open Data Toronto portal (Gelfand 2022). Maggie Zhang conducted the cleaning, analysis, and presentation as part of the Toronto Fire Incident Analysis Project. The processed dataset and related materials can be found in the repository TFS_Toronto_Fire_Incident. This datasheet follows the structure outlined by (Gebru et al. 2021), with the questions in this document derived from their published paper.

2 Motivation

1. **Purpose of Dataset Creation**: *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 the main factors influencing the severity of fire incidents in Toronto, while providing more in-depth insights into the current fire infrastructure system. Results are intended to better inform Toronto Fire Services (TFS) future decision-making and raise public awareness of fire incidents.

2. **Creators**: *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

    - The original dataset was provided through the Open Data Toronto portal (Gelfand 2022) by the Ontario Fire Marshal (OFM) and Toronto Fire Services (TFS). Data preprocessing and analysis were conducted by Maggie Zhang, a third-year undergraduate student at the University of Toronto.

3. **Funding**: *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.*

    - No external funding was used for this analysis.

4. **Additional Comments**: *Any other comments?*

    - The dataset reflects information available as of the latest update from the Ontario Fire Marshal (OFM). The original dataset will be updated annually, currently consisting of 32,929 rows and 43 columns of variables, providing information on fire incidents responded to by Toronto Fire Services (TFS). The analysis dataset for this study is based on the updated data from 2024.

3 Composition

1. **Instances Represented**: *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 fire incident that Toronto Fire Services responded to.

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

    - The analysis dataset contains 14,623 unique fire incidents.

3. **Dataset Type**: *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 can be considered complete with respect to the reported fire incidents during the collection period from 2011 to 2023, aside from some incidents removed for privacy protection purposes (Fire Services 2024). Therefore, it is representative of the larger population of fire incidents.

- Personal data have been removed, and exact addresses are aggregated, following the Municipal Freedom of Information and Protection of Privacy Act (MFIPPA).
- There are incidents that are still under investigation, thus some data points are missing (NAs).

4. **Features**: *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

    - The raw dataset consists of 32,929 rows and 43 columns of variables, providing information on fire incidents responded to by Toronto Fire Services (TFS).
    - In the cleaned analysis dataset for this study, there are 14,623 rows and 15 features. Each feature's name and description are listed below, extracted from the Fire Incident Data Portal (Fire Services 2024).
    - id : Unique row identifier for Open Data database
    - Area_of_Origin: OFM Area of Origin code and description
    - Civilian_Casualties: Count of civilian casualties. A causality can be a fire related injury or fire related fatality.
    - Count_of_Persons_Rescued: Number of persons rescued
    - Estimated_Dollar_Loss: Estimated Dollar Loss

- Fire_Alarm_System_Operation: OFM Fire Alarm System Operation code and description
- Ignition_Source: OFM Ignition Source code and description
- Initial_CAD_Event_Type: First event type in CAD system of this incident
- Number_of_responding_apparatus: Number of TFS responding apparatus
- Number_of_responding_personnel: Number of TFS responding personnel
- Possible_Cause: OFM Possible Cause code and description
- Sprinkler_System_Presence: OFM Sprinkler System Presence code and description
- TFS_Alarm_Time: Timestamp of when TFS was notified of the incident
- TFS_Arrival_Time: Timestamp of first arriving unit to incident
- TFS_Firefighter_Casualties: Count of TFS casualties. A causality can be a fire related injury or fire related fatality.
- Response_Time (Created Variables): time difference (in minutes) between TFS_Alarm_Time and TFS_Arrival_Time
- Visualizations of all features are contained in the paper, Appendix A section.

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

   - Fire severity is the target variable; it was categorized based on three key factors: the number of responding fire apparatus, the estimated dollar loss, and the total number of casualties.

6. **Missing Information**: *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 in the analysis fire incident dataset.
   - The original dataset includes all fire incidents as defined by the Ontario Fire Marshal (OFM) from 2011 to December 31, 2023.

7. **Relationships**: *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.*

   - `number_of_responding_personnel` and `number_of_responding_apparatus` are highly correlated. This can be explained by the fact that personnel usually carry a fixed number of apparatus.
   - `response_time` is related with `tfs_arrival_time` and `tfs_alarm_time` by the way its was defined

8. **Data Splits**: *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 has been split into training (80%) and testing (20%) subsets for analysis.

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

    - There are no redundancies in the dataset; each instance corresponds to a unique fire incident.
    - All variables in the dataset are categorical. Errors may exist in measurements and data entry, such as locations or ignition causes.
    - Bias may arise based on the category settings, as discussed in the Paper Discussion Section.

10. **External Resources**: *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 does not rely on external resources.

11. **Confidentiality**: *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 original dataset removes all data that may trigger privacy issues.

12. **Safty**: *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - Yes. There is a record of casualties and fire situations. People with trauma or fear of such incidents might feel anxious.

13. **Sub-Population**: *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.*

    - No. There is no information linked to specific groups of people in the analysis data, as the focus of the study is to understand the city of Toronto as a whole.

14. **Identifying Individuals**: *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

15. **Sensitive Data**: *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.

16. *Any other comments?*

    - None.

4 Collection Process

1. **Data Acquisition**: *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 was obtained directly from the Open Data Toronto portal. The original data are directly from TFS reports to the OFM.

2. **Collection Mechanisms**: *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 were collected through reports by fire departments and digitized for public use.

3. **Ethical Review**: *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 ethical concerns were identified since the dataset contains no personal information and no specific geographical information.

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

    - Data are from Open Data Toronto (Gelfand 2022), a publicly available resource.

5. **Consent**: *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.*

    - Not applicable, as the data are from reports.

5 Preprocessing, Cleaning, Labelling

1. **Steps Taken**: *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.*

    - Store raw dataset in its original csv format and later analysis data after cleaning and processing is saved into parquet file using `write_parquet` function in `arrow` package (Wickham and Hester 2024).
    - Columns of interest were selected, as explained in the Data section and Appendix A of the Paper. The column names were then standardized.
    - Rows with "Undetermined" ignition sources were removed, as they would not provide useful information for predicting severity.
    - Time-related variables were converted to the POSIX format.
    - All columns with minimal information were checked, and columns with more than 20% missing data were removed, considering their relevance to the study. More details on relevance decisions can be found in Appendix A of the Paper.
    - Rows with more than 30% missing data were removed.
    - Create new variable, Response_Time

2. **Raw Data**: *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 original raw dataset is archived for reference. It is stored in the [01 Raw Data] (https://github.com/MaggieZ111119/TFS_Toronto_Fire_Incident/tree/main/data/01-raw_data).

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

- Preprocess/clean/label was conducted using R (R Core Team (2023)). Packages include: The `tidyverse`(Wickham et al. 2023b), `dplyr`(Wickham et al. 2023a), and `tidyr`(Wickham, Bryan, et al. 2023) were used for data manipulation. `ggplot2`(Wickham 2016) facilitated data visualization, while `lubridate`(Grolemund and Wickham 2011) was used for date-time handling. `knitr`(Xie 2023) helped integrate code into reports.

4. *Any other comments?*

   - No

6 Uses

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

   - The original dataset is stored on an open data source; therefore, other studies can be conducted accordingly.

2. **Limitations**: *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?*

   - No. However, the current analysis dataset is derived from incomplete fire incident records, some of which are still under investigation..

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

   - No, except that the dataset should not be used for commercial purposes without proper attribution.

4. *Any other comments?*

   - None.

7 Distribution

1. **Third Party**: *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.
2. **Distribution Method**: *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 shared through a public GitHub repository TFS_Toronto_Fire_Incident_data_f in its `data` folder.
3. **Date**: *When will the dataset be distributed?*
    - December 15th, 2024.
4. **License**: *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 distributed under the MIT License.
5. **Restrictions**: *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
6. **Export Controls**: *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.
7. *Any other comments?*
    - None.

8 Maintenance

1. **Responsible Party**: *Who will be supporting/hosting/maintaining the dataset?*
    - The dataset will be hosted in Maggie Zhang's GitHub repository: TFS_Toronto_Fire_Incident_data repository. Maggie will be responsible for maintaining the dataset.

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

    - To contact Maggie, email: maggie.zhang@mail.utoronto.ca.
    - To access the original dataset, visit the Open Data Toronto Portal: (Fire Services 2024).

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

    - No.

4. **Updates**: *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)?*

    - No further updates to the analysis dataset for this study are planned after the release date.
    - The original dataset on Open Data Toronto is refreshed annually.

5. **Collaborations**: *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.*

    - Collaboration inquiries are welcome via email to Maggie. If they want to build on or contribute to the dataset, they can fork the public GitHub repository: TFS_Toronto_Fire_Incident_data_folder
    - If they are interested in the dataset and would like to start from the original data, they can access it through the Open Data Toronto Portal at Fire Incident Data (Fire Services 2024).

6. *Any other comments?*

    - None.

References

Fire Services. 2024. "Fire Incidents." https://open.toronto.ca/dataset/fire-incidents/.

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.

Gelfand, Sharla. 2022. "Opendatatoronto: Access the City of Toronto Open Data Portal." https://CRAN.R-project.org/package=opendatatoronto.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with Lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://CRAN.R-project.org/package=lubridate.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Jennifer Bryan, et al. 2023. "Tidyr: Tidy Messy Data." https://CRAN.R-project.org/package=tidyr.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023a. "Dplyr: A Grammar of Data Manipulation." https://CRAN.R-project.org/package=dplyr.

———. 2023b. "The Tidyverse." https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, and Jim Hester. 2024. "Arrow: R Interface to Apache Arrow." https://cran.r-project.org/package=arrow.

Xie, Yihui. 2023. "Knitr: A General-Purpose Package for Dynamic Report Generation in r." https://CRAN.R-project.org/package=knitr.