Datasheet for 'Donor Preferences in Canadian Political Financing'*

Understanding Donor Support for Ruling vs. Opposition Parties Between 2013 - 2024

Maria Mangru

December 3, 2024

This study investigates how the ruling political parties in Canada's federal and Ontario provincial governments affect the donation behaviors of individual contributors. By analyzing donation records from 2013 to 2024, the research explores whether a party being in power influences the total amount of financial support it receives. The findings reveal that opposition parties often receive more donations than those in power, highlighting donor preferences to support challengers. This insight enhances our understanding of political financing and can help parties and policymakers develop more effective fundraising strategies.

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 enable analysis of Australian politicians. We were unable to find a publicly available dataset in a structured format that had the biographical and political information on Australian politicians that was needed for modelling.

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

   - The Investigative Journalism Foundation created the dataset, they are an independent organization.

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.*

*Code and data are available at: https://github.com/MariaMangru/Ontario_Political_Donors.

- The IJF is funded by several foundations: Mitacs, Social Sciences and Humanities Research Council, 3858278 Canada Foundation, Canadian Internet Registration Authority, Data Driven Reporting Project, Avanti Foundation, Balsillie Family Foundation, C4C Canada, Data Science For Social Good, Toronto Metropolitan University and Individuals.

4. *Any other comments?*

    - NA

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.*

    - The instances that comprise the dataset represent political donations made by individuals or entities to political parties or entities in Canada.

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

    - There are 9,269,201 instances in total.

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 does not contain all possible instances since it only covers from 1993 to the present for federal data, but British Columbia's data only spans from 2005 onwards. The larger set would be all public data that references donations both federally and provincially. The sample is not representative of the larger set as it only covers a very short period of time since record availability and disclosure practices have withheld lots of public data.

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.*

    - The data comprising each instance is an processed json file with features, these features include region, date added, electoral event, political entity, political party, recipient, donor location, donation date, donation year, donor full name, donor type, amount and the monetary/non monetary amount as part of the total amount.

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

    - There is no label or target associated with each instance.

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 information due to data conversion (using OCR technology to convert pdf records into text), older recordes or the laws in different jurisdictions allowing corporate donations or not.

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 not explicitly defined but can be inferred.

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 data does not come with recommended data splits.

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

    - Yes, inconsistent naming conventions, data entry mistakes and OCR errors (which is the process of converting pdf to text) can occur.

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 does rely on public federal and provincial datasets to exist, but doesn't explicitly link to them. There are guarantees it will exist since it is compiled from publicly accessible records. There are archival version of the dataset that the IJF maintains. Also, the government of Canada and other provinces have archival versions as well. There are restrictions, which are fees to be able to access this data from teh website.

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 is comprised of public records and information of political donations which are legally required to be disclosed and accessible to the public.

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 it does not contains offensive, insulting, threatening or otherwise anxiety-inducing information.

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 explicitly identify sub-populations for example specific regions and political parties and corporations which could be considered subpopulations. [X]

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.*

    - Yes it is possible to identify individuals directly from the dataset by reading their full name from the row. All donations that exceed a certain threshold must be identifiable.

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 political opinions can be inferred from the donation records as it shows who donated to which party.

16. *Any other comments?*

    - NA

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 from official records that are maintained by election agencies at municipal, territorial, provincial and federal levels. All of these records are publicly accessible and include the detailed information on political donations.

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 mechanisms used to collect the data were automated retrieval, manual retrieval with OCR (Optical Character Recognition). These mechanisms were validated using manual verifications to correct any errors.

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 sample strategy was to fetch as much information as legally possible.

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 team was done by the IJF team which are staff membesr and potentially contractors. Compensation for this was not publicly disclosed.

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 timeframe was 1993 to 2024, the data is updated constantly as soon as its released publicly, so it does match the creation timeframe.

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.*

    - Any ethical review processes are not explicitly mentioned since this is scraped public data from different levels of government in Canada.

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 by making API calls to their server, therefore it was obtained by third parties.

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 individuals were not notified of the data collection.

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 individuals did consent to the collection and use of their data, since it is publicly disclosed data mandated by the law. Moreover, for the IJF's use of this publicly available data, additional consent is not required for this dataset.

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).*

    - NA

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 analysis has been done.

12. *Any other comments?*

    - NA

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, filtered the data for Federal and Ontario, then ensured the columns are the right datatype, records with incomplete information with key variables were excluded. Names were sometimes combined into one a single name for consistency, binary variables were also introduced and others for normalization and logarithmic transformations.

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 can be accessed at: https://github.com/MariaMangru/Ontario_Political_Donors/b 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.*

- The software was preprocessed and cleaned with RStudio https://posit.co/products/open-source/rstudio/

4. *Any other comments?*
    - NA

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - The dataset has been used for the current study, which investigated the impact of party leadership on donation patterns.
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.*
    - This repository https://github.com/MariaMangru/Ontario_Political_Donors/tree/main is the only one that links to a paper that uses the dataset.
3. *What (other) tasks could the dataset be used for?*
    - The dataset could be used for policy analysis and data journalism such as creating stories to highlight donation patterns or analyzing political financing trends.
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?*
    - A dataset consumer could conceal the names of the donors which donated large sums of money to political parties as it shows their full name or corporation name on the dataset.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - The dataset should not be used to harass or target individual donors.
6. *Any other comments?*
    - NA

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 is distributed to third parties through its website and public asccess to political donation records.

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 mainly accessible by their website where users can search and view records. There is currently no digital object identifier on the dataset.

3. *When will the dataset be distributed?*

   - The dataset is already distributed as it is publicly available and updated with the most recent data.

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 full dataset is behind a paywall, but other than that since it is all publicly available data it will stay publicly availabe with no intellectual property license on it.

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 third parties have imposed IP-based restrictions since it is public available government data that they use.

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.*

   - Yes there are, you cannot download or export the dataset like you would on other sites, you can only view it through the table on the search page.

7. *Any other comments?*

   - NA

Maintenance

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

   - THe IJF maintains, supports and hosts the political donations dataset.

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

- They can be contacted through their contact page on their website which is a form for communication. Email addresses are not publicly listed.

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

   - There is no mention of an erratum.

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)?*

   - The dataset is constantly updated to reflect new political donations, the IJF does not describe how often they do this.

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 includes information on political donations which contains donor names and details for large enough sums. This information is publicly accessible from the government and maintined by election agencies which have their own respective data retention policies.

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 IJF does not mention whether older version of the dataset will continue to be supported.

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.*

   - The IJF does not allow any mechanisms for external contributions to the dataset.

8. *Any other comments?*

   - NA

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.