Datasheet for Predicting Probability of Participants' Vote of Conservative and Liberty in 2025 Canadian election*

Jingyi Shen

April 19, 2024

In our study, we use binary logistic regression to analyze voter preferences for Canada's Conservative and Liberal parties based on demographic and socioeconomic factors, adjusting the models with the AIC for optimal predictor selection. The Canadian electoral system elects parliamentarians through a multi-step process where, unless a candidate achieves an outright majority, votes are redistributed from the lowest-ranking candidates until one secures 50%. The Liberal Party plans to increase its refugee quota to 15% of all immigrants and admit 500,000 migrants annually by 2025. The Conservative Party emphasizes economic growth and attracting skilled immigrants. Our hypothesis predicts a likely Conservative victory in the 2025 elections, with the Liberals closely behind, based on trends and electoral mechanics. This analysis uses post-stratification to enhance the representativeness and accuracy of our predictions.

The questions of the data sheet is from Gebru et al. (2021)

0.1 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 for the purpose of analyzing voter preferences in the Canadian political landscape, specifically targeting the Conservative and Liberal parties. The objective was to forecast the election probabilities for these parties and assess the influence of various sociodemographic factors, such as age, gender, province, religion, income, and education, on the probability of an individual's vote preference through a binary logistic regression model.

*Code and data supporting this analysis are available at https://github.com/CSCmaster/Final-Project

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

    - The data were sourced from two primary datasets: the General Social Survey (GSS) and the Canadian Election Study (CES), each providing a wealth of information on societal trends, behaviors, and attributes. The GSS data were adjusted to align with the survey data from the CES, ensuring a coherent dataset for analysis. The GSS captures a broad demographic, while the CES is more targeted toward understanding voter behavior and attitudes within the Canadian electoral context . The combined use of these datasets was to ensure a comprehensive and accurate representation of the Canadian population and their voting behaviors.

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 General Social Survey (GSS) is a nationwide survey program that collects data on societal trends, behaviors, and attributes across the entire population of Canada, aiming to provide insights into the changing dynamics of communities. It's conducted by Statistics Canada, a government entity tasked with producing statistics to help better understand Canada, its population, resources, economy, society, and culture. It usually funded from government.

The Canadian Election Study (CES), on the other hand, focuses on voter behavior, attitudes, and the electoral process within the Canadian context. It's conducted by a consortium of political scientists from The University of Toronto, McGill University, and The University of British Columbia, and it's often funded through research grants that may include the Social Sciences and Humanities Research Council of Canada (SSHRC) .

4. *Any other comments?*

    - NA

1 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 individual respondents to the surveys conducted by the Canadian Election Study (CES) and the General Social Survey (GSS). These instances include demographic and socioeconomic information about the individuals, such as age, sex, province, education, religion, and pre-tax income. The CES collected data through a sample of 20,968 individuals from across

Canada, aiming for representation from various regions and demographic balances in terms of gender and age demographics within each region. The purpose of these surveys is to gather and analyze data that reflects the political behavior, preferences, and social characteristics of the Canadian population, particularly focusing on the likelihood of voting for specific political parties—either the Conservative or Liberal party in this context .

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

    - 1777

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 a sample of instances from a larger set rather than all possible instances. Specifically, the survey data from the Canadian Election Study (CES) includes responses from over 37,000 participants. It was collected using the Leger Opinion panel to ensure representation across various regions of Canada and aimed to maintain a balance in gender and age demographics within each area. The CES data is not a census but rather a purposive sample aiming to infer characteristics of the larger Canadian population based on responses from selected individuals, particularly focusing on voter behavior, attitudes, and the electoral process within the Canadian context.

The census data from the General Social Survey (GSS) is designed to collect data on societal trends, behaviors, and attributes across the entire population, aiming to provide a comprehensive snapshot of demographic, economic, and social characteristics of communities. This data represents a systematic collection of information from the entire population at a specific point in time, in this case, conducted on August 12, 2022.

The representativeness of the CES sample was validated through the use of quotas reflecting population proportions in different regions (e.g., 7% for Atlantic, 23% for Quebec, 38% for Ontario, and 32% for the West) and language preferences (e.g., aiming for 80% French-speaking and 20% English-speaking participants in Quebec). For the GSS, the sampling strategy was stratified at the level of provinces and census metropolitan areas (CMA), using probability sampling techniques to ensure a representative sample of non-institutionalized individuals aged 15 and above residing in Canada's ten provinces

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 Canadian Election Study 2021 dataset consists of structured and processed survey data. This includes responses to a wide range of questions about political opinions, voting behavior, demographic information, and attitudes on various societal and political issues.

The survey collected detailed demographic information such as citizenship status, year of birth, gender identity, province or territory of residence, education level, and postal code (though the postal code is not released publicly for privacy reasons). Additionally, participants were asked about their interest in politics, likelihood to vote in the federal election, preferred voting method, and comfort with voting in person during the COVID-19 pandemic.

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

    - Yes, there is a label or target associated with each instance in the dataset, specifically related to the respondents' voting preferences and behaviors in the 2021 Canadian federal election. The survey collected data on which party the respondents intended to vote for or had already voted for during the election. This information serves as the target variable for analyses related to voting behavior.

The responses to this question include the choice of major parties like the Liberal Party, Conservative Party, New Democratic Party (NDP), and others, including the Bloc Québécois for respondents in Quebec. Additionally, respondents had the option to specify another party if their choice was not listed among the standard options. This label allows researchers to analyze patterns in party preference across different demographic groups and to study factors influencing voter decisions in the election.

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.*

    - The dataset from the Canadian Election Study 2021 contains instances where specific information may be missing due to various quality checks and data cleaning criteria. Here are some reasons why information might be missing from individual instances:

    Incomplete Responses: Instances where participants did not complete all parts of the survey were removed, resulting in missing data for those cases. Duplicate Responses: Responses identified as duplicates, either by panel ID or by a combination of IP address and demographic information, were removed except for the first occurrence. Speeders: Responses completed unusually quickly were scrutinized and possibly removed, under the assumption that they may not have been thoughtfully completed. Straight-lining: Instances where respondents provided the same answer across many questions, suggesting a lack of engagement, were removed. Mismatch Issues: Discrepancies in year-of-birth or province of residence provided at different times in the survey led to exclusion, thus

missing such instances. Attention Checks: Responses failing built-in attention checks were considered low quality and removed. Postal Code-Province Mismatches: Instances where postal codes did not match the reported province were removed.

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.*

    - The dataset from the Canadian Election Study 2021 does not explicitly model relationships between individual instances like user ratings or social network links. Instead, it focuses on individual respondents and their survey responses concerning political preferences, demographic data, and opinions on various issues.

There is no indication of relationships such as interactions between users (e.g., social networking connections) or between items (e.g., movie ratings influencing one another). Each survey response is treated as an independent instance, primarily concerned with capturing the individual's political preferences, behaviors, and demographic information.

The survey's design does not incorporate relational data between respondents but rather analyzes each response in the context of broader demographic and electoral insights. The dataset is structured to facilitate analyses of trends and patterns across the surveyed population, rather than interactions between the respondents.

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 2021 Canadian Election Study does not explicitly recommend specific data splits such as training, development/validation, and testing splits for data analysis. This is typical for survey data, where the primary focus is on the representativeness and comprehensiveness of the data for statistical analysis rather than on model training where such splits would be more critical.

However, the dataset itself is structured to allow for complex statistical analysis and could potentially be divided by researchers according to their specific needs. For example, researchers might choose to split the data into subsets for training and testing models of electoral behavior or voter sentiment analysis. This would depend on the research goals and the methodologies employed, such as predictive modeling or causal analysis.

The representativeness of the dataset is maintained through survey weights, which adjust for potential biases in the sample relative to the general population. These weights ensure that the findings from any analyses can be generalized back to the broader Canadian population, which would be a critical consideration in any split of the data for analytical purposes.

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

- The 2021 Canadian Election Study dataset includes some sources of noise and potential errors which are generally inherent to large-scale survey data. Here are some identified issues:

Incomplete Responses: Some responses were incomplete, possibly due to respondents starting but not finishing the survey.

Duplicate Responses: The dataset included duplicates where a single respondent may have taken the survey more than once. These duplicates were identified by matching panel IDs or a combination of IP address and demographic information, and only the first occurrence was retained.

Speeders: Some responses were removed because they were completed too quickly to be deemed reliable, indicating that the respondent may not have been fully engaged or thoughtful in their responses.

Straight-lining: This refers to respondents who chose the same response for a series of questions, which might indicate a lack of attention or engagement with the survey content.

Mismatch Issues: There were instances of mismatches in the data, such as discrepancies in the year of birth or province of residence provided at different times in the survey. These responses were typically removed to maintain data integrity.

Attention Checks: Responses that failed built-in attention checks were removed to ensure that only data from respondents who were paying attention and engaging with the survey questions were included.

Postal Code-Province Mismatches: Instances where the postal codes did not match the reported provinces were also removed, as these could indicate errors in data entry or falsification of information.

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 2021 Canadian Election Study dataset is essentially self-contained and does not rely on external resources such as websites, tweets, or other datasets for its primary content. The data collected and processed are self-sufficient for the analysis and insights that the study aims to provide regarding political opinions, voting behavior, and demographic information of Canadian voters.

However, it's important to note that while the dataset itself is self-contained, any analysis or comparison might involve external datasets or resources, especially for researchers looking to extend their findings or validate them against other studies. For instance, researchers might compare this data with other election studies or use census data to weight and validate their sampling methods.

No specific mention was made in the dataset documentation about dependencies on external resources that could change over time. Therefore, there are no explicit guarantees about the existence or constancy of external resources within the context of this dataset. Additionally, there are no restrictions like licenses or fees mentioned that would apply to a dataset consumer regarding external resources since the dataset does not rely on such resources directly.

If a researcher were to use external resources to supplement their analysis, they would need to ensure access and evaluate any licenses or fees associated with those resources independently of the CES dataset.

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.*

    - The Canadian Election Study 2021 dataset does not contain any data that might be considered confidential in the traditional sense, such as data protected by legal privilege or doctor-patient confidentiality. However, it does contain sensitive information that requires careful handling, including detailed demographic data and personal political opinions.

To ensure the confidentiality and privacy of participants, any personally identifiable information (such as exact postal codes) is not included in the public release of the dataset. Instead, broader geographic identifiers such as province or territory are used. Also, respondents' specific answers to open-ended questions are handled in a way that prevents the identification of individuals.

Moreover, the dataset is designed with privacy in mind, such that all data use must adhere to ethical guidelines and standards. The survey also included consent documentation where participants were informed about the data usage and their rights.

In summary, while the dataset contains detailed and sensitive information, it is managed in a manner that respects the privacy and confidentiality of the survey participants, adhering to ethical standards and legal requirements for data protection.

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 2021 Canadian Election Study dataset does not contain data that might be directly offensive, insulting, threatening, or cause anxiety. However, it includes responses to sensitive topics such as attitudes towards immigrants and refugees,

indigenous resentment scale, and questions on group discrimination which might contain opinions or language that some may find uncomfortable or disagreeable.

Given the nature of the topics covered, especially those touching on political, social, and cultural issues, some of the views expressed in the responses could potentially be provocative or distressing to some individuals. This is inherent in any dataset collecting opinions on a wide range of societal issues. The survey, however, is structured to ensure that such data is handled with respect for privacy and ethical guidelines, aiming to provide a neutral and analytical perspective on public opinion and democratic engagement within Canada.

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, the 2021 Canadian Election Study identifies several sub-populations within its dataset, based on demographic variables such as age, gender, province of residence, and education. Here's how these subpopulations are identified and described within the dataset:

    Age: Participants were categorized into three age groups: 18-34, 35-54, and 55 or older. This categorization helps in analyzing voting behavior and political opinions across different age cohorts.

    Gender: The dataset includes gender identification, allowing responses from men, women, and non-binary individuals. This distinction enables analyses that may reveal differences in political views and voting behavior across gender lines.

    Province of Residence: Respondents are identified by their province or territory of residence. The dataset includes specific quotas for different regions to ensure representativeness, such as Atlantic Canada, Quebec, Ontario, and Western Canada, reflecting their distribution in the Canadian population.

    Education: Respondents are categorized based on the highest level of education completed, ranging from no schooling to professional degrees. This variable can be crucial in understanding how educational background influences political preferences and election participation.

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.*

    - The 2021 Canadian Election Study dataset is designed to avoid directly or indirectly identifying individuals. The survey contains demographic and opinion data, but the dataset released for public use is anonymized to ensure confidentiality. Key measures taken include:

Removal of Direct Identifiers: Direct identifiers such as names and exact addresses are not collected or included in the dataset.

Use of Broad Categories: Instead of specific data that could be traced back to an individual (like exact age or income), broad categories are often used. For example, age groups rather than specific ages, and income brackets instead of exact figures.

Postal Code Data: Only the first three characters of the postal code are retained in some versions of the dataset to prevent location-based identification, while still allowing for regional analysis. Even these are only provided upon request.

Data Aggregation: Data is aggregated in such a way as to ensure that no individual can be identified through a combination of variables.

Ethical Compliance: The study adheres to ethical guidelines that ensure participants' information is kept confidential and used solely for research purposes.

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.*

    - The 2021 Canadian Election Study dataset contains data that could be considered sensitive, as it includes demographic information such as age, gender, and province, as well as personal opinions on various political, social, and economic issues. Here are some types of sensitive data included:

    Political Opinions: Respondents provided information on their political affiliations, voting intentions, satisfaction with democracy, and opinions on political leaders and parties.

    Demographic Information: Information on respondents' age, gender identity, education level, and province of residence was collected. Such demographic data, especially when combined, can be considered sensitive because they relate to personal attributes.

    Socio-Economic Status: Questions related to personal finances and perceptions of economic issues, which can be sensitive as they pertain to an individual's financial situation.

    Ethnic and Cultural Background: The survey included questions about respondents' attitudes towards immigrants and refugees, which are sensitive topics that can reveal personal beliefs and biases.

    Religious Beliefs: Information about respondents' religious affiliations was also collected, which is generally considered sensitive due to its personal and private nature.

16. *Any other comments?*

    - NA # Collection Process

17. *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 Collection Method: The data was collected directly from participants through se`

    Validation Checks: Various validation checks were integrated into the survey design to ensure data quality. For instance, respondents were asked to re-confirm their year of birth at the end of the survey, and their responses were checked for consistency. Additionally, attention checks were included within the survey to ensure that respondents were paying attention and not randomly filling out the survey.

    Data Cleaning: Post-data collection, a rigorous data cleaning process was applied. This included the identification and removal of duplicate responses, speeders (those who completed the survey too quickly to have given thoughtful responses), and straight-liners (those who selected the same answer for a series of questions). These measures help ensure that only high-quality, reliable data was retained.

    Survey Weights: To make the data representative of the Canadian population, survey weights were calculated and applied. These weights were based on demographic information such as province, gender, age group, and education level, compared against the 2016 Canadian Census. This process helps adjust for any sampling biases and ensures the findings can be generalized to the broader population.

18. *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 data associated with each instance in the 2021 Canadian Election Study was acquired through an online survey platform, specifically Qualtrics. The survey was conducted in two main phases: the Campaign Period Survey (CPS) and the Post-Election Survey (PES). Participants were Canadian citizens and permanent residents aged 18 or older. Mechanisms and Procedures for Data Collection:

    Online Surveys: Respondents completed the surveys online using Qualtrics, which is a comprehensive and widely used tool for administering online surveys. This method is advantageous for reaching a diverse, national sample quickly and cost-effectively. Panel Provider: The sample was procured through the Leger Opinion panel, which helped ensure a representative sample stratified by region, balanced on gender, and age within each region.

Validation of Mechanisms and Procedures:

```
Pre-Testing: The survey instruments were likely pre-tested to check for clarity, understandi
Pilot Testing: A smaller scale deployment before the main survey could be used to test the o
Data Quality Controls: Various quality controls were in place, such as checks for duplicate
Attention Checks: To ensure that respondents were paying attention and engaging thoughtfully
```

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 sampling strategy used for the 2021 Canadian Election Study involved a probabilistic approach with specific sampling probabilities to ensure representativeness across various demographic dimensions.

Sampling Design: 1. **Two-Wave Panel Structure**: The study was structured as a two-wave panel, consisting of a campaign period survey (CPS) and a post-election survey (PES). 2. **Modified Rolling-Cross Section**: During the campaign period, a modified rolling-cross section method was used. This approach allows for continuous sampling of different subsets of the population throughout the campaign period.

Detailed Strategies: - **Leger Opinion Panel**: The sample was sourced through the Leger Opinion panel, targeting stratification by region and maintaining balance in gender and age within each region. - **Demographic Quotas**: The survey aimed for a representative distribution across: - **Gender**: An equal split between men and women. - **Age Groups**: Distribution among age groups was 28% for ages 18-34, 33% for ages 35-54, and 39% for ages 55 and older. - **Regions**: Set quotas based on population proportions for Atlantic Canada, Quebec, Ontario, and Western Canada, accepting any respondents from the Territories without specific quotas. - **Language**: In Quebec, a target of 80% French-speaking and 20% English-speaking participants was set.

Validation of Sampling Method: - **Survey Weights**: To adjust for any potential sampling biases and to ensure the dataset is representative of the Canadian population, survey weights were calculated using an iterative raking process. This weighting accounted for regional, gender, age, and educational differences based on the 2016 Canadian census. - **Oversampling and Daily Targets**: The study included an oversample wave to gather sufficient data from categories at risk of underrepresentation. Daily sampling targets were adjusted during the campaign to ensure a robust sample size, particularly during the CPS modules wave.

The rigorous sampling design and subsequent weighting ensure that the dataset provides a representative and reliable basis for analyzing Canadian public opinion during the 2021 federal election.

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 for the 2021 Canadian Election Study was managed by Leger Opinion, which provided the sample from their panel. The survey was conducted

using the Qualtrics platform, ensuring a systematic and standardized data collection process.

1.0.1 Who was involved in data collection:

- **Panel Provider (Leger Opinion)**: The data collection was primarily managed by Leger Opinion, which utilized its panel of Canadian residents to provide a diverse and representative sample of the population.
- **Survey Respondents**: Individuals from the Leger Opinion panel participated in the survey. These respondents are Canadian citizens or permanent residents aged 18 or older.

1.0.2 Compensation:

- **Respondents**: Typically, participants in such panels are compensated through a points system, where they earn points for each survey completed. These points can usually be redeemed for rewards like gift cards, cash, or entries into draws. Specific compensation details for the Leger Opinion panel participants were not provided in the document but would be aligned with standard practices for survey panel compensations.

1.0.3 Validation of the data collection procedures:

- **Survey Platform (Qualtrics)**: Using a well-established survey platform like Qualtrics helps ensure the validity of the data collection process. Qualtrics provides tools for survey design, distribution, and data integrity checks.
- **Sampling Strategy**: The sampling strategy involved stratification by region and demographic balance, ensuring that the data collected reflected the diversity of the Canadian population. The sampling method was validated by applying survey weights to ensure representativeness based on the 2016 Canadian Census.

This comprehensive approach to data collection ensures that the dataset is robust, representative, and reliable for analyzing public opinion during the Canadian federal election.

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 data for the 2021 Canadian Election Study was collected over two main periods: the Campaign Period Survey (CPS) from August 17, 2021, to September 19, 2021, and the Post-Election Survey (PES) from September 23, 2021, to October 4, 2021. These collection periods directly align with the creation timeframe of the data

associated with the instances, as they are tied to specific activities and events of the 2021 Canadian federal election campaign and its immediate aftermath.

This direct observation of the election period through surveys ensures that the data reflects contemporaneous public opinion, behaviors, and attitudes related to the electoral process. The surveys captured insights as they unfolded, giving an immediate and relevant snapshot of the electorate's views, rather than compiling retrospective accounts or historical data which might not accurately represent the sentiment at the time of the election.

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.*

    - The 2021 Canadian Election Study underwent ethical review processes to ensure that the rights, safety, and well-being of participants were protected throughout the research. The study involved gathering data from respondents through surveys, which were conducted online using the Qualtrics platform. As part of the ethical oversight:

- **Consent Documentation**: Participants were provided with a consent form, which outlined the study's purpose, the voluntary nature of their participation, and the confidentiality of their responses. They were required to agree to participate before proceeding with the survey.
- **Ethical Approval**: The research design and methods, including data collection and handling procedures, were reviewed and approved by an Institutional Review Board (IRB) or an equivalent ethics committee. This ensures that the study complies with ethical standards in research involving human participants.
- **Privacy and Confidentiality**: Measures were taken to ensure that respondents' identities were protected. Direct personal identifiers were not included in the dataset, and sensitive information was handled according to strict confidentiality protocols.
- **Contact Information for Ethics Oversight**: The consent form provided contact information for the Office of Human Research Ethics, where participants could reach out with any concerns about their rights or the conduct of the study.

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 used in the study was primarily collected directly from individuals via surveys. Specifically, the study utilized data from two sources:

    Canadian Election Study (CES): This data comes from a survey conducted among a large sample of participants across Canada. The CES focuses on voter behavior, attitudes, and the electoral process within the Canadian context.

General Social Survey (GSS): This survey collects data from the general population on various social trends, behaviors, and attributes. It provides a broad snapshot of the demographic, economic, and social characteristics of the population.

Both datasets were acquired through direct surveys rather than via third parties or other external sources. This direct method of data collection helps ensure the reliability and relevance of the information to the study's objectives.

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.*

    - Yes, the individuals participating in the 2021 Canadian Election Study were notified about the data collection process. The participants were provided with consent documentation at the beginning of the survey, outlining the purpose of the study, the voluntary nature of their participation, and the confidentiality of their responses.

Here is the exact language of the notification provided in the consent documentation:

Letter of Information and Consent

Project Title: Canadian Election Study 2021

Principal Investigator: Laura Stephenson, PhD, Political Science, University of Western Ontario

Co-Investigators: Allison Harell, PhD, Political Science, Université de Québec à Montréal; Peter Loewen, PhD, Political Science, Munk School of Global Affairs, University of Toronto; Daniel Rubenson, PhD, Politics and Public Administration, Ryerson University

"Thank you for considering taking our survey. You can complete the survey in English or in French. Please choose your preferred language using the drop-down menu at the top right of this screen. We are doing a research study about Canadians' attitudes about current events, elections and democracy, and we would like your opinions. Participation in the research is voluntary. Your answers will be kept completely confidential and will be used for research purposes only. You must be 18 years or older and a legal resident of Canada to participate."

"Today's survey will take about 20 minutes to complete. If you participate today you may be re-contacted at a later date and invited to complete additional surveys as well. Participation in any additional surveys will also be voluntary."

"To read additional information and details about this study, including your rights as a research participant, how you will be compensated for your time, and how your data will be stored, please click HERE. If you would like to complete the survey, please click the appropriate link below."

"If you have questions at any time about the study or its results, you may contact Dr. Laura Stephenson at the Department of Political Science, University of Western Ontario."

"If you have any questions about your rights as a research participant or the conduct of this study, you may contact The Office of Human Research Ethics."

This detailed notification ensured that participants were fully informed about the study and their participation rights before they consented to participate.

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.*

    - Yes, the individuals participating in the 2021 Canadian Election Study provided explicit consent for the collection and use of their data. The consent process was integrated into the online survey platform (Qualtrics), where participants were presented with a consent form at the beginning of the survey.

Here is the exact language used in the consent form:

Letter of Information and Consent

Project Title: Canadian Election Study 2021

Principal Investigator: Laura Stephenson, PhD, Political Science, University of Western Ontario
Co-Investigators: Allison Harell, PhD, Political Science, Université de Québec à Montréal; Peter Loewen, PhD, Political Science, Munk School of Global Affairs, University of Toronto; Daniel Rubenson, PhD, Politics and Public Administration, Ryerson University

"Thank you for considering taking our survey. You can complete the survey in English or in French. Please choose your preferred language using the drop-down menu at the top right of this screen. We are doing a research study about Canadians' attitudes about current events, elections, and democracy, and we would like your opinions. Participation in the research is voluntary. Your answers will be kept completely confidential and will be used for research purposes only. You must be 18 years or older and a legal resident of Canada to participate."

"Today's survey will take about 20 minutes to complete. If you participate today you may be re-contacted at a later date and invited to complete additional surveys as well. Participation in any additional surveys will also be voluntary."

"To read additional information and details about this study, including your rights as a research participant, how you will be compensated for your time, and how your data will be stored, please click HERE. If you would like to complete the survey, please click the appropriate link below."

"If you have questions at any time about the study or its results, you may contact Dr. Laura Stephenson at the Department of Political Science, University of Western Ontario."

"If you have any questions about your rights as a research participant or the conduct of this study, you may contact The Office of Human Research Ethics."

The consent form provided participants with comprehensive information about the study, ensuring they understood their participation was voluntary, how their data would be used, and their rights as research participants. Participants were required to actively consent by selecting an option that indicated their willingness to participate, before they could proceed with the survey. If they chose not to consent, they were not allowed to participate in the survey.

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).*

    - In the consent documentation of the 2021 Canadian Election Study, participants were informed about their rights as research participants, including how to contact the research team and the Office of Human Research Ethics if they had any concerns or questions at any point. However, the documentation provided does not explicitly mention a mechanism for participants to revoke their consent once it has been given or to withdraw from the study after participating.

Typically, in research studies, participants can contact the principal investigator or the research ethics office if they decide to withdraw their participation or have their data removed. The contact details provided in the consent form include:

- **Dr. Laura Stephenson**, Department of Political Science, University of Western Ontario, laura.stephenson@uwo.ca
- **Office of Human Research Ethics**, which can be contacted for any ethical concerns about the study (contact details not specified in the excerpt).

Participants were likely able to discuss concerns and potentially revoke consent through these contacts. However, for specifics on the process of revoking consent after data collection, participants would need to directly contact these individuals or offices. This approach ensures confidentiality and adherence to ethical standards in handling participants' data and 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.* - The documentation for the 2021 Canadian Election Study does not explicitly mention the conduct of a data protection impact analysis (DPIA) or similar assessment specifically focused on evaluating the impact of the dataset and its use on data subjects. Typically, ethical reviews and consent procedures in academic research studies like the Canadian Election Study address privacy concerns and ensure the protection of participant

data, but they might not follow the same detailed impact analysis process as seen in other sectors, especially under data protection regulations like the GDPR.

However, the study's consent documentation and ethical considerations indicate that efforts were made to protect participant data:

- **Confidentiality and Data Handling:** The consent form assures participants that their responses will be kept confidential and used solely for research purposes. Personal identifiers are not included in the dataset to prevent identification of participants.
- **Ethical Oversight:** The study was reviewed and approved by an ethics committee, which typically evaluates the potential risks and benefits of the research to participants. This includes assessing how data is collected, stored, and processed.
- **Consent Process:** Participants were clearly informed about the purpose of the research, what participation involves, and their rights, including how to contact the research team or the ethics office if they have concerns.

While this indicates that considerations regarding data protection and participant impact were integrated into the study's design and execution, there is no specific mention of a formal DPIA. For more detailed information or to confirm whether a DPIA was conducted, one would need to contact the study's principal investigator or the relevant ethics office. These details are typically not publicly disseminated in academic research documentation but are handled internally within the governance frameworks of the institutions conducting the research.

12. *Any other comments?*

    - NA # Preprocessing/cleaning/labeling

13. *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, preprocessing, cleaning, and labeling of the data were conducted as part of the 2021 Canadian Election Study. Here are some specific tasks performed:

14. **Removal Criteria**: Respondents were removed from the dataset for various reasons, including incomplete responses, duplicates (identified by IP address and demographic details or panel ID), speeders (those who completed the survey too quickly to have given thoughtful responses), and straight-liners (those who selected the same response for many questions). Additionally, those who failed attention checks or had mismatches in their year of birth or province of residence between different survey points were also removed.

15. **Data Quality Checks**: The dataset underwent several quality checks to ensure the reliability of the data. This included verifying the consistency of responses, especially for critical data like the year of birth and province of residence.

16. **Processing of Missing Values**: Responses with missing data were flagged, and in some cases, missing data was due to respondents skipping questions or opting not to answer.

17. **Variable Naming and Standardization**: Variables were named descriptively to ensure clarity and ease of use in analysis. For instance, variables related to demographic information, survey responses, and data quality were clearly labeled.

18. *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 documentation for the 2021 Canadian Election Study does not specifically mention whether the "raw" data is preserved separately from the preprocessed/cleaned/labeled data. Typically, in large-scale studies like this, raw data is stored securely to enable verification of findings or additional analysis in the future. However, the public release versions of the dataset usually contain cleaned and anonymized data to protect participants' privacy and ensure the integrity of the data for research use.

For specific details about access to raw data or its availability, it would be appropriate to contact the study administrators or the data custodian, typically the principal investigator or the institution hosting the data. In this case, the contact information provided in the codebook would be the point of entry for such inquiries. 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 dataset from the 2021 Canadian Election Study, as provided, typically includes preprocessed, cleaned, and labeled data suitable for public release and research purposes. The raw data, collected via the survey platform Qualtrics, often undergoes significant processing to ensure privacy, remove errors, and format the data for analysis, as detailed in the codebook.

Typically, the raw data is not publicly available due to privacy concerns and the sensitive nature of the information collected. The provided dataset has been cleaned of identifiers and sensitive information that could potentially compromise participant confidentiality. Raw data might be retained by the researchers for internal use under strict ethical guidelines but is not distributed or accessible to the public to protect respondents' privacy and comply with data protection regulations.

For access to any form of raw data or to inquire about specific data handling practices not covered in public documentation, direct contact with the principal investigators or the institution hosting the dataset would be necessary. This is often the case where detailed data management information is required, which might not be explicitly detailed in the publicly available codebook or dataset documentation. 4. *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.* The 2021 Canadian Election Study dataset is intended for broad academic use and public release as indicated by its public release version detailed in the codebook. The dataset is made available to researchers and entities outside of the original study

team under specific usage guidelines which require appropriate citation and crediting when used for further analysis or publication.

1.0.4 Distribution Policy:

- **Broad Access**: The dataset is distributed for use by other researchers and institutions, which means it can be accessed and used for academic purposes beyond the original study team.
- **Usage Guidelines**: Any use of the dataset must be accompanied by proper citation to give credit to the original creators of the study. The citation ensures that the original contributors receive acknowledgment for their work and that the data usage is tracked.

1.0.5 Example Citation:

As stated in the codebook, any use of the dataset should cite it as follows: "Stephenson, Laura B., Allison Harell, Daniel Rubenson, and Peter John Loewen. The 2021 Canadian Election Study. [dataset]"

This citation requirement is a common academic practice designed to maintain the intellectual integrity and traceability of the data usage.

1.0.6 Access Point:

- While a direct link to the dataset is not provided in the provided excerpts, the data is typically accessible through academic data repositories or directly from the study's official website where researchers can request access.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?* The 2021 Canadian Election Study data is distributed in .dta format, which can be utilized with a variety of software programs such as STATA, SPSS, and R. This indicates that the dataset is provided as downloadable files, typically through an academic or research institution's repository or the official study website.

Regarding the digital object identifier (DOI), there is no specific mention of a DOI for the dataset in the provided documents. A DOI is a persistent identifier used to uniquely identify electronic documents, and it's not always applicable or provided for every academic dataset. If a DOI is necessary for your purposes, you might need to check the repository or website where the dataset is hosted or reach out to the principal investigators for more detailed information on accessing and citing the dataset. 3. *When will the dataset be distributed?* The documentation provided does not specify the exact distribution date for the 2021 Canadian Election Study

dataset. Typically, such datasets are released after the data collection and cleaning processes are complete and the initial analyses have been conducted by the primary research team.

If you need specific details regarding the release date or availability of the dataset, it would be best to contact the study administrators or the principal investigators directly. They would be able to provide the most accurate and up-to-date information on when the dataset will be made available for public use or academic research. 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 2021 Canadian Election Study data is distributed under a license that requires appropriate referencing and crediting when used. The dataset is intended for academic and research purposes, and any use of the data must include a proper citation. The citation for using the dataset is as follows:

Citation: Stephenson, Laura B., Allison Harell, Daniel Rubenson, and Peter John Loewen. The 2021 Canadian Election Study. [dataset]

No specific fees are associated with these restrictions, which makes the data broadly accessible for academic and research use. The focus is on ensuring that the original researchers receive proper acknowledgment for their work, and that the data's integrity is maintained in subsequent uses.

For accessing the dataset, typically, it would be available through academic data repositories or through the official website of the study where interested parties can request access. There is no mention of a digital object identifier (DOI) in the documentation provided, which suggests that accessing the dataset would be through these conventional academic channels.

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.* The 2021 Canadian Election Study dataset does not specify any intellectual property (IP)-based or other restrictions imposed by third parties on the data associated with the instances. The documentation indicates that all use of the dataset must be appropriately referenced and credited, with a specific citation provided for any publications or presentations that utilize the data:

Citation: Stephenson, Laura B., Allison Harell, Daniel Rubenson, and Peter John Loewen. The 2021 Canadian Election Study. [dataset]

There are no explicit IP restrictions or fees associated with the use of the data mentioned in the documentation. The dataset is intended for academic and research purposes, and its use is conditioned upon proper citation to acknowledge the contributions of the original researchers. This approach facilitates the ethical and responsible use of the data within the academic community. If there were any third-party restrictions, these would typically be detailed in the license or terms of use associated with the dataset, none of which are specified in the provided

documents. 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 documentation for the 2021 Canadian Election Study does not explicitly state any export controls or regulatory restrictions applied to the dataset or to individual instances. The data collected through this survey typically adheres to local laws and regulations concerning data privacy, research ethics, and data protection, as governed by the institution hosting the research and the jurisdictions in which the data is collected and used.

The lack of mention of specific regulatory restrictions or export controls suggests that the primary considerations for using the dataset revolve around ethical research practices, participant consent, and privacy protections, which are standard for academic research involving human subjects.

For more specific details regarding any potential regulatory restrictions or to confirm the absence of such, direct communication with the research team or the ethics board that reviewed the study would be necessary. They would be able to provide detailed guidance on any restrictions or compliance requirements that might not be explicitly documented in the public release of the dataset. 7. *Any other comments?* NA # Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?* The entity supporting, hosting, and maintaining the 2021 Canadian Election Study dataset is typically the academic or research institution affiliated with the principal investigators and research team responsible for the study. This institution serves as the custodian of the dataset, ensuring its accessibility, integrity, and long-term preservation for academic and research use.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?* Contact information for the owner, curator, or manager of the 2021 Canadian Election Study dataset is typically provided through the dataset documentation or associated publications. In academic research, the principal investigators or project leads are usually the primary points of contact for inquiries related to the dataset.

Based on the information provided in the documentation, you can contact the principal investigators of the study for further assistance or inquiries regarding the dataset. Here are the contact details for the principal investigators:

1. **Dr. Laura Stephenson**
    - Department of Political Science, University of Western Ontario
    - Email: laura.stephenson@uwo.ca
2. **Dr. Allison Harell**
    - Department of Political Science, Université de Québec à Montréal
3. **Dr. Daniel Rubenson**

- Department of Politics and Public Administration, Ryerson University

4. **Dr. Peter John Loewen**
    - Munk School of Global Affairs, University of Toronto

You can reach out to them via email with any questions or inquiries you may have regarding the dataset, its usage, availability, or any other related matters. They should be able to provide further guidance or direct you to the appropriate resources for assistance.

3. *Is there an erratum? If so, please provide a link or other access point.* There is no specific mention of an erratum related to the 2021 Canadian Election Study dataset in the provided documentation. An erratum typically refers to a correction or update issued after the initial publication of a dataset or associated documentation to rectify errors or provide additional information.

If you suspect that there may be errors or updates related to the dataset, it is advisable to review any accompanying documentation, such as codebooks or data release notes, for any relevant announcements or corrections. Additionally, contacting the principal investigators or project leads of the study directly can provide clarification or updates on any issues related to the dataset. They would be able to inform you of any errata or corrections that have been issued for the dataset and provide guidance on how to access them.

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 provided documentation does not specify whether updates will be made to the 2021 Canadian Election Study dataset. However, it is common practice in academic research to periodically update datasets to correct errors, add new instances, or improve data quality based on feedback or new research findings.

If updates are planned for the dataset, the frequency, process, and communication methods for these updates would typically be determined by the dataset's custodians or the research team responsible for its maintenance. Here are some general considerations:

1. **Frequency**: Updates may occur on an as-needed basis or periodically, depending on the nature and extent of changes required. For example, updates might be released annually, after each election cycle, or in response to identified errors or improvements.
2. **Responsibility**: The responsibility for updating the dataset lies with the custodians or managers of the dataset, typically the principal investigators or the institution hosting the research.

3. **Communication**: Updates to the dataset can be communicated to dataset consumers through various channels, such as email notifications, announcements on the dataset's website or repository, mailing lists, or social media platforms. Academic repositories like GitHub or institutional data repositories may also provide versioning and change logs to track updates and revisions.

4. *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 documentation for the 2021 Canadian Election Study dataset does not explicitly mention any limits on the retention of data associated with the instances. However, in academic research involving human subjects, data retention practices typically adhere to ethical guidelines and data protection regulations to ensure the privacy and confidentiality of participants.

Common practices for data retention in academic research include:

1. **Informed Consent**: Participants are typically informed about how their data will be used and retained during the consent process. If specific limits on data retention are in place, participants would be informed accordingly.

2. **Data Protection Regulations**: Data retention practices may be guided by local data protection regulations, which often specify retention periods or conditions for data storage and deletion.

3. **Ethical Oversight**: Academic research studies, especially those involving human subjects, are often subject to ethical review by institutional review boards (IRBs) or ethics committees. These bodies may provide guidance or set requirements for data retention and deletion.

4. **Research Protocols**: The research team may establish protocols for data management, including data retention and deletion procedures, to ensure compliance with ethical standards and regulatory requirements.

5. *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 documentation does not specify whether older versions of the 2021 Canadian Election Study dataset will continue to be supported, hosted, and maintained. However, in academic research, it is common for datasets to be preserved and maintained for long-term access, especially if they are considered valuable for ongoing research or historical reference.

Here are possible scenarios for how older versions of the dataset might be handled:

1. **Continued Hosting and Maintenance**: Older versions of the dataset may continue to be hosted and maintained alongside newer versions, ensuring ongoing access to historical data. This allows researchers to access and compare data from different time periods for longitudinal studies or trend analysis.

2. **Archiving**: If older versions of the dataset are no longer actively supported or maintained, they may be archived in repositories or data archives to ensure long-term preservation. Archived versions would remain accessible for reference purposes but may not receive updates or technical support.

3. **Communication of Obsolescence**: If older versions of the dataset are no longer supported or maintained, this information would typically be communicated to dataset consumers through announcements on the dataset's website, repository, or other communication channels. Consumers may be advised to migrate to newer versions of the dataset or alternative data sources for updated information.

4. *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 documentation does not explicitly mention a mechanism for others to extend, augment, build on, or contribute to the 2021 Canadian Election Study dataset. However, academic research datasets are often considered valuable resources for further analysis and research, and contributions from other researchers may be welcomed under certain conditions.

Here are potential considerations regarding contributions to the dataset:

1. **Mechanism for Contributions**: If contributions to the dataset are welcomed, researchers may be invited to submit their additions or modifications through a designated process, such as contacting the dataset's custodians or submitting proposed changes through a repository's contribution mechanism.

2. **Validation/Verification**: Contributions to the dataset may undergo validation and verification processes to ensure data quality, consistency, and adherence to ethical standards. This could involve peer review by experts in the field or validation against existing data sources.

3. **Communication/Distribution of Contributions**: If contributions are accepted and validated, they may be communicated to dataset consumers through updates to the dataset documentation, release notes, or versioning information. Contributions may be distributed alongside the original dataset, either as part of updated versions or as supplementary datasets.

4. **Process for Reviewing Contributions**: A structured process may be established for reviewing and approving contributions to the dataset. This process could involve criteria for assessing the relevance, accuracy, and ethical compliance of proposed contributions.

5. *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.