

2024 USA Presidential Election Prediction*

Ruiyang Wang

November 3, 2024

This PDF discusses the 2024 prediction dataset, including motivations, composition, Preprocessing/cleaning/labeling, Uses, Distribution, and Maintenance.

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
 - The dataset was created to analyze polling data from the 2024 U.S. presidential election with the specific aim of predicting the likely winner and understanding how factors like sample size, poll score, and transparency influence candidate support.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
 - The dataset was created by me, using data sourced from FiveThirtyEight.
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.*
 - None.
4. *Any other comments?*
 - No.

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.*

*Code and data are available at: [<https://github.com/Ruiyang-Wang/STA304-Paper2.git>].

- Each instance in the dataset represents an individual poll providing data on candidate support rates, with variables such as sample size, poll scores, and transparency scores to understand how these factors influence public support.
2. *How many instances are there in total (of each type, if appropriate)?*
 - 9
 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 is a sample drawn from a larger set of polling data available publicly (e.g., from FiveThirtyEight). While representativeness is implied, specific methods for validating this are not detailed.
 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 consists of both raw data representing polling percentages and processed data in “analysis data” that includes features like log-transformed sample sizes to improve analysis accuracy
 5. *Is there a label or target associated with each instance? If so, please provide a description.*
 - Yes, there is a target variable associated with each instance, specifically the percentage of support for each candidate in a given poll.
 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.*
 - No
 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 between poll score, transparency, sample size, and support rates are explicitly recorded in the cleaned data to explore their effect on candidate support.
 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.*

- No
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
- Potential sources of noise in the data include low R-squared values and non-constant variance issues in the cleaned “analysis data” for certain candidates.
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 is self contained, have no other external links or other datasets.
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.
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.
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.*
- Sub-populations are segmented by polling data on support levels for different candidates and the quality metrics of polls
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
- No
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.*

- No
16. *Any other comments?*
- None

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 for each instance was acquired directly from publicly available polling sources form FiveThirtyEight, with the raw data and cleaned dataset used for statistical analysis.
2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
 - Data collection and analysis were conducted using R.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
 - The dataset represents a sample of U.S. election polling data, likely selected to reflect the general voting population.
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 paper does not specify who was involved in the data collection process or if they were compensated.
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 for the data covers polling from the 2024 election period, consistent with the analysis’s objective of understanding current trends.
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

- No.
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 dataset was compiled from polling data obtained from third-party sources rather than directly from individuals.
 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.*
 - Since the dataset consists of aggregated polling data, there is no mention of notifying individuals about 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.*
 - Consent procedures are not applicable, as the dataset consists of publicly available, aggregated data.
 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).*
 - No
 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
 12. *Any other comments?*
 - None

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, the data underwent preprocessing and cleaning. This process involved removing irrelevant instances, processing missing values, and applying transformations such as logarithmic transformations of sample sizes to ensure model assumptions were met. This cleaned data is saved as “analysis data,” derived from the original “president poll” dataset.
2. *Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.*
 - The raw data, “president poll,” is saved separately from the cleaned “analysis data” to allow for unanticipated future uses or verification needs.
 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 preprocessing and analysis were performed using R, and the code is available on GitHub at <https://github.com/Ruiyang-Wang/STA304-Paper2>
 4. *Any other comments?*
 - None

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
 - No.
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.*
 - The code is available on GitHub at <https://github.com/Ruiyang-Wang/STA304-Paper2>
3. *What (other) tasks could the dataset be used for?*
 - Beyond election predictions, the dataset could be used for research into polling methodologies, bias analysis, and the impact of polling transparency and sample size on public opinion.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

- the dataset's composition and preprocessing may impact future uses, especially as it incorporates transformations and selection criteria specific to 2024 election polls.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
 - No
 6. *Any other comments?*
 - None

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 dataset may be distributed to third parties
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
 - By Github, no DOI.
3. *When will the dataset be distributed?*
 - The dataset is available immediately, as it is part of an GitHub project.
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.*
 - No
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
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
 - No
7. *Any other comments?*

- None

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
 - Ruiyang Wang
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
 - Contact through github issue
3. *Is there an erratum? If so, please provide a link or other access point.*
 - No
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
 - Yes, if future updates is required.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
 - No
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.*
 - No, since we only submit one version fo dataset. There might be more versions if there is future updates on the dataset
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
 - Contributions to the dataset may be possible via GitHub
8. *Any other comments?*
 - None