India Electoral Rolls Project (2024-25): Documentation (including Codebook)

Sharik Laliwala (University of California, Berkeley)*

Version 1.0: 18 August 2025

Project Overview

This project parses India's publicly available electoral (voter) rolls into a readily analysable dataset. Although the Election Commission of India (ECI) publishes these rolls publicly, they are stored in over a million PDF files, which hinders at-scale analysis. This project programmatically downloads multilingual electoral rolls and applies OCR using a state-of-the-art model with over 95% accuracy in English, Hindi, and Gujarati. The resulting datasets support research on population and demographic trends (e.g., migration), inter alia.

I have released a parsed dataset for the state of Haryana's electoral rolls prepared for the 2024 Vidhan Sabha election. I will periodically update this dataset to address errors and inconsistencies. I aim to release one state's dataset every month, depending on resource availability.

I use the terms polling stations (PS) and parts interchangeably in this document. Each part has a PDF file corresponding to a PS, the smallest polling unit in India, normally consisting of 800 to 1500 voters. Moreover, when referring to a specific electoral or voter "roll," I am referring to the voter list at the PS/Part-level.

Citation

This work has been tremendously challenging (and also rewarding in improving my coding skills and exploring the frontier of machine learning on high-performance computing clusters). Hence, I sincerely urge you to cite the dataset whenever you use it for any kind of analysis. A suggested format is:

> Sharik Laliwala. 2025. *Indian Electoral Rolls*. Harvard Dataverse. DOI: *to be added*.

*Contact: sharik@berkeley.edu

Technical Approach

I use the open-source Surya-OCR engine to parse non-OCR rolls. Surya-OCR provides strong multilingual recognition but requires high-end GPUs (for example, NVIDIA A100s or H100s) to work at this scale. To meet these requirements, I rely on the following high-performance computing resources:

1. University of California, Merced's Pinnacles cluster
2. NSF ACCESS-CI allocations (Grant No. SOC250025), which provided me access to:
    - San Diego Supercomputer Center's Expanse cluster,
    - Texas A&M University's ACES and FASTER clusters,
    - Purdue University's Anvil cluster, and
    - Pittsburgh Supercomputing Center's Bridges-2 cluster

Workflow

1. Mass download of electoral rolls via the ECI's API
2. OCR pipeline's implementation for text parsing from non-OCR PDFs using Surya-OCR engine
3. (Very light) cleaning and descriptive metadata checks.

I am **NOT** releasing scripts used to download and extract texts from files at the moment. There is sufficient documentation for cross-validation and reproducibility available at the moment. I plan to release these scripts whenever I release the full dataset as part of a peer-reviewed journal article.

Dataset Outputs

In the main, my dataset provides three sets of files per electoral roll (PS/Part):

1. **Analyzable CSV:** Each PS has a CSV file with individual-level voter records. You can merge these files at the assembly constituency, parliamentary constituency, city, district, or state-level for aggregate-level analysis.
2. **Rendered JSON:** Each PS has a JSON file with a structured representation from which that particular PS's CSV is derived. This JSON enables reproducibility.
3. **State-level Metadata (JSON & CSV):** Extracted from each roll's first and last page. It contains location details and PS-specific voter totals (male/female/other/total) and cross-checks the corresponding PS's CSV for discrepancies.

Known limitations and issues

While I have repeatedly sampled my script to maximise accuracy, no OCR pipeline can guarantee 100% accuracy. I strongly recommend that you verify and cross-validate your calculations using this dataset before publication. Here are some common issues that I've come across:

- **Additions/Deletions codes:** These flags can be inconsistently rendered, so they are not always reliable.
- **Deleted names:** Deletion markings and strikethroughs are hard to capture, therefore some deleted voter records may be missing in my dataset.
- **Symbols next to names:** There may be square checkboxes or emptyboxes next to voter names. This issue is not because of a fault in my OCR pipeline, but is an artifact of the PDFs made available by the ECI.
- **Roll numbers:** When a roll number is not detected on a given line (typically when the OCR pipeline fails to capture a deleted voter record), numbering may shift forward to the next detected entry.
- **NAs in the metadata file:** In rare cases, small values (often 0/1) have not parsed clearly from the last page of each roll. To avoid cross-row/column contamination, addressing this issue did not make sense and often worsened extraction in the state-level metadata file.
- **Wrong numbers in the metadata file:** In some cases (under 1% of PS), gender-wise voter numbers are contaminated across rows/columns. You can either exclude such PS from your analysis or re-extract numbers from JSON files.

Potential validation checks

You can validate data fields by applying some basic checks. I have not implemented these because they might exclude voter records that are problematic within the roll itself, rather than errors from my OCR pipeline (for example, missing names or house numbers, duplicate EPICs, or ages greater than 125). Here is a suggestive list for such checks:

1. **Voter/Parent/Spouse Names:** Are there names with just one character? Or voter names that do not make sense (such as "." or "_")? Are there missing entries?
2. **Roll No & Voter ID:** Are there roll numbers that exceed the typical maximum length of a roll (1500 voters)? Any missing roll numbers or EPIC ID? Any EPIC ID that do not make sense (under 5 characters or over 30 characters)?
3. **Part Number:** Any characters in this field instead of numeric values?
4. **Age:** Unusual age numbers such as under 18 or over 110 years?
5. **Totals in Metadata file:** Any extremely skewed gender ratio (5:1 : M:F) or over 200 voters in "Other" gender column? Additions or deletions exceeding 15-20 % of the original totals?

Acknowledgements

This work would not have been possible without the high-performance computing resources made available via:

1. University of California, Merced's Pinnacles cluster
2. NSF ACCESS-CI allocations (Grant No. SOC250025), which provided me access to:
    - San Diego Supercomputer Center's Expanse cluster,
    - Texas A&M University's ACES and FASTER clusters,
    - Purdue University's Anvil cluster,
    - Pittsburgh Supercomputing Center's Bridges-2 cluster

The Haryana release was primarily OCR-processed on UC Merced's *Pinnacles* and Purdue's *Anvil* clusters.

My text extraction script uses the Surya-OCR engine developed by Datalab. Existing work done on India's voter rolls by Raphael Susewind, as well as Gaurav Sood and his colleagues' work, served as a prototype to follow.

While I wrote the core scripts, I received troubleshooting support from a vendor. I also used generative AI tools, especially OpenAI's o4-mini-high model, for coding assistance. All errors are my own, of course.

Contact

For queries about the dataset, feedback, or corrections, please contact:

Sharik Laliwala
PhD Candidate, Political Science
University of California, Berkeley
sharik@berkeley.edu

Codebook: PS Specific Files

Variable name	Description	Notes
Voter_ID	Alphanumeric identifier or EPIC assigned by ECI.	Fairly high accuracy whenever extracted
Roll_No	Sequential number of the voter within a part/PS.	May shift forward if a preceding line was unreadable (check known issues subsection above)
DelOrAdd_Code	Status flag for deletion, addition, or other update.	Has not been captured always.
Name	Voter's full name.	As printed. No normalisation applied beyond OCR cleaning.
Father_Name	Father's name.	
Mother_Name	Mother's name.	
Husband_Name	Husband's name.	
Others	Other guardian/relation if provided.	
House_Number	House number or address token(s).	
Age	Age in years.	Fairly high accuracy whenever extracted
Gender	Gender as printed.	{Male, Female, Other}. Fairly high accuracy whenever extracted
Assembly_Number	Assembly Constituency (AC) number.	From each page's header/footer
Assembly_Name	Assembly Constituency name.	From each page's header/footer
Part_Number	Part number of the roll.	From each page's header/footer
Section_Number	Section number within the part.	From each page's header/footer
Section_Name	Section name within the part.	If available on the page
Roll_Updated_On	Last update date of the roll.	From each page's header/footer
Qualifying_Date	Eligibility cut-off date (age).	From each page's header/footer
Total_Pages	Total pages in this part/PS.	From each page's header/footer
Current_Page	Page on which the record appears.	From each page's header/footer. For tracing and cross-validation.

Table 1: Variable dictionary for the PS-level files.

Codebook: State-level Metadata File

Note: I have clubbed related fields to keep this table compact. Actual column names follow the patterns indicated. I have also excluded the description of location identifiers (assembly constituency, part/PS number, parliamentary constituency, district, city/village, ward, police station, pincode) that are self-explanatory from any PS roll's first page.

Variable (grouped)	**Description**	`Notes / Column pattern`
Male/Female/Other/Total voters	Totals printed on the first page (by gender and overall).	`total_male, total_female, total_other, total_voters`
Last-page totals	Totals printed on the last page (by gender and overall). Cross-check these numbers with the first page's totals.	`total_male_lastpage, total_female_lastpage, total_other_lastpage, total_voters_lastpage`
CSV totals	Totals computed from the part/PS-specific CSV file (by gender and overall).	`extract_csv_total, extract_csv_male, extract_csv_female, extract_csv_other`
Differences	Differences: last-page-counts minus CSV-counts.	`diff_male, diff_female, diff_other, diff_total`
Missing percentages	Percent differences computed in all cases.	`missing_male_pct, missing_female_pct, missing_other_pct, missing_total_pct`
Additions (phase & gender-wise)	Additions to the roll broken down by revision phase and gender.	`Columns follow additions_phase#_{gender/total}. Phases aggregated as additions_{gender/total}`
Deletions (phase & gender-wise)	Deletions to the roll broken down by revision phase and gender	`Columns follow deletions_phase#_{gender/total}. Phases aggregated as deletions_{gender/total}`

Table 2: Dictionary for select variables for the state-level metadata file (see, meta_checked_STATECODE.csv file)