Parsing Bihar: How *The Hindu's* Data Team cracked the SIR puzzle

Updated - September 18, 2025 10:42 am IST

AREENA ARORA

The enumeration phase of the Special Intensive Revision (SIR) of electoral rolls in Bihar has excluded around 65 lakh electors from the draft electoral roll published on August 1, 2025. | Photo Credit: SHASHI SHEKHAR KASHYAP

This article forms a part of the Data Point newsletter curated by The Hindu's Data team. To get the newsletter in your inbox, subscribe here.

Hello data enthusiasts!

It's not every day that a data journalist has access to data that includes millions of rows and a ton of potential to tell robust stories, albeit riddled with challenges.

Such was the case after the completion of Election Commission of India's (ECI)'s Bihar Special Intensive Revision (SIR) exercise last month. The resultant data showed that a total of 7.24 crore electors are part of the latest electoral rolls — over 56 lakh electors fewer than the rolls prepared in January this year. Naturally, The Hindu Data Team decided to scrutinise the new rolls for patterns and we found some interesting trends.

Women, for instance, were deleted in disproportionately higher numbers and there was no evidence of religious bias in the rolls, but to get there was a lengthy data-analysis process. Here's how.

We first downloaded August rolls – a seemingly routine task. However, each record was one of about 600 in a PDF file and we had about 80,000 such PDF files to parse through, some of them being image-based and therefore difficult to be read by a machine. This data were used to find patterns in deletions by gender.

Then, for our latest story, we first wrote code to mass download data (again). This latest data was lists of electors' names, EPIC numbers, age, gender and reason for deletion, which was released by the ECI after a Supreme Court directive August 14.

It took several hours to download the data, on several computers. The second round of coding involved processing the data from PDFs into a structured, SQLite database to get it ready for analysis.

We relied heavily on Python to handle the sheer volume of the data, using it to automate the mass download, extracting structured data from PDFs, cleaning and normalising columns.

We now had names of people deleted along with their age, gender and reason for deletion, which we ran through a character-based machine learning models developed by Rachana and Sugat Chaturvedi in their 2023 paper, "It's All in the Name: A Character-Based Approach to Infer Religion." Put simply, the algorithm used large datasets to train itself to infer people's religion from their name.

The analysis predicted people's religion with high accuracy and was manually checked for accuracy randomly. We found no bias on religious basis. Of the 52.8 lakh names analysed,

the model was able to identify with very high probability that about 9.7 lakh (18.4%) were Muslims, which follows the State's population trends.

We also compared reason and gender-wise breakdown of deletions between Muslims and non-Muslims, which showed no significant variation.

Data journalism peaks when there's robust data available to be questioned and scrutinised, like any other human source.

The Bihar SIR is also an especially important data drive in the absence of census figures in the country, offering an official statewide lens into the state's demographic churn. The granularity of the data available will be a robust resource for future stories as well.

While you're here, have a look at what we've published over the past two weeks:

Last month, China's Ministry of Industry and Information Technology introduced interim measures to tighten controls on 'rare earth' mining and processing. While China's trading partners such as India and the U.S. are seeking alternative sources to reduce dependency, data shows that China's dominance in rare earths stems not only from resource availability but more so from its longstanding strength in mining and research capacity.

China digs in on 'rare earth', commands global market

The share of Indian schoolchildren enrolled in private coaching has risen sharply over the past seven years, according to government data. Notably, at the secondary level, the sharpest rise was recorded among rural girls.

Private tuitions are picking up in India's rural areas too, especially among girls

Recent protests in Nepal that took a turn towards violence did not happen in isolation. History and data show that Nepal is one of the most politically unstable democracies, with the country raking in the most number of government tenures in the world since 1990.

Political instability and economic difficulties behind Gen Z rage in Nepal

Here are this week's News in Numbers:

Demonstrators gather as smoke rises from the Parliament complex following fire set during a protest against the killing of 19 people after anti-corruption protests. | Photo Credit: ADNAN ABIDI

15,000

Number of CSS employees who staged a state-wide protest in Mizoram

Employees under the Centrally Sponsored Scheme (CSS) in Mizoram on Tuesday staged a state-wide protest demanding that the government fulfil their long-standing demand to be regularised under the state government.

Source: PTI