

Computational Social Science and Digital Methods: Rethinking the Study of Society in India's Data Age

Abhinav Biswal

Independent Researcher — Sociology & Law LL.B. (Madhusudan Law University, Cuttack); M.A. Sociology (Ravenshaw University); UGC NET (Sociology)

DOI: <https://doi.org/10.51584/IJRIAS.2026.11090096>

Received: 23 September 2026; Accepted: 28 September 2026; Published: 16 October 2026

ABSTRACT

Somewhere between the last census and the next one, India quietly became a country that leaves a digital footprint before it leaves almost anything else on paper. This paper takes that fact seriously. It offers a humanised, evidence-grounded overview of computational social science (CSS) and digital methods for scholars and students attending ICCSDM-26, tracing the field's conceptual origins, its explosive but institutionally uneven growth, and the methodological toolkit — web scraping, natural language processing, social network analysis, and agent-based modelling — that now sits alongside the interview and the questionnaire. Drawing on recent secondary data, the paper shows that India closed 2025 with roughly 1.03 billion internet users (70.0 per cent penetration) and around 500 million social media “user identities,” even as 440 million people, disproportionately rural, older, and female, remained offline. Read together with global data-volume estimates — some 173–182 zettabytes generated in 2025 alone — these figures make the case that computational methods are no longer optional for social research, but they also expose a quiet danger: a discipline that studies only what is digitised risks mistaking the loudest, most connected fraction of society for society itself. The paper closes by arguing for a grounded computational social science — one that treats big data as a companion to fieldwork and doctrinal-legal method, not a replacement for either, particularly in a linguistically and socially plural country like India.

Keywords: Computational Social Science; Digital Methods; Big Data; Digital Divide; India; Social Media Analytics; Data Ethics; Socio-Legal Research

INTRODUCTION

Consider an ordinary evening in an ordinary Indian household. Someone checks a WhatsApp forward before dinner, another scrolls Instagram while the rice cooks, a third argues about a court judgment on X. None of this feels like data. It feels like life, mundane and unremarkable. Yet every one of those small acts is captured, timestamped, stored, and, increasingly, studied. This is the quiet premise of computational social science: that the residue of ordinary digital life — clicks, posts, searches, judgments uploaded to e-court portals, satellite images of a flooded village — has become one of the richest archives of social behaviour ever assembled, and that new computational tools are needed to read it.

Computational social science (CSS) is usually described as the marriage of social-scientific questions with computational method — the use of large-scale digital data and algorithmic tools to explain, rather than merely predict, human behaviour. As Edelmann, Wolff, Montagne and Bail note in their widely cited review, the term itself emerged in the final quarter of the twentieth century, carrying two rather different meanings depending on where one stood. Within sociology, it first meant agent-based modelling — building small artificial societies inside a computer to see what kinds of collective patterns simple rules could produce. Within engineering and computer science, by contrast, almost any study built on a very large behavioural dataset came to be called computational social science. The field we recognise today is the uneasy, productive merger of both lineages.

This paper, prepared for seminar presentation at the International Conference on Computational Social Science and Digital Methods (ICCSDM-26), does three things. First, it traces the conceptual and institutional trajectory

of CSS, including how rapidly — and how unevenly — it has grown. Second, it lays out the digital methods toolkit now available to social researchers, anchoring the discussion in numerical secondary data on India's digital landscape rather than treating “big data” as an abstraction. Third, and most importantly for a discipline trained to worry about who gets left out of the sample, it asks what computational social science misses when it studies only the digitally visible half of a country as vast and unevenly connected as India.

Computational Social Science: Conceptual Foundations and Growth

In an influential 2020 article in *Science*, Lazer and colleagues observed that computational social science had “exploded in prominence” over the preceding decade, with thousands of papers published using observational trace data, large-scale online experiments, and simulations that would have been technically unfeasible only a generation earlier. Conferences, workshops and summer schools in the field multiplied across disciplines and continents. Yet the same authors were careful to note that the institutional scaffolding around CSS — research ethics review suited to platform data, doctoral pedagogy, shared data infrastructure — remained comparatively nascent, lagging well behind the pace of substantive publication.

Bibliometric work bears this asymmetry out in numbers rather than impressions. A systematic Web-of-Science analysis of the field's literature found that the annual volume of CSS publications has risen steadily and shows every sign of remaining on an upward trajectory, even as the thematic content of the field diversifies and fragments across periods — network analysis giving way to text-as-data, which in turn gives way to multimodal and platform-specific methods. A separate thematic review by Meckin and Elliot, working through citation analysis, was able to identify only forty-one papers as the field's most highly cited “canon” as of 2021 — a strikingly small core for a field whose surrounding literature runs into the thousands, suggesting a discipline still in the process of settling on its own foundational texts.

For a paper presented at a hybrid conference explicitly built around this theme, the lesson is worth stating plainly: CSS is not a finished toolkit waiting to be applied, but a field still negotiating its own boundaries, ethics and canon. That openness is precisely what makes it worth entering from a sociological and socio-legal vantage point rather than leaving it entirely to computer science.

Digital Methods: A Working Toolkit

“Digital methods,” in Richard Rogers' well-known formulation, are methods native to the digital medium — built to follow how platforms themselves organise, rank and circulate information — as distinct from “digitised methods,” which merely move an offline technique such as the survey online. In practice, the working toolkit available to a social researcher today includes at least the following: application-programming-interface (API) based data collection from social platforms; web scraping of public forums, news portals and government data releases; natural language processing (NLP) and sentiment analysis applied to text at a scale no team of human coders could manage; social network analysis to map the structure of online communities and information flow; geospatial and satellite-image analysis for questions of land, migration and disaster; and agent-based modelling to simulate how simple individual rules can generate complex collective outcomes.

The scale that makes these tools necessary, rather than merely convenient, is worth stating in numbers. Current estimates suggest that roughly 402.7 million terabytes of data — about 0.4 zettabytes — are created, captured, copied or consumed worldwide every single day, with the annual total estimated at somewhere between 173 and 182 zettabytes in 2025 and projected to climb toward 221 zettabytes in 2026. Instagram alone is estimated to carry some 70 million new posts a day, generating roughly 10.5 terabytes of fresh data; YouTube adds around 720,000 hours of video daily. No interview schedule, however well designed, can be administered at this scale. This is the honest justification for computational method: not novelty for its own sake, but a simple mismatch between the size of the phenomenon and the size of the traditional toolkit.

The Indian Digital Landscape: A Numerical Portrait

It is one thing to speak of “big data” in the abstract and another to see what it looks like in a specific, familiar society. India offers an unusually vivid case because its digital population is now large enough to be globally

significant while still leaving a very large minority — hundreds of millions of people — outside the digital net altogether. Table 1 draws together recent secondary data to sketch this picture numerically.

Indicator	Figure	Reference Point
Active cellular mobile connections in India	1.06 billion (72.5% of population)	Late 2025
Individuals using the internet in India	1.03 billion (70.0% penetration)	End of 2025
Growth in internet users, year-on-year	+223 million (+27.7%)	Oct 2024 – Oct 2025
Population remaining offline	440 million (30.0%)	End of 2025
Social media “user identities” in India	500 million (34.1% of population)	Oct 2025
Facebook users in India	706.9 million (47.6% of population); 69.7% male	June 2026
Instagram users in India	553.4 million (37.2% of population); 69.5% male	June 2026
Largest age cohort on Facebook & Instagram	25–34 years (37.8% and 38.5% of users)	June 2026
Global data created/captured/copied per day	≈ 402.7 million terabytes (≈ 0.4 zettabytes)	2025 estimate
Global data volume, annualised	≈ 173–182 zettabytes in 2025; ≈ 221 zettabytes projected	2025–2026

Table 1. India's digital landscape and global data scale, compiled from DataReportal's Digital 2026: India report, NapoleonCat platform statistics, and independent big-data industry estimates (2025–2026).

A few things stand out. First, the sheer velocity of change: India added an estimated 223 million internet users in a single year (October 2024 to October 2025), a growth rate of 27.7 per cent that few large economies could match. Second, the scale of what remains a minority in proportion but a vast number in absolute terms — 440 million people, roughly the population of the United States and Indonesia combined, still without internet access at the end of 2025. Third, a gendered asymmetry that a purely technological reading of “digital India” tends to erase: nearly seven in ten Facebook and Instagram users in the country are recorded as male. A discipline that draws its evidence from what platforms make visible is, whether it intends to or not, listening mainly to men in their twenties and thirties in a country where women, older citizens and rural residents remain systematically under-represented in the very data trail it studies.

Methodological Opportunities for Indian Social Science and Law

None of this is an argument against computational method — it is an argument for using it with eyes open. The opportunities are genuine and, for a scholar working at the intersection of sociology and law, particularly rich. Digitally-mediated social movements — the kind of youth mobilisation seen around examination-reform protests or campus-based agitations in recent years — leave behind a continuous, time-stamped record of how a grievance travels from a small forum to national visibility, something ethnography alone struggles to reconstruct after the fact. India's e-Courts platform and the broader digitisation of judicial records likewise create, for the first time, a corpus large enough for computational text analysis of case law, sentencing patterns and delay in the justice system — questions of genuine socio-legal importance that were previously answerable only through painstaking manual coding of a handful of judgments.

Sentiment and discourse analysis of social media can, carefully used, offer an early empirical window into how communal narratives, gender discourse or economic anxiety circulate — provided the analyst remembers, per

the data above, that the sample skews young, male and urban. Used this way, digital methods become less a replacement for the sociologist's traditional instincts and more an extension of them: a way of listening to a much larger, if imperfect, crowd before deciding where to go and ask questions in person.

Challenges: The Underside of the Data Deluge

The same numbers that make the case for computational social science also mark out its limits. Four challenges deserve particular attention in the Indian context. The first is sampling bias by omission: with roughly 30 per cent of the population offline at the end of 2025, and platform use skewed heavily toward men aged 18–34, any study built purely on scraped or platform-provided data is, structurally, a study of a specific demographic slice wearing the costume of “society at large.”

The second is institutional and infrastructural lag. As Lazer and colleagues observed of the field globally, ethics review, data-sharing norms and pedagogy have not kept pace with substantive output; in the Indian university system, where computational training remains largely confined to STEM departments, this gap is wider still, leaving many sociology and law departments without the infrastructure to either teach or ethically govern this kind of research.

The third is linguistic and algorithmic bias. Most natural-language-processing tools are trained overwhelmingly on English and a handful of other high-resource languages; applying them uncritically to a country with twenty-two scheduled languages and hundreds of dialects risks a kind of computational myopia, in which Hindi, Odia, Tamil or Bengali social discourse is either mistranslated, flattened, or simply excluded from the analysis for want of a usable model.

The fourth is access and consent. Platform application-programming-interfaces that once made research data relatively easy to obtain have, in the years since 2022–23, become progressively restricted or commercially paywalled, creating a two-tier research world in which well-funded institutions can buy data access that independent and early-career scholars in India often cannot. Privacy and consent frameworks, meanwhile, were largely written for interview-based research and translate awkwardly to a world of scraped, aggregated, often re-identifiable trace data.

Toward a Grounded, Humanised Computational Social Science

The argument of this paper is not that numbers should be distrusted, but that they should be read the way a good field researcher reads a village — attentively, and with an active sense of who is standing just outside the frame. A computational social science worth building in India would triangulate platform-scale data with the older, slower instruments of the discipline: fieldwork, doctrinal-legal analysis, interviews conducted in the languages people actually use to argue, mourn and organise. It would treat the 440 million Indians presently outside the internet not as a footnote but as a standing methodological caution, and it would invest, deliberately, in low-resource-language NLP so that the discipline's new instruments can hear more than one-fifth of the country's linguistic range.

This is, in the end, a fairly old sociological instinct wearing new computational clothes: the discipline has always known that who is counted shapes what is concluded. Computational social science simply raises the stakes of that old lesson, because the datasets are larger, the algorithms more opaque, and the temptation to mistake volume for truth correspondingly greater.

CONCLUSION

Computational social science and digital methods have, in barely two decades, moved from a curiosity at the margins of sociology to a central and unavoidable part of how society can be studied — a shift the numbers in this paper make difficult to dispute. India's own digital trajectory, adding hundreds of millions of new users within a single year while still leaving hundreds of millions more offline, makes the country both an extraordinary opportunity and a cautionary case for this emerging field. The task facing scholars gathering at

ICCSDM-26, this paper has argued, is not to choose between the old methods and the new, but to build a computational social science humble enough to know what its data cannot see — and generous enough to keep listening, in the field and in the archive, to the country that has not yet logged on.

REFERENCES

1. Edelman, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational social science and sociology. *Annual Review of Sociology*, 46(1), 61–81.
2. Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A., & Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, 369(6507).
3. Meckin, R., & Elliot, M. (2021). Computational social science: A thematic review. National Centre for Research Methods, University of Manchester.
4. Rogers, R. (2013). *Digital methods*. MIT Press.
5. DataReportal / Kepios / We Are Social. (2026). Digital 2026: India — essential headlines on internet, mobile and social media use.
6. NapoleonCat. (2026). Social media users in India — platform-level statistics, April–June 2026.
7. Statista. (2025–2026). Volume of data or information created, captured, copied, and consumed worldwide, 2010–2029 (forecast).
8. Exploding Topics / DesignRush / Digital Silk. (2025–2026). Compiled big-data-volume and internet-activity industry estimates.