Wikipedia as a political knowledge repository: Analyzing content on 2024 Indian Elections

Mike Gruszczynski
Associate Professor, The Media School,
Indiana University

Shobha S V
PhD Student, The Media School,
Indiana University

Abstract

Research examining non-Western, non-Global North political contexts is exceedingly rare in the field of communication; specific studies of information creation and diffusion within those contexts are even more so. Although the Global South is inclusive of the vast majority of the world's population (India alone has a population of around 1.4 billion as of 2025), most of our theories and empirical research originate in the Western – and especially American – context.

The aim of the study is to study the content production and consumption of articles related to Indian politicians - especially the members of parliament (MPs) of Lok Sabha (India's House of People) - with a focus on 2024 Lok Sabha elections. Multiple components of this research will examine how gender and partisanship are associated with the information present on candidate Wikipedia pages, reader and editor engagement around Indian elections in 2024, and citational quality and preference of the Wikipedia pages.

Introduction

According to the 'Internet in India report-2024' prepared by The Internet and Mobile Association of India (IAMAI) and Kantar (an international market research company), India has 866 million internet users with the number expected to hit 900 million by 2025 (*India's Internet Users to Exceed 900 Mn in 2025, Driven by Indic Languages*, 2025).

In addition to English language Wikipedia, India is home to about nineteen language Wikipedias, including Hindi, Tamil, Telugu among others ("Wikipedia in India," 2025). In 2023, English language Wikipedia had 4,700 active volunteer editors in India, third in comparison to the number of volunteers in the US and UK (Padilla, 2023). Despite these dizzying facts, however, academic scholarship focusing on the dynamics of Wikipedia in India is woefully inadequate. While there are several academic papers about US elections and Wikipedia, there is an acute paucity of Wikipedia-based research as far as Indian elections are concerned (but see Karandikar, 2012) . Our paper/s will be a small attempt at filling a big gap in the academic scholarship. Wikipedia is one of the most important sources of information and knowledge and plays an important and essential role in the contemporary digital information infrastructure. English language Wikipedia is one of the largest communities in India and information from the former often gets translated in other Indian languages as well. Our project directly contributes to Wikimedia Foundation's strategic direction of knowledge equity. Analysis about gaps in the information

quality can help the foundation and the volunteer communities towards designing adequate interventions in the future.

In this project, we examine the production and consumption of Wikipedia articles concerning Indian politicians, particularly the Members of Parliament (MPs) in the Lok Sabha (House of the People) and the other candidates that competed in the 2024 Lok elections.

The proposed grant project is three-tiered in its problem focus.

First, there is a serious lack of publicly-accessible data to study information politics in the Global South. We know that Wikipedia becomes active during elections (Agarwal et al., 2020) and hence we have chosen to focus on the quality articles in English language Wikipedia of all the candidates that have fought the Lok Sabha elections across all the constituencies in India. This research will involve the collection of a dataset spanning 2024 India Lok Sabha candidates' Wikipedia pages, including full text, metadata, candidate characteristics, the network structure of all candidate pages within Wikipedia, their electoral competitiveness, readers' and editors' engagement with the pages, quality of citations used, and pageview statistics for each page.

This component of the project will not only allow us to undertake the rest of this study, but also offer other researchers a rich dataset through which to examine the information politics of this portion of the Global South. No research questions or hypotheses will be tested for this portion of the project, as it is entirely focused on the generation of a dataset.

The second tier of the proposed research assesses how Wikipedia page quality and methodological rigor differ according to the political party affiliation of the candidates involved. We will also analyze this dataset from the perspective of gender.

The third and final tier of this proposed research utilizes this dataset to test the extent to which electoral competitiveness, candidate information quality and demographics, and connectivity within the Wikipedia graph are associated with information-seeking behavior as manifest in Wikipedia pageview statistics. Drawing from research in the American context showing an association between candidate viability and information-seeking behavior (Utych & Kam, 2014), we hypothesize that electoral competitiveness will be associated with higher page views in the Indian election context as well. We also hypothesize that the extent to which candidates' Wikipedia pages are central within the network of broader Lok Sabha pages will be associated with higher numbers of page views.

Date: Proposed work will start on July 1, 2025 and conclude on June 30, 2026.

Related work

The first tier of the proposed research builds on past research the PI has published in the domains of communication research and data science, most notably large-scale data collection efforts in the domains of gendered visual imagery in American electoral campaigns (Gruszczynski et al., 2023), the media-public relations nexus in media coverage of new scientific discoveries (Comfort, Gruszczynski, & Browning, 2022), the role of expertise and citation in political debates over science-based knowledge (Gruszczynski & Michaels, 2014; Comfort, Tandoc, & Gruszczynski, 2020), as well as the role of online technologies in the orientation of public attention (Gruszczynski & Wagner, 2017; Hunt & Gruszczynski, 2021).

The second tier of this project as described above is related to the questions of gender within the dataset. As in much of the world, women politicians in India face several barriers limiting their entry into the political sphere, and we expect those candidates' Wikipedia entries to be less rich - in multiple dimensions - as a result.

These barriers are reflected in Wikipedia as well. Wikipedia has a documented problem of gender imbalance both in terms of content and the number of women editors on the platform (Bear & Collier, 2016). For instance, in a study of Wikipedia pages about sociologists, researchers found that most of the articles on the platform were about white male sociologists. Similarly, in an important study co-written by a Malayali woman Wikipedian (Hussein), Indian women editors reported problems related to Internet access, family restrictions owing to sexism, lack of free time, psychological barriers and harassment within the community (Chakraborty & Hussain, 2022).

Another study observed that of all the biographies on the platform, less than 19% of them are women's biographies (Tripodi, 2021). It is often seen that Wikipedia's principle of notability on Wikipedia — "a test used by editors to decide whether a given topic warrants its own article" ("Wikipedia," 2025) — often works against women. Most of the women's biographies that were marked for deletion because they weren't considered notable enough (Tripodi, 2021; Adams et al., 2019).

We hypothesize in this work that candidate Wikipedia pages will vary in their quality and rigor as a function of gender, particularly as regards men and women candidates.

The third tier of this project as described above pertains to the positioning of candidate Wikipedia pages within the network graph of all candidate pages, in particular the centrality of candidate pages within the 2024 Lok Sabha election network. As stated earlier, we expect that those candidate pages with more centrality - this is to say, those pages with a high degree of connectedness within the network - will receive more page views. This is in line with previous research demonstrating that Wikipedia page quality emerges with more interconnectedness (Ingawale et al., 2013); as such, we expect that with increases in connectedness we will observe higher quality. Flowing from this, research has also shown Wikipedia browsing to exhibit less depth (e.g., users are less likely to continue browsing the site) when pages are of lower quality (Piccardi et al., 2023). Additionally, we expect that users browsing to high-connectivity candidate pages will be facilitated by the simple fact that more inbound links to those pages will make them more likely to be encountered by Wikipedia "browsers."

We also expect that electoral competitiveness will be systematically related to Wikipedia users' page view statistics. Past research has found that perceived candidate "viability" is associated with greater information-seeking behavior (Utych & Kam, 2014).

Methods

Data collection for this project will proceed as follows. First, we will pull all Wikipedia pages for Lok Sabha constituencies, each which includes electoral outcomes. Every candidate who ran for office will be scraped from those pages, and those candidates who had Wikipedia biographies will have those biographies pulled from Wikipedia as well. In addition to candidate biographies, we will pull the pages as they appeared at the time of candidates' announcement of a political run as well as on the final day of the election.

All information from candidate pages at those two time points will be saved as initially unstructured data. We will also pull the edit history for candidate biographies and the pageview statistics for those pages (from WikiData).

We will create a structured dataset by combining the electoral results, constituency names, and any contextual information present on Wikipedia (for example, candidate party, age, gender). Trained coders will additionally code the data for broader party coalition (e.g., whether a candidate was associated with a multiparty coalition) and candidate demographics not present on Wikipedia.

Data on citation quality will be collected by aggregating all of the external citations on Wikipedia pages and subjecting them to human coders, who will categorize them based on source type (news, blog posts, social media posts, official government sources, books, journal articles). The PI will then, in conjunction with coders, develop a measure of citation quality based on previous research (CITE) and an initial 10 percent random sample of citations.

We will then construct a network graph of the 2024 Lok Sabha campaign by pulling every Wikipedia citation from candidate biography pages, creating a graph consisting of all Wikipedia links emanating from candidate pages. This will allow us to calculate measures of network centrality, closeness, and degree.

These methods will be used to create the dataset proposed in the first project.

For the second project focused on citation quality and editorial rigor as a function of candidate gender and partisanship, the analyses will encompass both descriptive, mixed-methods analyses and quantitative statistical modeling (correlation and regression, as appropriate) to test our hypotheses.

The third project will make use of a combination of network analyses and the correlational measures to examine the extent to which page views of candidate Wikipedia biographies are a function of candidates' gender, partisanship, party coalition, and degree and centrality within the 2024 Lok Sabha Wikipedia network.

Expected output

The dataset derived from this research will be open access published on the Harvard Dataverse under a Creative Commons license. We will submit an accompanying manuscript on the dataset to the open access *Journal of Quantitative Description: Digital Media*.

Production of this dataset will benefit academics studying information politics in the Global South. Additionally, data journalists and data enthusiasts will be able to make use of the resulting data to better understand digital platforms' place within politics.

We plan on submitting the proposed manuscript on citation quality, candidate gender, and partisanship to the *Journal of Information Technology and Politics*. If not accepted at that venue, we will seek publication in *Information, Communication, and Society*. We plan on submitting this manuscript to the 2026 Annual Meeting of the *International Communication Association*, as well as the 2026 *Wiki Workshop* meeting. The latter presentation will focus on interventions to close the gender gap in candidate Wikipedia biographies.

Publication of this manuscript will be primarily oriented at academics. However, as we are planning on disseminating this research beyond the academic community through both media

and the venue of the *Wiki Workshop*, we are aiming at illuminating editorial gaps in political candidates' pages and offering ways to intervene to minimize those gaps.

The proposed project on the Wikipedia network graph surrounding the 2024 Lok Sabha elections and public attention to Wikipedia pages will be submitted to the *Journal of Computer-Mediated Communication*, an open access journal, with plans to submit to *Mass Communication and Society* if not accepted in that venue.

This manuscript will have primary impact on academics, though through our assessment of network connectivity within Wikipedia election graphs, we expect to be able to offer mechanisms through which Wikipedia editors can more thoroughly integrate the pages on which they work into the broader information ecosystem.

Risks

As the proposed research neither involves primary data collection using human subjects nor sensitive topic areas, we anticipate this research to be of minimal or no risk.

Community impact plan

The researchers will work with the Indiana University Media School to promote the dissemination of research findings to national and international media outlets through the creation of press release materials and coordinated media campaigns. We plan to coordinate with the Indiana University India Gateway in New Delhi to increase engagement beyond the United States.

Our plan to publish the resulting dataset in an open access format will impact not only researchers, but also data journalists and data enthusiasts, who will be able to derive insights beyond our own analyses through which to understand Wikipedia creation and usage in myriad ways.

The researchers plan to submit the manuscript focusing on candidate page quality, gender, and partisanship to *Wiki Workshop 2026* in order to disseminate our findings to an audience that includes the broader Wikipedia community. Our aim is to connect our research findings with interventions in the Wikipedia editing community so as to increase recognition of gender gaps in political candidate biography quality and to offer interventions to decrease those gaps, through for example Edit-a-Thons.

Evaluation

We will evaluate our proposal through rigorous data analysis, publications in peer reviewed journals, making the datasets publicly available, and engagement with local Wikipedia communities about the study. The research fund chairs should evaluate based on the following criteria: our project is feasible and demonstrates methodological soundness, publication in peer reviewed journals, and contribution to scholarship. Measurement of success will include the following outputs and outcomes, as well as process-based metrics we will use to gauge the project's ongoing success.

Outputs and Outcomes

- Creation and publication of a large-scale open access dataset of Wikipedia editorial rigor and use in the context of the 2024 Lok Sabha elections.
- Presentation of the resulting research at national (*American Education in Journalism and Mass Communication*) and international (International *Communication Association*) conferences will help to disseminate our research findings to a broad academic audience.

- Presentation of research at Wikimedia conferences (*Wiki Workshop, Wiki Mania*) will give the researchers opportunities to offer interventions aimed at mitigating gender gaps as they manifest in community-curated information sources.
- Publication of three research articles derived from the dataset will increase the amount of academic work done within the underserved context of politics in the Global South.

Process based success

We plan to evaluate success of the process of this research through its adherence to the proposed budget and the following timeline:

- Initial unstructured data collection completed by August 1, 2025
- Human coding of dataset completed by October 15, 2025
- Submission of gender, partisanship, and page quality manuscript to *International Communication Association* conference by November 1, 2025.
- Publication of dataset to Harvard Dataverse by November 7, 2025
- Dataset manuscript submitted to journal by December 15, 2025
- Gender, partisanship, and page quality manuscript submitted to journal by February 1, 2026
- Network graph and page attention manuscript submitted to *Association for Education in Journalism & Mass Communication* conference by April 1, 2026

Budget

Please see our proposed budget at this link.

References

1. Agarwal, P., Redi, M., Sastry, N., Wood, E., & Blick, A. (2020). Wikipedia and Westminster: Quality and Dynamics of Wikipedia Pages about UK Politicians. *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, 161–166. https://doi.org/10.1145/3372923.3404817
2. Adams, J., Brückner, H., & Naslund, C. (2019). Who Counts as a Notable Sociologist on Wikipedia? Gender, Race, and the "Professor Test." *Socius, 5*, 2378023118823946. https://doi.org/10.1177/2378023118823946
3. Bear, J. B., & Collier, B. (2016). Where are the women in Wikipedia? Understanding the different psychological experiences of men and women in Wikipedia. *Sex roles, 74*, 254-265.
4. Chakraborty, A., & Hussain, N. (2022). Documenting the gender gap in Indian Wikipedia communities: Findings from a qualitative pilot study. *First Monday*. https://doi.org/10.5210/fm.v27i3.11443
5. Comfort, S. E., Tandoc, E., & Gruszczynski, M. (2020). Who is heard in climate change journalism? Sourcing patterns in climate change news in China, India, Singapore, and Thailand. *Climatic Change, 158*, 327-343.
6. Hunt, K., & Gruszczynski, M. (2021). The influence of new and traditional media coverage on public attention to social movements: the case of the Dakota Access Pipeline protests. *Information, Communication & Society, 24*(7), 1024-1040.
7. *India's internet users to exceed 900 mn in 2025, driven by Indic languages*. (2025, January 16). https://www.business-standard.com/ind

ia-news/india-s-internet-users-to-exceed-900-mn-in-2025-driven-by-indic-languages-125011600835_1.html
8. Ingawale, M., Dutta, A., Roy, R., & Seetharaman, P. (2013). Network analysis of user generated content quality in Wikipedia. *Online Information Review*, *37*(4), 602-619.
9. Gruszczynski, M., & Michaels, S. (2014). Localized concerns, scientific argumentation, framing, and federalism: the case of Devils Lake water diversion. *Journal of Natural Resources Policy Research, 6*(2-3), 173-193.
10. Gruszczynski, M., & Wagner, M. W. (2017). Information flow in the 21st century: The dynamics of agenda-uptake. *Mass communication and society, 20*(3), 378-402.
11. Karandikar, M. M. (2012). Media convergence and communication features in websites of political parties in india. *Unpublished doctoral dissertation, Department of Communication and Journalism, University of Mumbai, Mumbai, India*.
12. Padilla, S. (2023, December 5). *Wikipedia releases its top 25 most-viewed pages of 2023 | CNN Business*. CNN. https://www.cnn.com/2023/12/05/tech/wikipedia-chatgpt-oppenheimer-indian-entertainment/index.html
13. Piccardi, T., Gerlach, M., Arora, A., & West, R. (2023). A large-scale characterization of how readers browse Wikipedia. ACM Transactions on the Web, 17(2), 1-22.
14. Tripodi, F. (2023). Ms. Categorized: Gender, notability, and inequality on Wikipedia. *New Media & Society*, *25*(7), 1687–1707. https://doi.org/10.1177/14614448211023772
15. Utych, S. M., & Kam, C. D. (2014). Viability, information seeking, and vote choice. The Journal of Politics, 76(1), 152-166.
16. Wikipedia in India. (2025). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia_in_India&oldid=1285212086
17. Wikipedia:Independent sources. (2025). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:Independent_sources&oldid=1281082788