Reducing the gender gap in AfD discussions: an evidence scoring approach

Giovanni Luca Ciampaglia, Khandaker Tasnim Huq
University of Maryland, College Park
EMAIL ADDRESS REDACTED

Abstract

The goal of this research is to investigate the impact of Articles for Deletion (AfD) discussions in the gender gap on Wikipedia. Prior work has emphasized the role of collective deliberation processes as potential factors behind the gap in gender representation between women and men on Wikipedia. AfDs offer an open forum for editors to decide whether content meets the criteria for inclusion of the project. To gauge notability, discussants are expected to cite reliable evidence from independent sources outside of Wikipedia. Thus biases in how individual discussants assess notability may perpetuate a gender gap in AfDs. Here we propose to investigate these deliberative processes from the lens of information foraging and develop methods to identify, collect, and score AfD discussions by the type and amount of external evidence used by AfD discussants in assessments of notability. This research could form the basis for the development tools to promote consistent outcomes regardless of gender in deliberations about content inclusion.

Introduction

A well-known form of bias in Wikipedia is gender bias: across all Wikimedia projects worldwide, only 18% of the content is about women, and on the English Wikipedia, only 19% of the biographies are about women [1]. There are many theoretical and empirical contributions on the systematic factors behind this gap in content representation. Prior work used community surveys to determine the demographic makeup of the Wikipedia community and found that women are only a small fraction of the contributors; follow-up surveys identified potential behavioral and psychological factors that limit female contribution on Wikipedia.

However, a gender disparity in contribution may not be the only factor behind such a stark gap in content coverage; broader editorial processes may perpetuate this gap even in the absence of such an imbalance. For example, Tripodi [2] found that biographies of women are nominated for Articles for Deletion (AfD) discussions at a higher rate than those of men. This is puzzling, especially since earlier work on gender asymmetries in Wikipedia has found that the encyclopedia tends to cover women that are more notable than men [3].

What is the cause for this differential treatment between women and men? Within Wikipedia, AfDs offer an open forum for editors to decide whether content meets the criteria for inclusion of the project. Owing to their deliberative nature and the fact that discussants are self-selected, their outcome typically reflects the editorial stances toward content inclusion (often referred to as “inclusionism” and “deletionism”) of those taking part in the discussion [4]. In recent work, we have studied the role of editors with an “inclusionist” and “deletionist” stance in AfD debates, and found that the number of deletionists in a debate is a strong predictor of its outcome [5].

Of course, the importance of the stance of individual discussants does not mean that these deliberations are mere headcounts. The goal of an AfD discussion -- like that of other deliberative processes in Wikipedia -- is to reach a consensus about a specific editorial outcome. In the case of AfDs, this often hinges upon the notability (i.e., level of external attention) of the subject whose entry is under discussion.

To gauge notability, discussants are expected to cite reliable evidence from independent sources outside of Wikipedia. Thus, when it comes to deliberating about the inclusion of biographical content on Wikipedia, gender bias may not only arise from the self-selection patterns of discussants to individual deliberations; it may also stem from biases in how individual discussants assess the notability of the subjects under discussion.

In summary, gender may act as a catalyst for the deletion of content about women. If this is the case, we would expect deletionists to be over-represented in AfD debates about biographies of women, and these debates to be shorter and less developed than those about biographies of men, with less evidence cited in support of their notability, and from sources of lower reliability and quality. Furthermore, we may expect content inclusion decisions in AfD to be a possible contributing factor for the observed gender gap in content representation on Wikipedia as a whole. Thus, our research questions for this project are the following:

RQ1. Are decisions about inclusion of biographies of women deliberated in a different way from those of men? In particular, we are interested in comparing deliberations along both quantitative (e.g. group size, length of the discussion, etc.), and qualitative dimensions (e.g., degree of development, strength of consensus);

RQ2. Can differential outcomes between deliberations on (the biographies of) men and women be ascribed to different stances toward content inclusion? In particular, are AfD deliberations on biographies of women targeted by deletionists more than those of men?

RQ3. What is the role of notability assessments in determining differential outcomes between men and women in the AfD process? Can differences in notability assessments explain known gender-based asymmetries, like the aforementioned tendency for women to be nominated in AfDs more often than men [2], [6] and the imbalance in notability between the coverage of men and that of women in Wikipedia [3], [7]?

As mentioned before, the gender gap in content representation is not limited to the English Wikipedia only. It is reasonable to assume that the deliberative biases we seek to understand may act in addition to, or as a reflection of, any pre-existing gender bias of the broader cultural and social context in which these discussions take place. Thus, to make sure our results are not dependent on the particular community under study, we propose to explore the above research questions in two smaller Wikipedia communities in addition to the English one: the Italian and Bengali Wikipedias. The choice of these languages is dictated both by practical considerations (our team is fluent in both languages), and by substantive reasons: they are both large projects (1,000,000+ articles for Italian, while the Bengali recently crossed the 100,000+ threshold [8]) at different stages of development.

Because we are interested in characterizing the role of gender in determining deliberation outcomes of interest (e.g. deletion), we need to be able to match AfD discussions of different genders so that other contributing factors that could determine the same outcomes are not responsible for the observed patterns. In particular, besides describing key deliberation metrics (RQ1) we want to compare AfDs of men and women with similar stance composition among discussants (RQ2), and with similar external evidence for notability assessments (RQ3).

To do so, we propose to develop machine learning methods for matching AfD discussions based on the type and amount of external evidence available to discussants. This *evidence scoring* approach would form the basis for the development of a AfD matching tool that discussants could use to review the outcomes of previous discussions, and check whether they are consistent with the current one. In doing so, we want to study the information foraging practice of AfD participants and understand how the language-specific affordances of the AfD process affect notability assessments. For example, in both the English and Bengali Wikipedia, discussants are provided with direct links to external sources (e.g. Google, NYT, JSTOR), unlike in the Italian. We will thus conduct a content analysis of a sample of AfDs to determine the most frequent external sources used in each community.

Of course the above questions require an operational definition of the gender of the subjects of AfD discussions. Here, we choose to leverage Wikidata as the ground truth about gender in AfD discussions. Our preliminary analysis on AfDs in English analyzed by Tripodi [2] shows that this approach achieves >60% coverage on average. Furthermore, even though our focus is on the gap in coverage between men and women, it allows us to extend our analysis beyond traditional genders (e.g. transgender men, women, etc.) due to the availability of rich gender information.

Date: July 1, 2023 -- June 30, 2024

Related work

Biographies are one of the most common entries in Wikipedia. In the English Wikipedia, for example, there are 1.5 million biographies about people who are notable in various fields like literature, politics, academia, and sports [9]. However, only 19% of them are about women, which is a major indicator of the gender gap in the encyclopedia ([1], [10]–[12]). Reagle & Rhue [10] compared the coverage and representation of gender in the English Wikipedia and *Encyclopedia Britannica*. They found that, even though Wikipedia covers more biographies than *Britannica*, a significant number of biographies of women are missing from Wikipedia [10]. Konieczny and Klein [1] used data from Wikidata to measure the gender gap longitudinally across cultures. Currently, their indicator, which is the ratio between the number of female biographies and total biographies in a given Wikipedia edition, shows that only 19% of biographies in the English Wikipedia are about women [1].

What could be the cause of this observed disparity? One possibility is that this gap in content arises from a gender gap in participation: although the number of women contributors has increased in recent years, they still form a minority of the total community [13]. To understand the reasons behind the lower number of female contributors in Wikipedia, Hargittai and Shaw [11] surveyed a diverse (in terms of gender, age, and nationality) panel of Wikipedia editors and readers about their experience and skills with editing Wikipedia [11]. They found that Web-use skills are a significant predictor of contribution to Wikipedia, and that, on average, male report being more skilled than women at contributing to Wikipedia. However, psychological factors may also affect the differences in contribution beyond differences in skills and experience. For example, drawing from an international survey with Wikipedia contributors, Collier and Bear

[12] noted that even when women have similar (reported or observed) editing abilities, they are less confident about it and report more discomfort with editing [12], a finding that mirrors similar observations in the context of mathematics [14] and engineering [15]. Collier and Bear also found that when contributing to Wikipedia women face criticism from their male peers more often compared to other men contributors [12]. Thus, women may tend to focus less on editing as a way to avoid discomfort from conflict.

More recent work has highlighted that the Wikipedia gender gap is a complex phenomenon comprising of a number of asymmetries, discursive dimensions, and social concerns beyond the mere demographic makeup of the community of contributors. For example, Wagner et al. [3] used Google search results as a proxy for notability and found that women tend to be more notable than men, suggesting a subtle glass ceiling effect. More generally, Beytía and Wagner [16] argue that the gender gap manifests itself in three distinct phases over the lifecycle of content in Wikipedia: a) the *selection* phase, which is about the creation and selection of new content by contributors, the choice of topic, and its notability evaluation, b) the *building* phase, which refers to the collaborative editing process itself, and c) the *positioning* phase, which relates to the structural placement of articles in terms of topic, language, occupation, region, historical era, etc.

Article for Deletion discussions play a major role in the selection of content in Wikipedia and thus have attracted considerable attention from the literature. Prior work has investigated its group composition and opinion dynamics [4], [5], [17]. Tasnim Huq and Ciampaglia [5] observed polarization based on the stance towards the inclusion or deletion of articles. Tripodi [2] compared the AfD discussions about biographies of men and women in the English Wikipedia and found that women are often miscategorized as non-notable: their biographies are nominated for deletion more often than males. And although other studies have already investigated the role of behavioral factors in the gender bias of AfDs [18], less is known about notability assessments from an information foraging perspective [19], and in particular large-scale investigations on how AfD discussants frame the notability of women.

Methods

There are a number of methodological challenges when it comes to investigating the role of gender in AfD discussions, as this requires first to determine whether the subject of an AfD is a biography or not, and then the gender of its subject. To determine whether an AfD is about a biography we propose to apply natural language processing (NLP) techniques to the text of the discussion itself, which is available even if an article has been deleted. Specifically, we will develop a machine learning classifier to determine whether an article is a biography or not, and then search Wikidata for a matching entry (either via an interwiki link or by querying Wikidata directly using the title of the AfD). To train our biography detection model, we will also use ground truth from Wikidata, and in particular look for instances of the “Human” class (Q5). Since this ground truth is incomplete, we will use semi-supervision to account for missing labels.

Our preliminary analysis shows that this approach is viable: our semi-supervised biography detection classifier achieves 90% accuracy, and the simple query-based matching technique sketched above yields gender labels for >60% of the entries in the same corpus of AfDs of biographies used by Tripodi [2] to investigate miscategorization rates by gender in the English Wikipedia.

Note that we *do not* plan to use machine learning to predict the gender of individuals; we will rely on ground truth from Wikidata instead.

See Table 1 for a preliminary breakdown of available gender labels for a sample of AfDs from the English Wikipedia using this method. In the table, only genders with >5 entries in Wikidata are listed.

Table 1: AfDs of biographies with >5 Wikidata entries

	Kept	Deleted	Other	Total
Male	17,744	33,849	11,505	63,098
Female	6,878	10,165	4,156	21,199
Transgender female	43	11	22	76
Trans woman	9	47	5	61
Non-binary	21	24	12	57
Transgender male	8	7	3	18

Our next step will be to perform a content analysis of biographicals AfDs to understand how discussants evaluate external sources. Using an open coding approach, we will train human annotators to identify notability assessments made by discussants within the AfD. These assessments will typically include citations to external sources, which will allow us to identify what sources AfD discussants use in practice, and thus to define metrics that operationalize the concept of external evidence. For example, based on prior work by Wagner et al. [3] one possible metric could be the number of Google search results associated with the subject of the AfD. The definition of these metrics will form the basis for the implementation of a Web scraper that will allow us to collect data for the evidence scoring step.
Our next step will be an analysis of AfD debates of biographical articles, in which we will compare debates about biographies by gender. As the vast majority of Wikidata labels covers men and women, we will focus first on a comparison of these two genders. However, we will also test how our approach performs when using gender labels beyond these traditional genders (see Table 1). To match AfDs based on the strength and type of external evidence, we will experiment with propensity score matching, though we may consider alternative matching methods, like Coarsened Exact Matching [20]. Thanks to the scoring method, we will investigate the effect of the gender of the AfD subject on group composition and stance of the debates (RQ2), and on how editors assess external sources to gauge the notability of the subject of a biographical article (RQ3). We will pre-register our study on OSF or a similar repository.
Finally, to achieve our goal of a cross-cultural study of AfD discussions, we will develop a suite of multi-lingual tools for AfD debates. Our initial goal will be to support three languages of interest (English, Italian, and Bengali) including a parser for AfD discussions, and a set of scrapers of core external sources used in each language community.

Expected output

We plan to publish 1-2 articles for the project described above. We will target interdisciplinary venues in the areas of Social Computing and Human--Computer Interaction, such as CHI, CSCW, ICWSM, Nat. Comm, Nat. Hum. Beh., and Sci. Adv. All these venues provide Open-Access publishing options for which we have budgeted funds.
We also have a track record of dissemination in the media (our research has been covered in the WSJ, NBC, NPR, SciAm, The Conversation, etc.) and we will pitch an editorial for a general audience describing our main findings.
The propensity score matching technique will form the basis for an evidence scoring service that we will deploy on Wikimedia cloud services so that AfD discussants can identify matching AfDs by gender based on the amount of evidence at hand. Such a tool could help discussants identify outcomes that are consistent with previous consensus on similar cases.
Finally, we will release all code of our tools and for the replication of the findings from this research under an open source license on

Github or GitLab. All corpora and datasets will be uploaded on an institutional repository such as figshare or Zenodo.

Risks

This research presents minimal risk for AfD discussants since it will rely in its entirety on publicly available data. We will maintain IRB oversight throughout the duration of the project. There are also potential sources of societal risk associated with our tools. One possible risk is associated with reliance on Wikidata for ground truth on gender of subjects of AfD discussions, and in particular the risk of misgendering individuals by relying on labels that may be erroneous or vandalized. To mitigate this risk, which is admittedly low, we will make sure to periodically refresh the corpus of labels and manually review any change we see in it.
Another risk stems from the possibility that our matching tool may be misused by AfD discussants in a way that perpetuates existing bias. To mitigate this risk our tool will not provide any recommendation on the outcome to take in the AfD under consideration and will not provide a "default" match or rely on default external sources. Instead, users will be provided with the opportunity to customize the matching criteria and the sources of external evidence used to score prior AfD discussions.
Finally, there is risk to the research itself, and in particular the risk that AfD discussants may not adopt our tool. To mitigate this risk we will post invitations to engage in participatory design sessions with relevant WikiProjects (e.g., Women in Red) prior to committing to a particular design.

Community impact plan

This project could help researchers and Wikipedia contributors gain a better understanding of AfD debates on biographies of women and other genders not typically considered when dealing with the gender gap in content representation. Our AfD discussion matching service could enable researchers and contributors to compare the potential outcome of ongoing discussions with that of similar discussions based on the availability and type of external evidence. We envision such a tool could promote consistency in outcomes across debates regardless of gender. For example, WikiProjects devoted to closing the gender gap, like "Women in Red", may benefit from the ability to identify gaps in outcomes between discussions of biographies of women and men. As a proof of concept, we will build a dashboard keeping track of AfDs of biographies with relevant stats broken down by gender.
We will take a number of steps to maximize the chances of adoption of our tools. We will list our project on the Wikimedia Research Index and post regular updates there. We will deploy our tools on the Wikimedia cloud services and make the source available on Github. We will advertise our research on relevant WikiProjects related to gender and inclusion, like "Women in Red", and "LGBT Studies" and invite their members to attend participatory design sessions. We will submit a proposal to run a demo or workshop on our tools at Wikimania 2024. Finally, we will submit a pitch for an article on The Conversation or similar outlet for outreach to the broader public.

Evaluation

We will follow standard evaluation approaches for various parts of our research. For example, for the semi-supervised biography detection task, we will consider the task successful if the classifier achieves AUC >95% in multi-lingual settings. We will use standard methods from causal inference to evaluate the ability of our matching methods to match AfDs on simulated data. We will also elicit feedback from users of our tool on the quality of the matches and on whether it helps achieve a consensus in AfD deliberations. Finally, to evaluate the progress of adoption of our work, we will track a number of standard project analytics, such as web visits, API calls, downloads on Github, etc.

Budget

[Full budget: LINK_REDATED]
Total for this request: $29,326.

Staff Costs

Senior Personnel

PI Professor Giovanni Luca Ciampaglia, of the University of Maryland's College of Information Studies, will provide overall direction and oversight of this research project and outreach to mission organizations. Responsibilities include supervision of analysis, data collection, reporting, publication, and dissemination along with responsible archival and management. He will lead evaluation, drafting, and dissemination of results. He will commit 2 summer weeks at a cost of $7,256.

Undergraduate annotators

This project will support 3 undergraduate students for annotation and translation tasks from Bengali and Italian. For this, we plan an amount that works out to 3 students at $20 per hour, 10 hours per week, for a period of 10 weeks, for a total of $6,000. The students staffing these positions will be recruited from the international student population at UMD.

Fringe Benefits

Fringe benefits include health insurance, FICA, unemployment, workers' compensation, retirement, terminal leave payout and employee assistance. Amounts for the sponsor's contribution to employee fringe benefits are calculated using UMD's U.S. Department of Health and Human Services (DHHS) approved Fringe Benefit Rates effective July 1, 2022. The approved rates are as follows: 29.9% for Faculty, 35.6% for Staff, 27% for Graduate Assistant and 7.6% for Contractual Faculty/Staff, hourly students and most Faculty/Staff additional pays. Tuition Remission is a UMD fringe benefit but is not included in the fringe calculation and is budgeted separately as applicable. Additional information about fringe benefits can be found at: ora.umd.edu/resources/benefits-stipends. The Fringe Benefit Rate Agreement can be found at: ora.umd.edu/resources/fa. Fringe rates could be adjusted in future years. Total requested funds for fringe benefits across the project period is $1007.

Travel

The PI will present research from the project at meetings and conferences at locations to be determined.

Domestic Travel

Travel funds in the amount of $2,226 is budgeted for domestic travel and research conferences. Costs for conference attendance can be broken down into registration costs ($350), round trip airfare ($650), lodging ($810 for 3 nights), and food ($79/day per diem for 4 days).

International Travel

Travel funds in the amount of $4,011 is budgeted for international travel and research conferences. Costs for conference attendance can be broken down into registration costs ($500), round trip airfare ($2,091), lodging ($820 for 4 nights), and food ($100/day per diem for 5 days)

Other Direct Costs

Publication fees

To cover fees related to publishing in Open Access fees, the project requests $5000 to cover the publication of 1-2 articles.

Indirect Costs (F&A)

The max indirect cost rate allowed by the funding entity is 15% of the Total Direct Costs (TDC) base. The total indirect cost requested is $3,825.

Response to reviewers and meta-reviewers

REDACTED

References

[1] P. Konieczny and M. Klein, "Gender gap through time and space: A journey through Wikipedia biographies via the Wikidata Human Gender Indicator," *New Media Soc.*, vol. 20, no. 12, pp. 4608–4633, Dec. 2018, doi: 10.1177/1461444818779080.

[2] F. Tripodi, "Ms. Categorized: Gender, notability, and inequality on Wikipedia," *New Media Soc.*, p. 14614448211023772, Jun. 2021, doi: 10.1177/14614448211023772.

[3] C. Wagner, E. Graells-Garrido, D. Garcia, and F. Menczer, "Women through the glass ceiling: gender asymmetries in Wikipedia," *EPJ Data Sci.*, vol. 5, no. 1, p. 5, Dec. 2016, doi: 10.1140/epjds/s13688-016-0066-4.

[4] D. Taraborelli and G. L. Ciampaglia, "Beyond Notability. Collective Deliberation on Content Inclusion in Wikipedia," in *2010 Fourth IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshop*, Sep. 2010, pp. 122–125. doi: 10.1109/SASOW.2010.26.

[5] K. Tasnim Huq and G. L. Ciampaglia, "Characterizing Opinion Dynamics and Group Decision Making in Wikipedia Content Discussions," in *Companion Proceedings of the Web Conference 2021*, Ljubljana Slovenia: ACM, Apr. 2021, pp. 632–639. doi: 10.1145/3442442.3452354.

[6] M. Lemieux, R. Zhang, and F. Tripodi, "'Too Soon' To Count? The impact of gender and race on perceived notability".

[7] C. Wagner, D. Garcia, M. Jadidi, and M. Strohmaier, "It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia," *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 9, no. 1, pp. 454–463, Aug. 2021, doi: 10.1609/icwsm.v9i1.14628.

[8] A. G. Dastider, "A Brief Analysis of Bengali Wikipedia's Journey to 100,000 Articles," in *Companion Proceedings of the Web Conference 2021*, Ljubljana Slovenia: ACM, Apr. 2021, pp. 552–557. doi: 10.1145/3442442.3452340.

[9] "Why it's so hard for biographies about women to stay on Wikipedia," *Marketplace*. https://www.marketplace.org/shows/marketplace-tech/why-its-so-hard-for-biographies-about-women-to-stay-on-wikipedia/ (accessed Mar. 31, 2023).

[10] J. Reagle and L. Rhue, "Gender Bias in Wikipedia and Britannica," *Int. J. Commun.*, vol. 5, no. 0, Art. no. 0, Aug. 2011.

[11] E. Hargittai and A. Shaw, "Mind the skills gap: the role of Internet know-how and gender in differentiated contributions to Wikipedia," *Inf. Commun. Soc.*, vol. 18, no. 4, pp. 424–442, Apr. 2015, doi: 10.1080/1369118X.2014.957711.

[12] B. Collier and J. Bear, "Conflict, criticism, or confidence: an empirical examination of the gender gap in wikipedia contributions," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, Seattle Washington USA: ACM, Feb. 2012, pp. 383–392. doi: 10.1145/2145204.2145265.

[13] "Community Insights/Community Insights 2021 Report - Meta." https://meta.wikimedia.org/wiki/Community_Insights/Community_Insights_2021_Report (accessed Mar. 22, 2023).

[14] E. H. Fennema and J. A. Sherman, "Sex-Related Differences in Mathematics Achievement and Related Factors: A Further Study," *J. Res. Math. Educ.*, vol. 9, no. 3, pp. 189–203, 1978, doi: 10.2307/748997.

[15] J. B. Bear and B. Collier, "Where are the Women in Wikipedia? Understanding the Different Psychological Experiences of Men and Women in Wikipedia," *Sex Roles*, vol. 74, no. 5, pp. 254–265, Mar. 2016, doi: 10.1007/s11199-015-0573-y.

[16] P. Beytía and C. Wagner, "Visibility layers: a framework for systematising the gender gap in Wikipedia content," *Internet Policy Rev.*, vol. 11, no. 1, Mar. 2022, Accessed: Mar. 13, 2023. [Online]. Available:

https://policyreview.info/articles/analysis/visibility-layers-framework-systematising-gender-gap-wikipedia-content

[17] J. Schneider, A. Passant, and S. Decker, “Deletion discussions in Wikipedia: decision factors and outcomes,” in *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration,* Linz Austria: ACM, Aug. 2012, pp. 1–10. doi: 10.1145/2462932.2462955.

[18] Z. Worku, T. Bipat, D. W. McDonald, and M. Zachry, “Exploring Systematic Bias through Article Deletions on Wikipedia from a Behavioral Perspective,” in *Proceedings of the 16th International Symposium on Open Collaboration,* Virtual conference Spain: ACM, Aug. 2020, pp. 1–22. doi: 10.1145/3412569.3412573.

[19] P. Pirolli and S. Card, “Information foraging,” *Psychol. Rev.*, vol. 106, pp. 643–675, 1999, doi: 10.1037/0033-295X.106.4.643.

[20] S. M. Iacus, G. King, and G. Porro, “Causal Inference without Balance Checking: Coarsened Exact Matching,” *Polit. Anal.*, vol. 20, no. 1, pp. 1–24, 2012, doi: 10.1093/pan/mpr013.