Tracking Biases in the Wikimedia Ecosystem

Marco Antonio Stranisci
Università degli Studi di Torino, Italy
Viviana Patti
Università degli Studi di Torino, Italy
Mirko Lai
Università del Piemonte Orientale
Rossana Damiano
Università degli Studi di Torino

Abstract

Despite the great attention of academics and activists to the issue of bias in the Wikimedia Ecosystem, there is a lack of approaches to systematically track this phenomenon.
Our proposal aims to develop the first framework for bias detection that jointly considers three forms of potential discrimination: *i.* the stereotypical representation of people belonging to certain communities; *ii.* their quantitative underrepresentation in Wikidata and English Wikipedia; *iii.* the lack of editors for the curation of pages about them.
The framework combines Information Extraction (IE) techniques with Semantic Web (SW) technologies to detect and analyze the above-mentioned types of bias and these analysis are performed over three different point in time (2015, 2020, 2025), in order to jointly ensure a synchronic and diachronic analysis of bias.
The resulting technology, models, and datasets will be released through a public and Open Source API endpoint that will be also used to produce interactive graph representations of data for a non-expert audience.
A dissemination plan focused on academic conferences and workshop and Wikimedia events aims at promoting the adoption of our framework by Wikimedia editors and the scientific community.

Introduction

The presence of bias in the Wikimedia ecosystem is an open issue acknowledged by several academic research work (Graells-Garrido *et al*, 2015; Field *et al*, 2022) and by the community itself[1]. As a matter of fact, one of the core pillars of Wikimedia 2030 Strategic Direction is to foster knowledge equity by facilitating diversity in accessing and contributing to Wikimedia projects. *Bias* is often adopted as a vague term, though (Blodgett *et al*, 2020): studies and initiatives on this topic are fragmented, hindering the implementation of a thorough evaluation of the problem.
Our proposal addresses the issue by proposing a **multidimensional evaluation framework to track the presence of bias in Wikidata and English Wikipedia**. The framework is built upon the distinction between *representational* and *allocative* harms provided by Barocas *et al* (2017). The former refer to stereotypical representations of vulnerable groups to discrimination; the latter to their underrepresentation in data or among data contributors.
Specifically, we develop a socio-technical solution that jointly considers three types of

[1] https://en.wikipedia.org/wiki/Racial_bias_on_Wikipedia

potential bias against women and non-Western people:
- stereotypical representations in their biographies (Sun *et al*, 2021);
- lack of representation of vulnerable groups in Wikidata and English Wikipedia (Adams *et al*, 2019);
- limited diversity among contributors (Flöck *et al*, 2011; Kaffee *et al*, 2018).

Our approach relies and systematizes our previous approach to bias detection based on biographical event extraction (Stranisci *et al*, 2023) and semantic modelling (Stranisci *et al*, 2023). Specifically, we develop a method focused on two research objectives.

RO1. Automatically identify the presence of biases against people characterized by specific intersections of socio-demographic traits (e.g., gender, race, age, occupation).
RO2. Track the evolution of biases against people through time through the comparative analysis of different Wikipedia and Wikimedia snapshots

The main contribution of our initiative is the release of a methodological framework that can support Wikimedia contributors in fostering knowledge diversity and that can be scalable to different languages and Wikimedia projects. The framework is released under an Open Source licence and delivered through public API endpoints.

The project duration is 12 months. It starts on the 1st of July 2025 and ends on the 30th of June 2026.

Related work

A complete survey of existing work on bias in the Wikimedia ecosystem is outside the scope of this submission. In this section we present related work on bias in Wikimedia that allows positioning our approach.
Natural Language Processing (NLP)-based methods is characterized by a wide range of approaches: Field *et al* (2022) provide an ensemble of traditional Machine Learning methods to compare biographies of people belonging to different socio-demographic groups. Sun *et al* (2021) adopt a token-based event detection classifier to get events from biographies and analyze the distribution of events related to certain demographics; Lucy *et al* (2022) propose an approach based on vector embeddings.
From a topic perspective, research on bias in the Wikimedia ecosystem focuses on different discrimination axes: gender (Sun *et al*, 2021), sexual orientation (Weathington et al, 2023), race (Adams *et al*, 2019).
Finally, most of existing research specifically focuses on a single source of bias, such as underrepresentation (Shaik *et al*, 2021) and stereotypical representation (Lucy *et al*, 2022) of vulnerable groups.

Our framework is characterized by multidisciplinarity and intersectionality (Crenshaw, 1991): we adopt methods from NLP and Semantic Web (SW) and do not limit our study to a single axis of discrimination but leverage all ones that can be gathered from Wikidata.

Methods

Our methodology is divided in 3 distinct steps.
Data collection and cleaning. We use the official Wikimedia APIs to gather all the information about people in Wikidata and English Wikipedia in three different points in time: 2015, 2020, and 2025. Specifically, we exploit a Wikimedia bot to download all the Wikidata triples about people and all their English Wikipedia pages (Stranisci *et al*, 2021) and clean them

Semantic modelling of metadata. Given the participatory nature of the Wikimedia ecosystem, knowledge stored in Wikidata is not homogeneous and some properties need to be remodeled to perform scientifically-sound analysis. Coherently with our previous work (Stranisci *et al*, 2024), we map all properties related to people to a semantic model that enables grouping individuals according to 9 coarse-grained categories. E.g., whether a person is linked to an organization or has a relation with a person. Additionally, we create novel properties to better align existing ones E.g., we identify all the countries to which places of birth belong to compute more meaningful statistics about people's origins .
Biographical event extraction. We leverage the corpora for biographical event extraction that we created through years (Stranisci *et al*, 2022; Stranisci *et al*, 2023) to improve our existing model for biographical event detection. We shift our approach from a token-based classification (Jijang *et al*, 2024) to a sequence-to-sequence approach (Cabot *et al*, 2021). We take as an input each English Wikipedia page and return as an output a list of events, which is the representation of the page itself in our framework. Such an approach is also adopted to Wikipedia and Wikidata editors public pages to compare them.
Intersectional analysis of *personas*. We take into account different features that contribute to the identities of people by adopting an intersectional approach (Crenshaw, 1991): we form clusters of people pages by combining different socio-demographic axes. Our approach is based on *personas* (Cheng *et al*, 2023): abstract sets of features that characterize people with similar traits and characteristics. For instance 'African Millennial actress' or 'European director with an immigrant background'. The analysis of biases is based on comparing the representations of different *personas* in Wikidata and Wikipedia along the following axes: *i.* which stereotypical events are associated with a specific *persona*?; *ii.* which *personas* are underrepresented in Wikidata?; *iii*. which *personas* are characterized by a lack of Wikimedia editors? All these analysis can be either performed synchronically (E.g., how African Millennial actress are represented against North American X-generation actors?) and diachronically (E.g., which representational changes have occurred to the African Millennial actress *persona* through time?).
The taxonomy and the classifier will be released under an open source licence. The processed data will be made publicly available through API endpoints on top of which we will develop graphical representations of results to engage with non-academic audiences.

Expected output

We plan to release the expected output of our research

1. Open Source Release of the Framework. We will release on Github all the code to replicate the framework to other Wikipedia and Wikidata snapshots. Models will be hosted on HuggingFace[2]. The audience of this output are computer scientists who can build their research on top of our library.
2. Datasets. Each Wikipedia and Wikidata dataset snapshot will be semantically encoded and published as an output of our research. We plan to release a new snapshot each year. The audience of these outputs are professionals on the topic of bias who have access to periodical release of data.
3. API endpoint and graphical representations. We release the framework through an Open Source API endpoint and graphs. The audience of the endpoint are Wikimedia contributors who can assess the impact of their activities to foster knowledge equity. Additionally, we plan to

[2] https://huggingface.co/marco-stranisci/wikibio

present our API endpoint to non-profits that are committed against discrimination.

Risks

We identified the following risks together with the strategies to prevent and mitigate them.

1. Lack of engagement with Wikimedians. Level: medium. Prevention: we plan to participate in Wikimedia events and engage with the community by discussing our progress on Meta-Wiki. Mitigation: we develop a tutorial for the usage of our framework by a non-expert audience.
2. Lack of engagement with non-profit. Level: medium. Prevention: we leverage the existing networks of the University of Turin to reach a wide number of associations that are actually partners of our institution. Mitigation: we record a set of video lectures on bias in Wikimedia and on how to face them.
3. Computational cost of the analysis. Level: low. Prevention: we perform all the computations asynchronously on the HPC4AI of the University of Turin with subsidized costs. Mitigation: we scale down our technological requirements for event extraction to Small Language Models.

Community impact plan

Our community impact plan aims to: *i.* operationalize our framework in downstream tasks that are useful for the Wikimedia community; *ii.* contribute to the advances of research on bias in Wikimedia.

1. We propose a session at the 2026th edition of Wikimania during which we will present a tutorial of our framework and public APIs.
2. We present a scientific contribution at the 2025th NLP for Wikipedia Workshop, during which we will discuss our framework with other NLP practitioners that work on Wikipedia

Evaluation

We design an evaluation that considers both the scientific value of the framework and its impact on the community

1. Scientific soundness. The framework is evaluated through recognized metrics by the research community:
   - F-1 score above 0.85 for event detection;
   - Competency Questions for the evaluation of the semantic alignment.

2. Community Impact (quantitative). 500 single users access the API endpoint 6 months after the release of the technology.

3. Community Impact (qualitative).
   - 100 respondents of a survey encountered at Wikimania and at the Wiki Workshop;
   - 7 non-profit organizations encountered.

Budget

The three main components of our budget are the following:
- personnel costs. The budget covers the work of a post-doctoral researcher who will lead the entire project. Supervision costs are also considered for a figure that ensures the quality of the work developed.
- computational costs. We allocate a specific part of the budget to the computational resources that are needed to train our model and to perform inferences.
- dissemination and community engagement costs. We hypothesize costs that cover our participation in the 2026 edition of Wikimania, NLP for Wikimedia Workshop, and in the Conference on Fairness, Accountability, and Transparency.
Here is it possible to consult the budget

References

Adams, J., Brückner, H., & Naslund, C. (2019). Who counts as a notable sociologist on wikipedia? gender, race, and the “professor test”. Socius, 5, 2378023118823946.

Barocas, S., Crawford, K., Shapiro, A., & Wallach, H. (2017, October). The problem with bias: Allocative versus representational harms in machine learning. In 9th Annual conference of the special interest group for computing, information and society (p. 1).

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020, July). Language (Technology) is Power: A Critical Survey of “Bias” in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5454-5476).

Cabot, P. L. H., & Navigli, R. (2021, November). REBEL: Relation extraction by end-to-end language generation. In Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 2370-2381).

Cheng, M., Durmus, E., & Jurafsky, D. (2023, July). Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1504-1532).

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. Stanford Law Review, 43(6), 1241-1299.

Field, A., Park, C. Y., Lin, K. Z., & Tsvetkov, Y. (2022, April). Controlled analyses of social biases in Wikipedia bios. In Proceedings of the ACM Web Conference 2022 (pp. 2624-2635).

Flöck, F., Vrandečić, D., & Simperl, E. (2011, June). Towards a diversity-minded Wikipedia. In Proceedings of the 3rd International Web Science Conference (pp. 1-8).

Graells-Garrido, E., Lalmas, M., & Menczer, F. (2015, August). First women, second sex: Gender bias in Wikipedia. In Proceedings of the 26th ACM conference on hypertext & social media (pp. 165-174).

Jiang, B., Li, Z., Asif, M. S., Cao, X., & Ma, Z. (2024, April). Token-based spatiotemporal representation of the events. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5240-5244). IEEE.

Kaffee, L. A., & Simperl, E. (2018, August). Analysis of editors' languages in wikidata. In Proceedings of the 14th International Symposium on Open Collaboration (pp. 1-5).

Lucy, L., Tadimeti, D., & Bamman, D. (2022, January). Discovering Differences in the Representation of People using Contextualized Semantic Axes. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing.

Shaik, Z., Ilievski, F., & Morstatter, F. (2021, October). Analyzing race and citizenship bias in Wikidata. In 2021 IEEE 18th international conference on mobile Ad Hoc and smart systems (MASS) (pp. 665-666). IEEE.

Stranisci, M. A., Patti, V., & Damiano, R. (2021). Representing the under-represented: A dataset of post-colonial, and migrant writers. In 3rd Conference on Language, Data and Knowledge (LDK 2021) (pp. 7-1). Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Stranisci, M. A., Mensa, E., Diakite, O., Radicioni, D. P., & Damiano, R. (2022).

Guidelines and a Corpus for Extracting Biographical Events. In Proceedings of the 18th Joint ACL-ISO workshop on Interoperable Semantic Annotation (ISA-18) (pp. 20-26). European Language Resources Association (ELRA).

Stranisci, M. A., Damiano, R., Mensa, E., Patti, V., Radicioni, D., & Caselli, T. (2023). WikiBio: a Semantic Resource for the Intersectional Analysis of Biographical Events. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 12370-12384). Association for Computational Linguistics.

Stranisci, M. A., Bernasconi, E., Patti, V., Ferilli, S., Ceriani, M., & Damiano, R. (2023, October). The World Literature Knowledge Graph. In International Semantic Web Conference (pp. 435-452). Cham: Springer Nature Switzerland.

Stranisci, M. A., Bassignana, E., Cabot, P. L., & Navigli, R. (2024). Dissecting Biases in Relation Extraction: A Cross-Dataset Analysis on People's Gender and Origin. In GeBNLP 2024-5th Workshop on Gender Bias in Natural Language Processing, Proceedings of the Workshop (pp. 190-202). Association for Computational Linguistics (ACL).

Sun, J., & Peng, N. (2021, August). Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 350-360).

Weathington, K., & Brubaker, J. R. (2023). Queer identities, normative databases: Challenges to capturing queerness on Wikidata. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 1-26.