# Hidden Voices: Reducing gender data gap, one Wikipedia article at a time

**Neeraja Kirtane[1], Anuraag Shankar[2], Chelsi Jain[3], Ganesh Katrapati[4], Senthamizhan V[1],**

**Raji Baskaran[5], Balaraman Ravindran[1],**

Robert Bosch Centre for Data Science and Artificial Intelligence, IITM Chennai[1], PICT, Pune [2]
College of Technology and Engineering, MPUAT[3], IIIT Hyderabad[4] Superbloom studios[5]
kirtane.neeraja@gmail.com

## Abstract

Wikipedia is the most widely available structured repository of information on the Internet. However, gender disparity has been observed in wiki articles, and it is a major issue. We aim to tackle this problem using Machine Learning methods to generate wiki-like biographies for notable women on Wikipedia. We present Hidden Voices, a project which will assist wiki editors and enthusiasts in writing more biographies about women, thereby increasing their representation on Wikipedia.

**Keywords:** Automation, Diversity, Gender Bias, Summarization, Women's biographies

## Introduction

Wikipedia is one of the top-most visited websites on the Internet. It is also easily accessible to all. It is run by volunteer contributors and editors who write articles. It is also one of the few large-scale webs of content that are free of advertisement revenue and commercial use.

There is a gender-based disparity on Wikipedia. In particular, gender-based asymmetry for the living person biography category is a considerable 19% in English [1]. As a global digital community, all participants - of all gender identities and ethnic, racial and national identities will be better served with a Wikipedia that has more balanced, diverse, and nuanced information about the people and topics of interest to all people. Despite being the largest democratic digital commons, Wikipedia entries and editorial participation are dominated by a small fraction of the world's demographic.

Wiki also has a learning curve for those interested in being new editors. To write a Wikipedia article, one has to curate the data, get enough citations, get relevant figures and tables, and ensure it is written in a wiki-style. The design of a fully opt-in volunteer editor network also presents other challenges for the participation of women. This is because they may have competing expectations on the use of "disposable" time and energy in the form of domestic chores, child care and social expectations. By automating the solution, there will be an increase in the diversity of the editors.

To address the gender disparity issue and to help a larger and more diverse section of people overcome the initial challenges of writing biographies, Our project (Hidden Voices) aims to create an automated solution that generates biographies.

### Related work

There have been notable efforts by individuals like Jess Wade[2] and groups like Project Women in red [3] which have made considerable progress in creating and maintaining biographies of women in Wikipedia. (Banerjee and Mitra, 2016) propose an automated solution to write Wikipedia articles. Recently, there has been work done by Meta [4] on generating wiki-like biographies for minority communities. The recent work in this field uses Large Language Modes (LLMs), which have proven very efficient in generating natural language text. However, these generative models suffer from "Hallucination," i.e., they have a tendency to produce inaccurate information, repetitive text and out-of-context information.

Our approach involves using artificial intelligence (AI) system that still leverages the advantages of LLMs while avoiding its pitfalls by using a human-in-the-loop.

## Experiments

Our pipeline is divided into three parts - web extraction to get relevant articles about notable women, generating factoids from the data collected and finally generating articles in wiki-like format from the factoids as in Fig 1. We create a user interface in which the user can enter the name of a person. After that the data is scraped and intermediate factoids are generated. Each generated factoid is mapped with the scraped data. The person can then monitor if the generated factoids are correct. After that, a wiki-like biography is generated with the factoids. The person who has entered the name can again check if the text is accurate and not hallucinating.

[1] https://bit.ly/3Nrnk9c

[2] https://www.washingtonpost.com/lifestyle/2022/10/17/jess-wade-scientist-wikiepdia-women/

[3] https://w.wiki/347

[4] https://about.fb.com/news/2022/03/generating-biographies-of-marginalized-groups/

### Web Extraction

To get a list of notable women, we ask people from various organizations and groups to suggest names and give information of notable women that they know. We also rely on lists of people who are in top of their field which are published by trusted sources [5]. We focus on women in STEMM in our first phase. We used multiple search engines to perform a keyword search and fetch relevant articles. The content was then ranked to filter out better results by assigning relevance scores to the documents. This would allow documents containing the name of the POI to be ranked higher than other noisy documents.

### Intermediate Knowledge Representation

Intermediate knowledge representation was done by extracting the relationship of the subject (POI) with various objects. The intuition behind going through this step was to get relevant information from the scraped data in a uniform format. Two approaches were used to identify relationships: Rule Based (RB) and Machine Learning (ML), described in Table 1. Three methods were explored to express relationships - triplets, a knowledge graph and factoids. Triplets were of the form subject-relation-object. Due to its simpler structure and easy interpretability, factoids were chosen as the method of representation. As traditional NLP methods were not giving good results we generated factoids from the scraped data using GPT-3.

### Text Generation

We implemented (Chen et al., 2020) as it is a table-to-text generation model especially for Wikipedia, which use LLMs. The generated text using this model hallucinates a lot by producing inaccurate and repetitive text as shown in Fig 2. One reason for this is that the set of triples used to generate the text is very pithy, therefore is not enough to generate meaningful text. We find a workaround for this by producing meaningful factoids with more information as our intermediate information. We do this with the help of GPT-3 to extract more dense factoids.

### Human-in-the-loop evaluation

After generating factoids from GPT-3, they were checked for how factually correct they were by having a human evaluate the data that was passed on to the next step. The human makes sure that factoids generated are all from the scraped content. This ensures that accurate data was passed on to the text generation step. After the text is generated a human evaluates the text to check its accuracy.

[5]https://www.psa.gov.in/article/she-75-indian-women-steam/3628

## Notability on Wikipedia

For a biography to be published on Wikipedia, the person has to be notable enough.Wikipedia has the following criterias for notability: one should have a sufficient online presence, have reliable secondary sources, and have a significant coverage.[6]. Wikipedia has a set of whitelisted and blacklisted websites which determines the reliability of a certain source. A secondary source is the one in which gives information about the primary thing that we are discussing. Online coverage of notable women is especially less in India. This narrows down the number of individuals who pass the notability criteria on Wikipedia. Therefore, Wikipedia's criteria of notability is a flawed metric when determining if the article should be published or not. There is also a notable feature that more than 41% of biographies marked to be removed as being "not-notable" enough are that of women [7]

## Limitations and Conclusion

One of the major limitations with the current method is that we are relying on GPT-3 which is not a freely available open-sourced model. There is also a risk of hallucinations that comes with training LLMs. The notability and reliability criteria on Wikipedia also creates a barrier to actually publish the article on Wikipedia.

We plan to replicate the experiments carried on GPT-3 with publicly available models for our task. We also plan to automate relevant inline citations linked to the text.

We would like to thank Santi Advani, Shreya Goyal, Ziryan Seddek and Tobi Odufeso for their valuable inputs. We would like to thank Denvr Dataworks for providing us with compute to run the experiments.

## References

[Banerjee and Mitra2016] Siddhartha Banerjee and Prasenjit Mitra. 2016. Wikiwrite: Generating wikipedia articles automatically. In *IJCAI*, pages 2740–2746.

[Chen et al.2020] Mingda Chen, Sam Wiseman, and Kevin Gimpel. 2020. Wikitablet: A large-scale data-to-text dataset for generating wikipedia article sections. *arXiv preprint arXiv:2012.14919*.

[Gardner et al.2017] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

[Han et al.2019] Xu Han, Tianyu Gao, Yuan Yao, Demin Ye, Zhiyuan Liu, and Maosong Sun. 2019. Opennre: An open and extensible toolkit for neural relation extraction. *arXiv preprint arXiv:1909.13078*.

[6]https://en.wikipedia.org/wiki/Wikipedia:Notability

[7]https://journals.sagepub.com/doi/10.1177/14614448211023772

| Method | Category | Shortcomings |
|---|---|---|
| Noun Chunking | RB | Limited to prepositional phrases and limited to named entities and noun chunks |
| Dependency Relation | RB | Relies solely on the dependency relations between tokens to extract information and does not take into account other linguistic features such as named entities or semantic role labeling |
| Dependency Tag | RB | Relies solely on the dependency tags to identify the subject and object of a sentence. This approach may not always work well, as the dependency tags can be ambiguous or incomplete in some cases. For example, in complex sentences with multiple clauses, it may be difficult to determine the main subject and object. |
| AllenNLP Open Information Extraction (Gardner et al., 2017) | ML | Extracted triples were not always complete or accurate |
| NLPTriples | ML | Not able to handle complex sentence structures or ambiguous language |
| Stanford OpenIE | RB | Directly extracts triplets given the text<br>Works only for simple sentences |
| User-defined Tree-based Parsing | RB | Infeasible since the number of rules to extract all relations is enormous |
| OpenNRE (Han et al., 2019) | ML | Performs multi-class classification given two entities<br>Works only for simple relations |

Table 1: Methods explored for relationship extraction

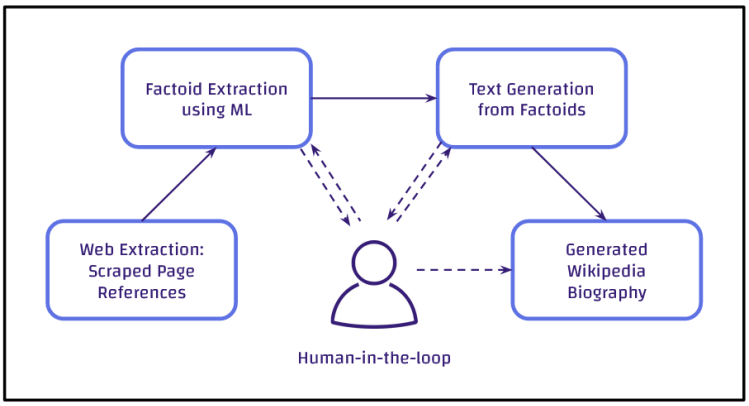


Figure 1: Wikipedia Generation Pipeline

Original:

She earned a degree in performing arts from Universidade Estadual de Campinas in 2005, then began her career at SESI and then worked with the theater company Os Satyros. Until 2013 she was one of the reporters of the "Custe o Que Custar" as the only woman in the team, after winning a competition to be the eighth member of the team in September 2009. She appeared on "Big Brother Brasil 14".

Output:

Monica Iozzi de Castro Paulo, born in September 2009 in Ribeiro Preto, Castro Paulo, Brazil, is performing arts and performing arts. In 2005, Monica Iozzi founded Os Satyros and Os Satyros. In 2005, Monica Iozzi founded the "Big Brother Brasil 14" (Universidade Estadual de Campinas). In 2013, Monica Iozzi founded Os Satyros (Universidade Estadual de Campinas). In September 2009, Monica Iozzi founded Os Satyros (Universidade Estadual de Campinas). In 2013, Monica Iozzi founded SESI (Universidade Estadual de Campinas).

Figure 2: Text generation example