Gender information gap on Wikipedia: A Knowledge graph solution

Project Report

Mariam Arustashvili, Thilo Dieing, Jusztina Judak, Aaron Koßler, and Petra Revesz

Group 03 - The Mining Minds
https://github.com/The-Mining-Minds/Gender-Gap-Wikipedia

Abstract. In this study, the authors investigate a possible data gender gap on Wikipedia. Using different methods such as sentiment analysis, web scraping and knowledge graphs, the authors test their formulated hypotheses. The results indicate that the gap is shaped differently to the authors expectations. While the gap between men and women seems to slowly close, the authors identify a big gap towards non-binary people, which future researchers should be aware of.

Keywords: *Knowledge graphs · Hyperlink graphs · Web mining*

1 Introduction

Gender equality has been a fundamental objective of societies worldwide and remains a compelling research topic even today. The Yentl Syndrome refers to the mistreatment or misdiagnosis of female patients due to differences in symptoms or illnesses compared to male patients, which can end deadly in some cases. In her book, Caroline Criado-Perez (2019) lists this type of syndrome as one of many examples of the gender data gaps, we can observe in today's society. These gender data gaps describe instances where a bias in the data collection process led to one or multiple genders being underrepresented. Most of the time, the genders that are left out or forgotten are female and non-binary people [Criado-Perez, 2019].

Our research paper aims to explore a potential gender data difference within the encyclopedia Wikipedia by analyzing three distinct areas (Five pages). These pages namely are Male, Man, Female, Woman and Non-binary. As there are many studies already available for comparing the gender bias between these pages by text analytics methods, we will analyze the topics using Knowledge Graphs. The authors rely on the following methods to test the hypotheses.

To create the graphs, we scrape the hyperlinks from DBpedia using SPARQL queries and investigate the gender gap based on the results. We calculate different measures such as centrality and density of the graphs in order to compare

the 5 distinct articles. Merging the 5 graphs and clustering will provide us with further insights for our analysis. In addition to that, we take a more in-depth look at the actual page contents via sentiment analysis. For this, we use the dictionary approaches by VADER and AFINN as validation.

The results indicate that the data gender gap on Wikipedia is shaped differently than expected by the authors. While there is a gap between men and women which manifests most in the woman text being more negative than the man text, the gap is not really existent regarding Backlinks and overall linkage. Nevertheless, the authors do find a significant gender gap regarding non-binary people.

2 Methodology

For the purposes of studying a possible gender data gap, the authors have identified three research assumptions prior to the project, from which they have derived testable hypotheses.

For hypothesis 1, the authors argue that sentiments expressed on Wikipedia pages can be influenced by the gender of the individuals they are about. Hence, we hypothesize that men-related pages are more positive, while women-related pages are less positive. The reason for this is that as the majority of Wikipedia authors (90%) are men [Khanna, 2012], one can expect them to write more positively in regard to pages of their own gender and less positively when formulating pages for women. As for the non-binary page sentiment, we hypothesize it to be more neutral. The argument for this expectation is that the non-binary gender and political correctness are highly intertwined, therefore leading pages on non-binary people to be more politically correct and making the text more neutral.

In the second hypothesis, we argue that the underrepresentation of women (and non-binary people) regarding achievements and contributions is a historical bias, leading us to hypothesize that man-male pages are mentioned more compared to women-female and non-binary pages since there are more biographies for men. According to a recently published article, the number of women's biographies has only increased by 4% up to 19.47% in the last eight years, even though foundations are making big efforts to close the gender gap on Wikipedia [Kandek, 2023]. This shows that links to man-male pages are more likely to appear on Wikipedia pages resulting in more remarks, compared to women-female and non-binary pages.

Furthermore, our third hypothesis suggests that a significant bias in linking patterns leads to a network structure where articles concerning men are disproportionately more central compared to articles concerning women or non-binary people. We theorize this is again due to the higher presence of male authors and pages (such as biographies etc.).

The investigation will look at distinct areas of Wikipedia for males, females, and non-binary people and compare them for differences. Based on these pages, we construct a Knowledge Graph to analyze the referenced Wikipedia pages for each original page. By doing so, the authors will look at possible differences caused by gender data gaps. With the proposed method, we hope that the possible existing gender biases on Wikipedia can be found and then dealt with accordingly, thereby closing part of the data gap.

We test our first hypothesis by analyzing the sentiment expressed within the pages and aim to uncover the general emotional associations and subjective perceptions surrounding these gender-related terms. The authors use the VaderSentiment library and validate the results with the AFINN library. Via the comparative analysis, we expect similar results so that we ensure reliability and validity.

Regarding our second hypothesis, we develop a backlink analysis by using the backlinks provided by DBpedia for our five target articles and building a negative binomial regression model as it enables the modeling of the relationship between a count response variable and the different gender pages while accounting for the extra variation observed in the data.

Lastly, we investigate the network structure of the five articles by merging the five separate hyperlink graphs, calculating the betweenness, and degree centrality, and removing the irrelevant nodes. With this combined graph, we also use the top prestige scores in order to list the highest-ranking entities within our merged graph. To classify clusters of the graph, we experiment with a divisive clustering algorithm, establishing four clusters.

3 Experimental setting

For the experiment, our data source is specific Wikipedia pages published in the English language (en.wikipedia.org). As mentioned before, the foundation of our Knowledge Graphs is the following five articles:

- "Male" (https://en.wikipedia.org/wiki/Male),
- "Man" (https://en.wikipedia.org/wiki/Man),
- "Female" (https://en.wikipedia.org/wiki/Female),
- "Woman" (https://en.wikipedia.org/wiki/Woman),
- "Non-binary gender" (https://en.wikipedia.org/wiki/Non-binary_gender)

In order to extract the embedded links from the listed pages, we use DBpedia (instead of Wikipedia) because it transforms the unstructured information from Wikipedia articles into a structured and organized format. Therefore, it is more suitable for our research, and it also provides a SPARQL endpoint that allows us to query and retrieve specific information from the dataset. The five pages were scraped from DBpedia using SPARQLWrapper. The queried data comes

without HTML tags and inconsistent formatting.

For our research we use the embedded links for the network structure (hypothesis 3), backlinks for the quoted analysis (hypothesis 2) and the Wikipedia text for sentiment analysis (hypothesis 1). Consequently, the format of our data is heterogeneous and we work with links as well as text. Regarding the backlink analysis, we use negative binomial regression (statsmodels package) and measure our results by comparing the coefficient and z-values.

The sentiment analysis is based on the five Wikipedia articles' content that needs to be extracted from Wikipedia's web pages, creating the base for our sentiment analysis. To simplify the task, we use the "Wikipedia" package that allows easy access to Wikipedia's content and is a convenient way to retrieve the relevant articles and remove the present HTML tags.

4 Evaluation and discussion of the results

During the scraping, we continuously review the data and evaluate the relevance of the reached pages in terms of our research. We assume that the built knowledge graphs reveal patterns in each and also differences between them, which would confirm a possible gender gap. Accordingly, we inspect the relationships between the Wikipedia pages and evaluate the significance of the links in the graphs. Furthermore, we analyze the features of the graphs separately, like centrality, communities, densities, and in-out degree distributions then we will compare these properties.

Some preliminary limitations of our research project are that we are not able to identify all the pages that are referencing our root pages because they can't be reached by them.

4.1 Sentiment Analysis

VADER To get a general overview of the article's Sentiments, we used the *vaderSentiment* library (Valence Aware Dictionary and Sentiment Reasoner), as it offers a wide range of application possibilities and does not require prepossessing that could bias the results. Regarding the first part of hypothesis 1, we can see that, while a similar percentage of the text (84.6% and 83.9%) is considered to be neutral, the Man page is judged a lot more positively than the Woman page (see Table 1). Furthermore, the page about the male human being is judged the most positive, while the one about the female counterpart is judged the most negative, therefore supporting our hypothesis. A similar pattern can be observed by the comparison of the Male and Female pages. Although the difference is within a 3% range for all three sentiments, the Female page is still judged more negatively, thus supporting our hypothesis as well. For hypothesis 1 we also focus on the neutrality of the articles regarding the non-binary gender

page. Here the Female page is judged the most neutral, but with only a difference of 0.7%, the non-binary page comes quite close. Since the non-binary page does not show any significant difference in the percentage of neutrally judged text in comparison to the other gender pages (Male and Female), we consider this part of hypothesis 1 not supported by our analysis.

AFINN To validate our VADER results and further investigate the articles by splitting them into sections, we analyzed the articles using the AFINN library, which is another wordlist-based library used for sentiment analysis. Contrary to VADER, AFINN required some preprocessing, such as cleaning up tags from the scraped article contents and splitting the text into sentences, since it only outputs one score per input string. The results of this analysis showed similar results to VADERs, although showing more polarized judgments overall. The articles judged the most positively, neutral and negatively also remain the same for the AFINN results, which validates the VADER results and therefore strengthens our hypothesis 1, that male pages have a more positive sentiment while female pages are more negative (see Table 1).

Article name	Labels	VADER	AFINN
Man	Positive	**11.5%**	**37.8%**
	Neutral	84.6%	51.1%
	Negative	3.8%	11.1%
Woman	Positive	6.8%	30.2%
	Neutral	83.9%	42.7%
	Negative	**9.3%**	**27.1%**
Non-Binary	Positive	4.2%	19.4%
	Neutral	93.1%	66.7%
	Negative	2.8%	14.0%
Male	Positive	6.7%	30.2%
	Neutral	91.9%	60.3%
	Negative	1.4%	9.5%
Female	Positive	3.9%	17.6%
	Neutral	**93.7%**	**72.1%**
	Negative	2.5%	10.3%

Table 1: Sentiment Analysis Results

To get a better understanding of how these sentiment judgments occur, we split the articles into sections using the major headlines, which required further preprocessing, because the headline priority could not be distinguished by only considering the articles' tags. Comparing the overlapping sections of the pages for Male and Female (see Table 2), we would argue that the differences in judgment are insignificant for all topics except for "Evolution" and "Sex Determination".

For "Evolution" the sentences on the Male page got significantly more negative scores than the ones on the Female page (27.3% and 17.4%). For the "Sex Determination" topic, we can observe the opposite, with a 9% more negative judgment for the section in the Female article (5.3% and 14.3%). The overlapping results do not show enough evidence for the gap between the overall scores, indicating that the reason for them lie within the other sections.

Topics	**Labels**	**Male**	**Female**
Introduction	Positive	12.5%	0.0%
	Neutral	87.5%	100.0%
	Negative	0.0%	0.0%
Evolution	Positive	9.1%	30.4%
	Neutral	63.6%	52.2%
	Negative	27.3%	17.4%
Sex Determination	Positive	47.4%	28.6%
	Neutral	47.4%	57.1%
	Negative	5.3%	14.3%
{Gender} across species	Positive	0.0%	13.6%
	Neutral	100.0%	81.8%
	Negative	0.0%	4.5%
Symbol (and usage)	Positive	22.2%	50.0%
	Neutral	77.8%	50.0%
	Negative	0.0%	0.0%

Table 2: Sentiment Analysis Topic Comparison Male / Female (AFINN)

We also compared the Man and Woman page (see Table 3), which led to a lot of overlapping topics. Big gaps in judgements can be seen for "Biology", "Social role / Culture and gender roles", "History", "Clothing (fashion and dress codes)", "Family / Fertility and family life" and "Education". "Biology" shows a lot more positive sentiments for the Woman page (45.2 %) in comparison to the Man page. Furthermore, sentences in the "Culture and gender roles" part of the Woman page are judged a lot more negatively than the "Social role" part of the Man page, indicating that there are more negative associations with the female role in society. Contrary to our expectations, the "History" section was rated more negatively for the Man page (33.3% / 0.00%). For "Clothing (fashion dress codes)" the part of the Woman page received a more negative (0.00% / 33.3%) and less positive score (25.0% / 16.7%). This could be traced back to existing beauty standards having a bad connotation in today's society. A similar judgment distribution can be observed for the "Family / Fertility and family life" section. The high negative score (27.3%) of the Woman page might be a result of a negative association towards the role of a woman within the family structure.

Topics	Labels	Man	Woman
Introduction	Positive	58.3%	50.0%
	Neutral	33.3%	28.6%
	Negative	8.3%	21.4%
Etymology	Positive	0.0%	16.7%
	Neutral	100.0%	83.3%
	Negative	0.0%	0.0%
Biology	Positive	18.8%	45.2%
	Neutral	81.2%	45.2%
	Negative	0.0%	9.7%
Social role / Culture and gender roles	Positive	35.0%	23.9%
	Neutral	65.0%	17.4%
	Negative	0.0%	58.7%
History	Positive	0.0%	21.4%
	Neutral	66.7%	78.6%
	Negative	33.3%	0.0%
Clothing (fashion and dress codes)	Positive	25.0%	16.7%
	Neutral	75.0%	50.0%
	Negative	0.0%	33.3%
Family / Fertility and family life	Positive	42.9%	0.0%
	Neutral	57.1%	72.7%
	Negative	0.0%	27.3%
Education	Positive	0.0%	28.6%
	Neutral	66.7%	61.9%
	Negative	33.3%	9.5%
Rights / Reproductive rights and freedom	Positive	28.6%	25.0%
	Neutral	14.3%	25.0%
	Negative	57.1%	50.0%
Sex Symbol / Gender Symbol	Positive	20.0%	0.0%
	Neutral	80.0%	100.0%
	Negative	0.0%	0.0%

Table 3: Sentiment Analysis Topic Comparison Man / Woman (AFINN)

Lastly, results for "Education" show a big gap for the negative scores between the Man and Woman page (33.3% / 9.5%). In conclusion, the analysis gives us interesting insights into the difference of the depiction of the two genders, male and female, and their corresponding human forms, man and woman. For future analysis, we suggest to look deeper into the libraries to further understand their judgment behavior as well as maybe training a sentiment model, instead of using a pre-trained one.

4.2 Backlinks

To test our second hypothesis, which states that male-man pages are mentioned more often compared to female-woman and non-binary pages, we have collected the backlinks for each of the five pages of interest from DBpedia. The backlinks were then transformed into a count variable. Overall, the male text was mentioned the most with a total of 5821 backlinks. To test the hypothesis, we have decided to use a negative binomial regression model. This model was chosen since using a dependent variable that is a count variable in an ordinary least squares regression would violate multiple OLS assumptions. For the regression, summing up the counts could've caused bias in the results, since there is only one non-binary page, while the other genders have two. Therefore, the pages were averaged together so that one observation for male-man pages, one for female-woman pages, and one for the non-binary page were generated (the output of the regression can be viewed in the notebook).

The negative binomial regression reveals that when the page changes from female-woman to male-man, this change negatively affects the number of backlinks. The change from female-woman to non-binary is also negatively associated but larger in its effect. The coefficients, therefore, show that female-woman pages are mentioned more based on the counted backlinks than male-man pages and non-binary pages. This is quite interesting since the male page alone has the most backlinks. However, the man page has 1000 backlinks less than the female page, which results in the female-woman pages having the most backlinks when averaged together. The negative binomial regression also shows that none of the coefficients are significant based on their z-values. Most likely, this is due to the low number of pages included in this study.

Concluding, this analysis led us to fail our second hypothesis, since female-woman pages are associated with the most backlinks. We recommend investigating this matter using a wider variety of pages and further retesting if new results are statistically significant.

4.3 Linking Patterns

Our third hypothesis stated that there exists a disproportionate centrality of articles about men compared to articles about women and non-binary people in terms of their linkage structure. To investigate this, we first analyzed individual

pages. For each page, we scraped the outgoing hyperlinks and created directed knowledge graphs with them. After, we iterated through the outgoing links of the new nodes and we added the edges between the existing nodes. The resulting graphs were then evaluated for their density, and we extracted the top 10 articles based on centrality, betweenness centrality, and prestige. (The constructed graphs can be viewed in the notebook).

The results (see Table 4) show that most nodes belong to the Woman page (307 nodes), and the second is the Non-Binary page (134) which has a little bit under half of the nodes compared to the nodes of the Woman's graph. Afterward, follows the Man page (131) with a similar value and then the Female page (96) and the Male page (102). In terms of density, the female and male pages are the densest, with values of 0.0451 and 0.0454, followed by the male page with 0.0272, and non-binary with 0.0232. The knowledge graph of the Woman's page has the lowest density, with a value of 0.0141. However, it has the highest number of nodes, meaning that the article has the most outgoing links.

	Man	**Woman**	**Female**	**Male**	**Non-binary**
Number of nodes	131	307	96	102	134
Number of edges	464	1324	411	468	414
Density	0.0272	0.0141	0.0451	0.0454	0.02323
Backlinks	1709	2776	5555	5821	1047
Top betweenness scores in merged graph	0.1081	0.2750	0.1323	0.0550	0.1207

Table 4: Wikipedia pages comparison

We also checked which articles had the most degree and betweenness centrality scores as well as prestige for each graph. In the Man's and Woman's articles, we see that there are topics about each gender's role and rights but in Woman's graph page "Marriage" is among the top 10 central articles unlike for Men's page. This could indicate that the Woman's article transfers a more old-style traditional picture of womanhood, where a woman was only considered to be a wife and mother, not equal to men.

To compare the graphs better, we merged the individual ones and observed the overall connections and interactions between the five Wikipedia articles. In this merged graph, we were able to compare how central each of the original articles is.

The merged graph has 652 nodes and 719 edges. As expected, the 5 initial pages have the most degree centrality scores, namely: Woman - 0.492, Non-binary gender - 0.220, Man - 0.240, Female - 0.220, Male - 0.200. The scores are decreasing in the same order as the nodes of the separate knowledge graphs of each article.

Hence, the Woman's page is more central because it has the biggest number of linked pages, while the Male page has the smallest. This finding depicts an upside-down picture of our hypothesis, where we hypothesized Male pages to be more central than their Female counterparts.

In addition, we tried to cluster the combined knowledge graph using the divisive clustering technique. Since this method is only applicable to undirected graphs, we changed the directed edges to the undirected ones. The following visualization shows the merged hyperlink graph (see Figure 1).

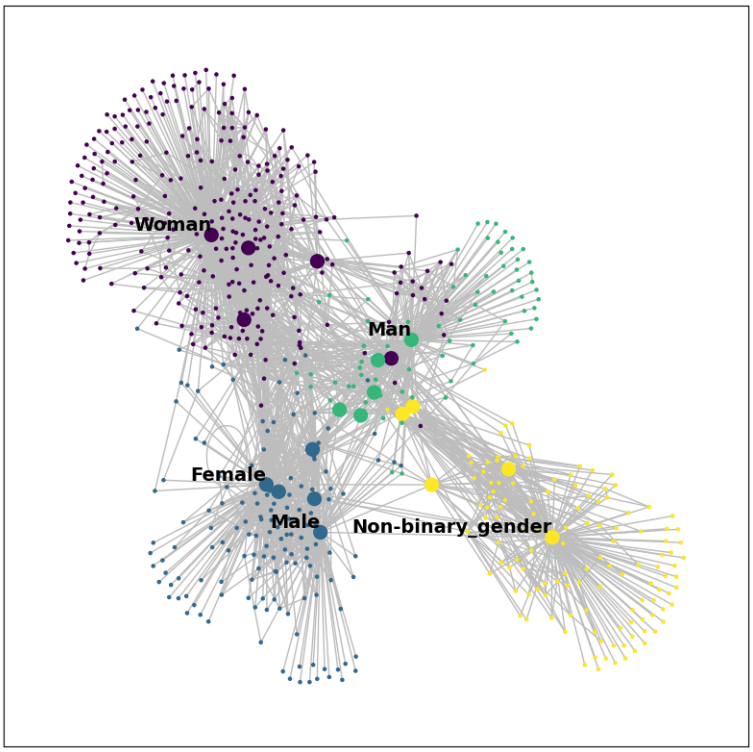

Fig. 1: Merged hyperlink graph clustered into four subgraphs. Bigger nodes indicate articles with the greatest betweenness centrality scores in each cluster.

We tested a different number of clusters, ranging from three to six, and four worked the best. Figure 1 also illustrates that four is the most logical number of clusters for the merged graph, since these four components seem well separated. Table 5 shows the top 5 articles with the betweenness centrality score for each cluster.

As one can see, each initial page has its own cluster except for Female and Male, which are in the same cluster. This can be caused by the fact that they are

biology-related terms rather than gender descriptions. This explanation seems more realistic if we look at the other articles in the same cluster.

Clusters			
Gender role	Female	Man	Gender identity
Pregnancy	Male	Prostate	Transgender
Sexism	Mammal	Semen	Non-binary gender
Woman	Sex	Human penis	Intersex
Women's rights	XY sex-determination system	Reproductive system	Gender

Table 5: Top 5 articles by betweenness centrality for each cluster. Colors indicate corresponding clusters from the graph.

The biggest cluster is the one containing the Woman page, while the smallest one is the cluster with articles about Man. It is also noticeable from the picture that the cluster with non-binary gender is farther away from the other ones, while there are a lot more connections between the other three clusters. Moreover, the topic distribution also shows that both man and woman clusters have similar topics on biology and women's rights. The topics found in the non-binary gender cluster, however, evolve more around the topic of gender identity.

An additional interesting fact that the authors identified during the clustering process is that the model splits woman's and men's clusters further when k=6. Namely, it separates two additional clusters containing the following articles: 1. Women's rights, Domestic violence, Violence against women; 2. Men's rights, Masculinity, Violence against men. Consequently, these kinds of topics are well-represented in Wikipedia articles for both genders. The female-male cluster, as the one with more biological topics and the non-binary gender cluster, as the farthest one, are stable. This analysis has led us to partially fail our third hypothesis, since Woman related Wikipedia pages were the most central, followed by males and lastly non-binary people.

5 Conclusion

In this work, the authors investigated a possible gender data gap on Wikipedia. For the purposes of this study, the concrete hypotheses were established and tested using multiple different methods such as sentiment analysis, web scraping, and knowledge graphs. The sentiment analysis of the gender pages has revealed that male pages have a more positive sentiment while female pages have a more negative sentiment, which agrees with our hypothesis. While we hypothesized the Non-Binary page to have the most neutral sentiment score, due to political correctness, the analysis has shown, that this is not the case, leading us to only partially accept our hypothesis.

The negative binomial regression model, used to test our second hypothesis, showed that Women's pages are expected to have the most backlinks, ergo being referred to the most compared to the other Wikipedia pages, while Non-Binary pages are expected to have the fewest. This finding contradicts our expectations, leading us to fail our second hypothesis.

The graph analysis showed us some interesting structural patterns of the different gender and gender-related pages. Our hypothesis, that Man's page would be more central, turned out to be wrong. The Woman's page had much more links and it stayed similar when we separated the woman's cluster from the merged graph. Also, it was notable, that the Non-Binary gender page has not many common links with the other four articles. Additionally, we concluded that the articles about violence against men and violence against women had significant centrality, while any similar article for non-binary people did not appear within any measure.

Overall, this paper has shown that there are significant differences between gender pages on Wikipedia. Nevertheless, as discussed, these differences appear to be less like the authors believed them to be. The gender gap between males and females on Wikipedia appears to be much smaller than previously anticipated, with female pages being more central and more referred to. The sentiment analysis, however, has also shown that the gap still exists. While the gap between males and females may be slowly closing, it's not the case for non-binary people. The authors, therefore, recommend further investigating the gaps on Wikipedia, to close them permanently.

6 Future Direction

Due to the limitations of our research, the authors only focused on English Wikipedia pages, however, an interesting future direction would be extending the five gender pages by adding these pages in different languages to the analysis. By comparing the new graphs for the distinct languages, interesting patterns could be expected, as the articles for these five pages can greatly differ because articles and topics vary by language.

Another intended direction could be adding more iterations in the hyperlink structure and exploring bigger topics related to the contents. The analysis of Wikipedia pages in different languages and the addition of further iterations of pages could potentially contribute to a more inclusive result for diversity and the existing gender gap on Wikipedia.

Lastly, gathering and analyzing the gender bias in Wikipedia biographies using knowledge graphs would be also interesting for future research, since the number of biographies is unbalanced regarding gender.

References

1. Perez, Caroline C. 2019. Invisible Women: Exposing Data Bias in a World Designed for Men. Random House.
2. Kandek, Barbara (2023): Closing the gender gap: Women in Red's efforts to add more women to Wikipedia.
3. Khanna, Ayush (2012): Nine out of ten Wikipedians continue to be men: Editor Survey