PAP

Hannah Schweren

2024-01-22

Contents

1. Summary: **1**

2. Motivation and background: **2**

3. Introduction: **2**

4. Research question: **3**

5. Data and methods: **3**

Bibliographie: **4**

1. Summary:

The project aims to measure gender bias in politicians' Wikipedia biographies on an international scale. Wikipedia ranks as the seventh most visited website worldwide in 2022 (statista), with over 4 billion unique global visitors per month (statista). Many individuals rely on it to quickly access information about celebrities, artists, and politicians. Therefore, it is crucial for Wikipedia's content to maintain neutrality and avoid reinforcing societal biases. One could argue that this importance is heightened within the subgroup of politicians, as Wikipedia's representation of political figures influences citizens seeking information before elections in democratic processes. Previous research has revealed a noticeable lexical gender bias in Wikipedia biographies. I specifically focus on politicians' biographies to investigate whether general findings can be applied to this specific subgroup. Additionally, much of the existing research concentrates on English articles, whereas I analyze all politicians' texts in their respective national languages. This approach allows me to compare the extent of gender bias across different countries. My research questions are as follows:

RQ1: Can the results of previous research on lexical gender bias in biographies be applied to politicians' biographies? depending on further capacity, a second question will be examined: RQ2: How do the results vary in an international context?

I plan to address these questions using available data from the "Comparative Legislator Database," (Göbel and Munzert 2022)which includes information about politicians, their Wikipedia names, and additional details such as traffic and edits from nine countries ('aut,' 'can,' 'cze,' 'esp,' 'fra,' 'deu,' 'irl,' 'sco,' 'gbr,' 'usa_house,' or 'usa_senate'). Initially, I will retrieve all articles in their original language and then assess the extent of gender bias in the articles, first using simple descriptive indicators and secondly either employing mutual pointwise information or a machine learning approach.

2. Motivation and background:

This research question is particularly intriguing to me, as I've been interested in gender equality topics for several years now. During my studies, I've delved into the concept of the child penalty and worked on addressing the gender care gap in my student job. These experiences have motivated me to explore gender-related inequalities further. In the scope of my master's thesis at Hertie School, I aim to delve deeper into such issues in the online world and utilize the techniques I've acquired during my studies. Thus, this research topic is very exciting for me, and I'm curious to dive deeper into gender inequalities in the political area. As a public policy student, the category of politicians is specifically interesting to me. Inequalities in this sphere are particularly noteworthy as politicians represent all citizens and have a real-life impact. So, misrepresentation in the form of gender bias can be seen as problematic for democratic processes, as citizens use online platforms like Wikipedia to inform themselves, expecting a neutral voice. I expect to acquire skills in the field of text data processing, which will be advantageous for my future career. I aspire to work in the political sphere after my graduation, and a lot of political data is in the form of text. Thus, I hope to gain some knowledge that will be useful for my next steps after university.

3. Introduction:

The area of research focussing on gender bias in wikipedia has developed in the last few years and is mainly focussed on English articles and general biographies (not focussing on one specific category of people). The following 3 papers are of a main interest, when dealing with lexical gender bias on wikipedia:

1. "Women through the glass ceiling: gender asymmetries in Wikipedia" (Wagner et al. 2016) aim at assessing potential gender inequalities in English Wikipedia articles along different dimensions using the The DBpedia dataset. Adopting an open vocabulary approach, they consider n-grams with $n \leq 2$ 2 to encompass multi-word concepts. The analysis involves exploring the association between the top 200 n-grams for each gender and the four topics (gender, relationship, family, or other), with Pointwise Mutual Information used for ranking the n-grams for men and women. Among other findings, they achieve to show that family-, gender-, and relationship-related topics are more present in biographies about women. Which is a clear indicator for a lexical bias.

2. "First Women, Second Sex: Gender Bias in Wikipedia" (Graells-Garrido, Lalmas, and Menczer 2015)deal with the question wether there is a gender bias in user-generated characterizations of men and women in Wikipedia and if so, how to identify and quantify it. For this, they use the DBPedia 2014 dataset and The Wikipedia English Dump of October 2014. They rely on several approaches to estimate the gender bias in the articles, among others they use Pointwise mutual information and find that "Sex-related content is more frequent in women biographies than men's, while cognition-related content is more highlighted in men biographies than women's".

3. "It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia" (Wagner et al. 2021) As in the previous paper, Wagner et al. assess potential gender inequalities in Wikipedia articles along different dimensions (coverage bias, structural bias, lexical bias and visibility bias). In this paper they also include several language editions and compare their results with the Gender Inequality Index of the World Economic Form (WMF) (Forum, n.d.) to draw a connection between bias in the offline and online world. The authors use collections of notable people as reference datasets, crawling the articles for the dataset's individuals. To assess the amount of lexical bias, they, again, use an open vocabulary approach. Instead of an analysis of the PMI, as in the previous paper, this time they use tfidf scores to train a Naive Bayes classifier. The classifier employs log likelihood ratios (L(word, g)) to assess which words are most effective in discerning the gender of the person mentioned in an article.The lexical analysis shows "that articles about women tend to emphasize the fact that they are about a women (...) while articles about men don't contain words like "man", "masculine" or "gentleman"."

4. Research question:

My Research Questions are the following:

QR1: Can results of previous research concerning lexical gender bias in biographies be applied to the subgroup of politican's biographies?

RQ2: How do the results vary in the international context?

The approach reflects previous research on this topic, specifying the examined group, focussing on politicians, instead of generic biographies on wikipedia. I want to see if the background of being a politician maybe changes the extent of gender bias.

5. Data and methods:

Data:

I am using the data of the Comparative legislature Database (Reference), containing information (including sex, wikititle, traffic, edits etc.) on more than 67,000 contemporary and historical legislators. To gather the according wikipedia articles of the politicians, I use the wikipdeiR package. I use a subset of politicians that don't have a date of death recorded in the data and thus are contemporarian politicians, as I am rather interested in the gender bias towards politicians that are known as acting politicians, not as historical figures. The data includes 9 available countries: Canada, Germany, France, Czech Republic, USA, Ireland, Scotland, Austria and Spain.

Method:

1. Decriptive analyis

The start of the analysis will be a simple descriptive analysis of the available data. Figures like the length of the articles have already been used in previous research to get an impression of possible bias (Reference reagle, Graells-Garrido). For this analysis, I plan to use the average monthly page traffic as a matching variable to only include comparable female and male politicians. This serves the purpose of reducing the influence of possible confounders - the biggest confounder in this case lies in the fame of certain politicians and monthly page views seems like a good proxy to reduce this confounding factor. Further, I want to compare not only the article length but also the number of page edits of the female and male politicians.

Further analysis

This descriptive analysis will be followed by a second technique to detect gender bias, focussing on the kind of bias that is pretty common based on previous research (reference, wagner), namely lexical bias. Lexical bias means that language is used differently when talking about men/women (reference. Lakoff RT (1973) Language and woman's place. Lang Soc 2(1):45-80). I want to propose two possible methods to analyse the extend of lexical bias in politicians biographies. Depending on the complexity of the method, a focus on 1 or few countries may be advantageous (e.g. focussing on german politicians and or compare to 1/few other countries) as the methods have to be applied to each country individually.

Option 1

The first proposed method to detect gender bias is inspired by Wagner (reference) and has been used by other authors as well to assess lexical bias (Graells-Garrido). For this method, Pointwise Mutual Information (PMI) is used to find out which words are strongly associated with articles of females/males. (Reference Kenneth Ward Church and Patrick Hanks. "Word association norms, mutual information, and lexicography". In: Computational linguistics 16.1 (1990), pp. 22–29.). The second method uses machine Learning techniques to detect gender bias in the biographies.

Previous research has often focused on the Wikipedia introduction text to detect bias. In my analysis, I plan to include the whole text as I expect the introduction texts for politicians to be rather simple, similar

and short, whereas the other sections include more relevant information for the analysis. First step of this approach is to tokenize the wikipedia articles and to create the vocabulary, containing "gender", "word" and "Number of biographies" containing this word. Next, stopwords are removed, and only words that are present in both genders are included. After creating a df that contains all common words of female and male politicians, witht he respective frequency for men/women, the PMI scores for the vocabulary are calculated. PMI is expressed as: $\text{PMI}(c, w) = \log$

C represents the gender and w represents the word. Further, the PMI score needs to be normalized for further analysis. PMI helps identify words that tend to occur together with a specific gender term more frequently than expected by chance - so, higher PMI values indicate a stronger association between a word and a specific gender. The next step is thus to sort the resulting PMI values decreasingly for each gender. Following the approach tested by Wagner et al. (reference), the top 200 words are manually annotated and put in categories: Family, Relationship, Gender, Other. To assess differences between the genders, a chi-square test is used for the categories of men and women. Further, word Clouds can be used to show the results of the analysis for each gender.

Option 2

The second possible method is a machine learning approach as applied before by Wagner et al (reference) or bruns (reference) before. Again, it would be an open vocabulary approach. Firstly, it includes stemming the words of the wikipedia articles and compute Term Frequency-Inverse Document Frequency scores (tfidf Scores), used to evaluate the importance of a word in a document relative to a collection of documents (corpus). These Scores are used to train a Naive Bayes classifier, which again, serves the purpose of identifying words that effectively differentiate the gender of the individual discussed in an article.

International analysis

As I am particularly interested in the german case, as a german speaker and having gained experiences in gender inequalities in Germany, I will firstly apply the respective methos to the german language. If there is enough capacity to apply the method to several countries, I will continue with english and french speaking countries. An alternative would be to group all countries with the same language (e.g. Germany and Austria) but as I ideally aim to compare my results on a country level, I decided against this option. If I can apply my method to several countries, I aim at comparing the country level of gender equality, meausured by the Gender Inequality Index of the World Economic Form (WMF) (Schwab et al. 2013 Reference) with the results of my lexical bias analyis to see, if there is a connection. This way of comparing bias in the offline world and the bias on Wikipedia has been conducted by Wager et al (reference)before. .....Comparing our results with the Gender Inequality Index of the World Economic Form (WMF) (Schwab et al. 2013) shows that a positive correlation exists between the bias in the offline world and the bias on Wikipedia. However, one needs to note that it is difficult to compare our Wikipedia based gender bias rankings of languages with the ranking of countries according to the gender inequality index since countries where the same language is predominantly spoken often reveal very different positions in the WMF ranking. We use the weighted average of the WMF rank positions of countries where the same language is spoken3 and weight countries by the size of the internet population4.

Bibliographie:

Forum, World Economic. n.d. "Global Gender Gap Report 2023."

Göbel, Sascha, and Simon Munzert. 2022. "The Comparative Legislators Database." *British Journal of Political Science* 52 (3): 1398–408. https://doi.org/10.1017/S0007123420000897.

Graells-Garrido, Eduardo, Mounia Lalmas, and Filippo Menczer. 2015. "First Women, Second Sex: Gender Bias in Wikipedia." In *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*, 165–74. https://doi.org/10.1145/2700171.2791036.

Wagner, Claudia, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2021. "It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia." *Proceedings of the International AAAI Conference on Web and Social Media* 9 (1): 454–63. https://doi.org/10.1609/icwsm.v9i1.14628.

Wagner, Claudia, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. "Women Through the Glass Ceiling: Gender Asymmetries in Wikipedia." *EPJ Data Science* 5 (1): 5. https://doi.org/10.1140/epjds/s13688-016-0066-4.