Wikigender: A Machine Learning Model to Detect Gender Bias in Wikipedia

Natalie Bolón Brun
École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
natalie.bolonbrun@epfl.ch

Sofia Kypraiou
École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
sofia.kypraiou@epfl.ch

Natalia Gullón Altés
École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
natalia.gullonaltes@epfl.ch

Irene Petlacalco Barrios
École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
irene.petlacalcobarrios@epfl.ch

ABSTRACT

The way Wikipedia's contributors think can influence how they describe individuals resulting in a bias based on gender. We use a machine learning model to prove that there is a difference in how women and men are portrayed on Wikipedia. Additionally, we use the results of the model to obtain which words create bias in the overview of the biographies of the English Wikipedia. Using only adjectives as input to the model, we show that the adjectives used to portray women have a higher subjectivity than the ones used to describe men. Extracting topics from the overview using nouns and adjectives as input to the model, we obtain that women are related to family while men are related to business and sports.

KEYWORDS

Wikipedia, gender bias, topic bias, linguistic bias, logistic regression, natural language processing, classification

Wikipedia has become a very popular source of information. By November 2019, the number of entries in the English Wikipedia was above 5M [7], and it is increasing every day at a rate of 500 entries on average.

In previous studies, Wagner et al. [8] show how gender bias manifest in Wikipedia in the way women and men are portrayed. In a different study, Graells-Garrido et al. [2] show that women biographies are more likely to contain sex-related content. Along with these studies, several other researchers have analyzed topic-related bias in the way women are portrayed, but we can also take a look from a linguistic perspective.

Linguistic bias is defined as a systematic asymmetry in word choice that reflects the social-category cognition that is applied to the described group or individual(s) [5]. We want to analyze how men and women are portrayed and, more specifically, the adjectives used to describe them to spot a possible bias from a linguistic perspective. To do so, we will use the overview of the biographies in the English Wikipedia together with other characteristics of the people we are analyzing.

The overview section of the biographies is the first information to encounter by the reader. According to Wikipedia, it should be written with a neutral point of view and contain a summary of the most relevant content [11]. Given this description, we center our study in this section, where the expected non-relevant content is minimal.

In this work, we model the problem as a prediction task to infer the gender of the person described by using the set of words used in the article. We base our prediction on a logistic regression model that provides interpretable insights on the importance of its features. The difference is manifested as the presence of different words given the sex of the person being described. This bias is also studied along with different occupations. Finally, we analyze those words that appear as most predictive for each gender and quantify their subjectivity and strength.

Results show that there is actually a distinction in the usage of words based on gender. As was already shown in [9], in terms of topics, women tend to be more related to family and marriage, while men are usually linked to sports and politics topics. Furthermore, results show that women tend to be described using more strongly subjective positive adjectives, while for men, there is a predominance of weakly subjective negative adjectives.

1 RELATED WORK

Gender bias on Wikipedia is a topic that has been widely explored from different perspectives [2–4].

The existence of a gender gap from the editors' perspective was already studied by Hill and Shaw [3], showing a predominant contribution of men who represent more than 70% of the authors' community.

In his work "First Women, Second Sex" [2], Graells-Garrido et al. explore the differences introduced by gender from different perspectives. From the linguistic point of view, they introduce a method to relate topics and gender by exploring the most important n-grams for each gender. Their results show a topical bias given that women are highly related to marriage and family, whether men are linked to sports and politics. These differences also show up in different language editions.

In [8], Wagner et al. assess the extent to which Wikipedia suffers from potential gender bias. Among others, they explore lexical bias and, by computing log-likelihood ratios, they show that female articles tend to describe romantic relationships and family-related issues much more frequently than male ones in most Wikipedia language editions.

In "Women through the glass ceiling: gender asymmetries in Wikipedia" [9], Wagner et al. analyze different dimensions of the gender gap in Wikipedia. They use Pointwise Mutual Information

(PMI) to show that words related to gender, relationships, and family are more prominent for women than men. On the contrary, words associated with men are mainly related to politics and sports. Additionally, for the linguistic bias, they reveal that more abstract terms are used for positive aspects of men’s biographies and negatives aspects of women’s biographies. This is calculated by computing the ratios of abstract positivity and negativity as the number of positive/negative adjectives over the number of positives/negative words. Apart from topical and linguistic bias, they also show a bias in other dimensions such as notability and structural properties, finding that women in Wikipedia are more notable than men and that structural differences in terms of meta-data and hyperlinks have consequences in information-seeking activities.

The linguistic bias on collaborative crowdsourcing biographies has also been expanded beyond Wikipedia, on the IMDB database by Otterbacher [6]. She also uses the Semin and Fiedler’s Linguistic Category Model (LCM) (Semin and Fiedler 1988) to analyze the biographies. The LCM model classifies terms (like nouns or adjectives) on a scale from abstract to concrete. The more abstract language implies stability over time and generalizability across situations.

2 DATA

2.1 Dataset

For the following study we use data from the following sources:

- Wikipedia Human Gender Indicators (WHGI) dataset [10]. This dataset contains all the biographic articles from all Wikipedia editions. The version used is that from November the 4th, 2019. From this dataset we extract all biographies appearing in the English Wikipedia together with the corresponding gender. Other data such as the unique identifier of the entry in Wikidata or the occupation of the person are also gathered from this dataset.
- Wikidata dataset. It allows us to link the people we want to explore with their corresponding articles in the English Wikipedia.
- Wikipedia dataset. Given the previous steps, we are able to extract the biographies of interest in our study. From them, we will keep only the overviews for the corresponding analysis.

2.2 Gender and Occupation

The dataset used is restricted to those entries matched with either male or female gender. In total, nine different gender categories appear along with the whole dataset but those cases not stated as either male or female such as *transgender* or *non-binary* gender represent less than 1% of the total number of entries and are not considered in the study. Finally, a total of 1,383,430 articles are used and only 16.58% of them correspond to female entries.

The dataset used is very diverse in terms of occupations. A total of 5,891 different occupations show up with a very different weight in the total representation. We limit the study to the 100 most common occupations since they cover 78.58% of all the biographies. Moreover, we group them into 10 different fields to extract more meaningful information. As shown in Figure 1, the most common occupation is Sports (i.e. footballer, midfielder, etc.) followed by Artist (i.e. actor, singer, painter, etc.) and Politics (i.e. senator, president, etc.). A deeper analysis of the gender by occupation shows again a great disparity with men outnumbering women in all fields except the Model category where the ratio is 5 female entries per each one corresponding to a male.

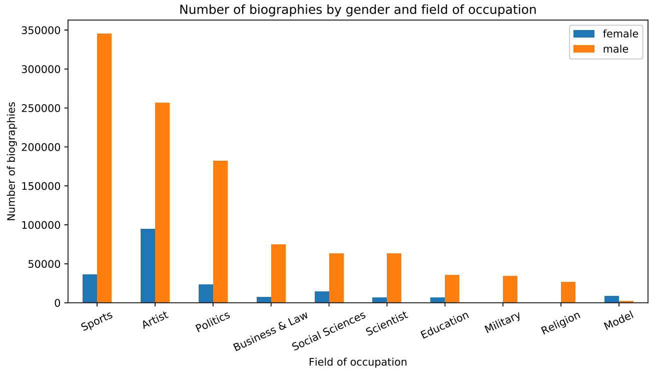

Figure 1: Distribution of biographies per occupation and gender

2.3 Subjectivity introduced through the usage of adjectives

We analyze the subjectivity introduced in the overviews through the usage of adjectives. For this purpose, we make use of the Subjectivity Lexicon version used in [12]. This allows to determine the degree of subjectivity of the vocabulary and if the given adjectives are usually employed with a positive or negative connotation.

A first exploration shows that the distribution of strengths and subjectivity of the adjectives is very similar for both genders. In both cases, the majority of adjectives are weakly subjective and positive while those adjectives neutral and strongly subjective are the least present. The final distribution is shown in Figure 2.

In terms of strength, the weakly subjective adjectives account for 68% of the total adjectives. In terms of subjectivity, the positive adjectives represent 64% of the adjectives, the negatives represent the 20% and the neutral ones the 15%.

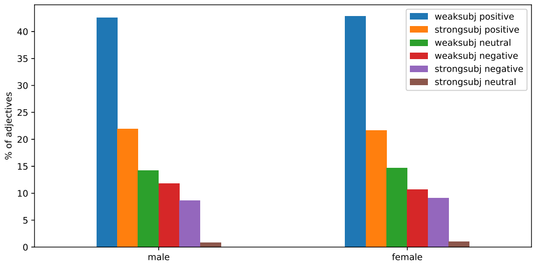

Figure 2: Percentage distribution over the strength and subjectivity of the adjectives

If we analyze deeper the most common adjectives for each gender, we can see that three of them (best, high, active) are common for both genders. Moreover, most of them are weakly subjective and have a positive connotation.

	MALE			FEMALE		
	adjective	**sentiment**	**subjectivity strength**	**adjective**	**sentiment**	**subjectivity strength**
1	**active**	positive	weak	**best**	positive	**strong**
2	**major**	neutral	weak	**high**	neutral	weak
3	**best**	positive	**strong**	**active**	positive	weak
4	**high**	neutral	weak	**popular**	positive	weak
5	**famous**	positive	weak	**long**	negative	**strong**

Table 1: Most common adjectives for each gender and their subjectivity and strength.

3 METHODOLOGY

In order to verify the presence of a bias linked to the gender in the overviews of the studied biographies, we develop a model using logistic regression that takes as input a vectorial representation of the text and aims to predict the gender of the person described.

We first start by encoding the text to obtain a vectorial representation. This process begins by removing stop words such as pronouns. Then we follow with the definition of a vocabulary. This vocabulary is composed by a combination of the top 100 most common words (words being adjectives or adjectives and nouns depending on the model) for each gender from which we obtain a final set of 114 words in the case of adjectives and 132 words when including also nouns. From the vocabulary, we encode the texts in a binary vector in which each entry of the vector represents the presence of a word from the vocabulary in the text.

The model is build using Logistic Regression. Given the wide diversity of the data in terms of occupation and the imbalanced character of it, we first balance the data per occupation, matching the same number of entries per gender for each occupation. A total of 187,698 entries are then included in the balanced dataset and then split into train and test sets in a proportion 70%-30%.

To verify the robustness and estimate the generalization error, the model is fit and test 50 times. In each case the original dataset is sampled to obtained new balanced version.

4 RESULTS

The bias is studied in two ways: a first model that only uses adjectives and a second model that includes adjectives and nouns.

4.1 Model using adjectives

The model that only uses adjectives achieves an accuracy of 54.6 ± 0.002% in the task of determining the gender based on the encoded text. Given that the dataset has been previously balanced, obtaining an accuracy higher than 50% shows a difference based on gender in the way people are described, i.e. the words used to portray men and women. Apart from discerning the existence of a bias, we can extract from the model those words that are highly indicative of each gender. The adjectives most correlated with female biographies are beautiful, profit, cross, creative and romantic, while the ones most correlated with males are offensive, certain, hard, defensive and diplomatic.

The analysis of the obtained words using the Subjectivity Lexicon [12] shows that adjectives related to men are weakly subjective and most of them have a negative connotation, whereas the ones related to women are mostly strongly subjective and have a positive connotation (see Table 2). Therefore, the overviews portraying females are more likely to contain subjectivity.

	MALE			FEMALE		
	adjective	**sentiment**	**subjectivity strength**	**adjective**	**sentiment**	**subjectivity strength**
1	**offensive**	negative	weak	**beautiful**	positive	**strong**
2	**certain**	neutral	weak	**profit**	positive	weak
3	**hard**	negative	weak	**cross**	negative	**strong**
4	**defensive**	negative	weak	**creative**	positive	**strong**
5	**diplomatic**	positive	weak	**romantic**	positive	**strong**

Table 2: Most predictive adjectives for each gender and their subjectivity and strength.

4.2 Linguistic bias per field of occupation

Once we have explored the general linguistic bias, we explore it for each field of occupation to see if there are some occupations with a higher bias than others. We measure the error by randomly balancing the data and computing the accuracy for each model which is fit using data only from the corresponding field of occupation. The fields of Military, Model, and Religion are the only ones where we cannot state that there exists a bias since the accuracy is not significantly above 50%. These fields are also the ones with fewest data, so this might be the reason behind the high variability of the results. The accuracy for the other fields of occupations is approximately the same, between 55% and 60%. The one with the highest bias (i.e. highest accuracy) is Business and Law.

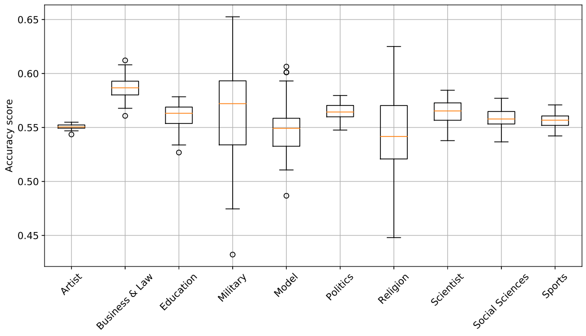

Figure 3: Results of accuracy from the model fit by occupation

4.3 Model using adjectives and nouns

A second model using adjectives and nouns is now analyzed to study the effect of the introduction of nouns in the bias. Again, we follow the procedure described in Section 3. Nevertheless, to cope with those words that include references of gender, we substitute them by a neutral form (e.g. actor and actress are substituted by act*).

The presence of nouns in the text is larger than the one of adjectives which leads to a vocabulary formed mainly by nouns. Among the top 100 most common words (using adjectives and nouns), nouns represent 92% in the case of females and 91% for males. The final vocabulary is composed of 132 words and most of them are nouns.

	MALE	FEMALE
1	footballer	person
2	war	marriage
3	officer	model
4	musician	dancer
5	football	midfielder

Table 3: Most predictive words for each gender

In this case, the accuracy achieved by the model rises to 62.9% ± 0.002. In this case, the words most correlated with females are person, marriage, model, dancer, and midfielder, while the ones most correlated with males are football, musician, officer, war and footballer. We should mention that "person" includes the words man and woman, but since the words themselves indicate the gender we transform them into a neutral gender one. Words such as spouse and child, which are related to family, also have a positive coefficient which means that they are more predictive for women than men.

4.4 Words representation

In order to verify the results, we analyze the presence of the most predictive words along with the biographies. As shown in Figure 4, those adjectives more correlated to female biographies are more frequent in this group of articles and the same effect occurs in male articles.

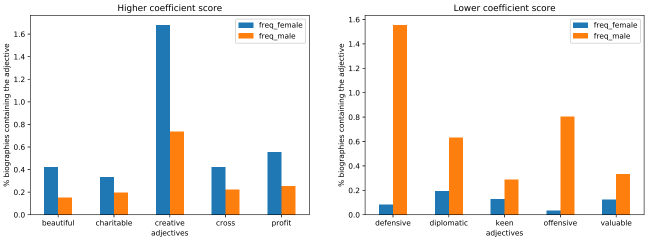

Figure 4: Frequency of appearance of most predictive adjectives along the biographies

In the case of the model developed using nouns and adjectives, the majority of the most predictive words for women and for men are more likely to appear in their corresponding overviews. However, the word midfielder is more likely to appear in a man's overview 5 than in a woman's one although it is among the most predictive words for women. This happens because this word is highly correlated with other words (football and footballer) that are predictive for males, as it can be seen in Figure 6

4.5 Topic extraction

Once we know the adjectives and nouns most predictive for males and females, we analyze them using Empath, a tool for analyzing text across lexical categories [1]. Using this library, we can extract the categories associated with the words highlighted from

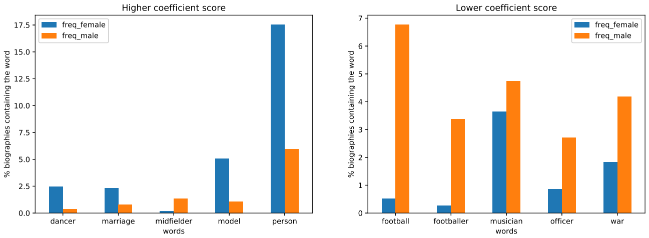

Figure 5: Frequency of appearance of most predictive words along the biographies

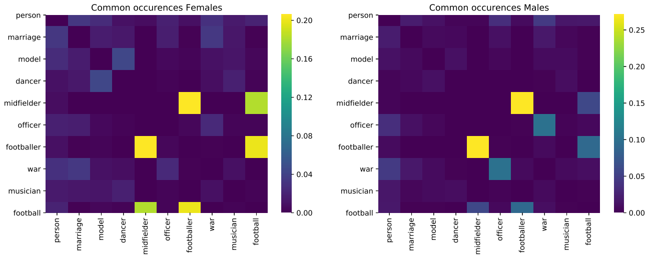

Figure 6: Correlation between words

the model. We analyze the categories for both the results using the model with only adjectives and the ones using nouns and adjectives.

The results with only adjectives show that women are portrayed as wealthy and men as heroic. Nevertheless, the analysis using both nouns and adjectives results in more insightful results since we are able to extract topics related to them. In this case, we can observe that women are related to family in the first place and other topics related to art (i.e. reading, music), whereas men are mostly related to business and sports.

5 CONCLUSION AND FUTURE WORK

In this work, we presented a different way to measure linguistic and topic bias based on gender in Wikipedia biographies. This new system based on a logistic regression model aims to predict the gender of the person described based only on the appearance of different words in the text. The model also allows us to extract those words that are more relevant for each gender and further analyze them.

The analysis performed cover from subjectivity introduced by the usage of different adjectives to the extraction of topics based on the most correlated nouns and adjectives for each gender.

These results show the existence of a difference in the usage of words and topics based on gender. Although these differences may be subtle and could be hidden inside such a great amount of information that Wikipedia constitutes, it is essential to highlight that they should not be normalized, and further steps to limit this issue should be taken.

Different areas are left as future work after this study. One is the introduction of other features such as year of birth or length of the biographies in order to balance the dataset using propensity score, and therefore eliminate as many confounding factors as possible from the study. Another step to be taken is to extend the study to the whole biographies and not just the overview to see to which extent this bias is present along with the text. In terms of language, extending the study by including other Parts of Speech such as verbs could open a new path to explore. Finally, expanding the work to other languages would be a great way to determine how cultures and their usage of words influence this bias.

ACKNOWLEDGMENTS

We thank Tiziano Piccardi who provided insight and expertise that greatly assisted the research, although he may not agree with all of the interpretations of this paper.

REFERENCES

[1] Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* 4647–4657.

[2] Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. 2015. First Women, Second Sex: Gender Bias in Wikipedia. *CoRR* abs/1502.02341 (2015). arXiv:1502.02341 http://arxiv.org/abs/1502.02341

[3] Benjamin Mako Hill and Aaron Shaw. 2013. The Wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PloS one* 8, 6 (2013).

[4] Andrew Lih. [n.d.]. Opinion | Can Wikipedia Survive? https://www.nytimes.com/2015/06/21/opinion/can-wikipedia-survive.html

[5] Oxford Research Encyclopedia of Communication. 2019. *Linguistic Bias.* Retrieved February 13, 2020 from https://oxfordre.com/communication/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-439

[6] Jahna Otterbacher. 2015. Linguistic Bias in Collaboratively Produced Biographies: Crowdsourcing Social Stereotypes? (2015). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10539/10513

[7] Wikipedia the free encyclopedia. 2020. *Wikipedia:Statistics.* Retrieved February 12,2020 from https://en.wikipedia.org/wiki/Wikipedia:Statistics#Page_views

[8] Claudia Wagner, David García, Mohsen Jadidi, and Markus Strohmaier. 2015. It’s a Man’s Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *CoRR* abs/1501.06307 (2015). arXiv:1501.06307 http://arxiv.org/abs/1501.06307

[9] Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science* 5, 1 (March 2016), 5. https://doi.org/10.1140/epjds/s13688-016-0066-4

[10] Wikidata Human Gender Indicators (WHGI). 2019. *Gender Index Data.* Retrieved December 13,2019 from http://whgi.wmflabs.org/snapshot_data/2019-10-07/gender-index-data-2019-10-07.csv

[11] Wikipedia. 2020. Lead paragraph — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Lead%20paragraph&oldid=941479766. [Online; accessed 19-February-2020].

[12] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT ’05).* Association for Computational Linguistics, USA, 347–354. https://doi.org/10.3115/1220575.1220619