Is Wikipedia succeeding in reducing gender bias? Assessing changes in gender bias in Wikipedia using word embeddings

Schmahl, Katja Geertruida; Viering, Tom Julian; Makrodimitris, Stavros; Naseri Jahfari, Arman; Tax, David; Loog, Marco

DOI
10.18653/v1/2020.nlpcss-1.11

Publication date
2020

Document Version
Final published version

Published in
Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science

Citation (APA)
Schmahl, K. G., Viering, T. J., Makrodimitris, S., Naseri Jahfari, A., Tax, D., & Loog, M. (2020). Is Wikipedia succeeding in reducing gender bias? Assessing changes in gender bias in Wikipedia using word embeddings. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science* (pp. 94-103). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.nlpcss-1.11

Is Wikipedia succeeding in reducing gender bias? Assessing changes in gender bias in Wikipedia using word embeddings

K.G. Schmahl[1], **T.J. Viering**[1], **S. Makrodimitris**[1], **A. Naseri Jahfari**[1], **D. M. J. Tax**[1] and **M. Loog**[12]
[1]Delft University of Technology, [2]University of Copenhagen
katjaschmahl@hotmail.com, {t.j.viering, s.makrodimitris, a.naserijahfari, d.m.j.tax, m.loog}@tudelft.nl

Abstract

Large text corpora used for creating word embeddings (vectors which represent word meanings) often contain stereotypical gender biases. As a result, such unwanted biases will typically also be present in word embeddings derived from such corpora and downstream applications in the field of natural language processing (NLP). To minimize the effect of gender bias in these settings, more insight is needed when it comes to where and how biases manifest themselves in the text corpora employed. This paper contributes by showing how gender bias in word embeddings from Wikipedia has developed over time. Quantifying the gender bias over time shows that art related words have become more female biased. Family and science words have stereotypical biases towards respectively female and male words. These biases seem to have decreased since 2006, but these changes are not more extreme than those seen in random sets of words. Career related words are more strongly associated with male than with female, this difference has only become smaller in recently written articles. These developments provide additional understanding of what can be done to make Wikipedia more gender neutral and how important time of writing can be when considering biases in word embeddings trained from Wikipedia or from other text corpora.

1 Introduction

Word embeddings are vectors that represent the meaning of words and their relation. They are the cornerstone of many NLP techniques. For example, word embeddings can be used to search in documents, to analyze sentiment and to classify documents [Mikolov et al., 2013a, Nalisnick et al., 2016, Parikh et al., 2018, Jang et al., 2019]. These embeddings are typically created using unsupervised learning from a large corpus of text [Krishna and Sharada, 2019].

Large corpora of text used for training word embeddings may contain stereotypical biases. Word embeddings can then inherit these biases [Mikolov et al., 2013a, Caliskan et al., 2017, Jones et al., 2020]. For example, stereotypical words such as 'marriage' can be more strongly associated with female words than male words. In fact, changes in word embedding can be useful for detecting minor changes in the meaning of words at small time scales [Kutuzov et al., 2018].

Biases in word embeddings may, in turn, have unwanted consequences in applications. Bolukbasi et al. [2016] show that when embeddings are used to improve search results, biased embeddings can lead to biased results. As an example, scientific research with male names may be ranked higher if male names have a stronger association with the scientific search words [Bolukbasi et al., 2016].

Another example of a downstream application with unwanted gender bias consequences is machine translation. When translating a sentence from a language with a gender neutral pronoun to English, a sentence about a nurse may be translated with a female pronoun while a sentence with the word engineer may be translated with a male pronoun [Prates et al., 2019]. Such stereotypical translations can be avoided by using a more gender neutral embedding [Font and Costa-Jussa, 2019].

Bolukbasi et al. [2016] have already proposed a method for debiasing word embeddings. However, it has been hypothesized that debiasing covers up biases instead of removing them [Gonen and Goldberg, 2019]. Stereotypical words remain clustered in the debiased embeddings and thus there is still a risk for algorithmic discrimination [Gonen and Goldberg, 2019]. A more robust debiasing procedure is yet to be proposed.

Gender bias, as measured in word embeddings trained on books, has been shown to decrease over time up to the year 2000 [Jones et al., 2020, Garg et al., 2018]. Whether the decreasing trend has con-

Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science, pages 94–103

https://doi.org/10.18653/v1/P17

tinued in more recent years has not been tested. If bias has continued to decrease, a straightforward way to obtain less biased word embeddings would be to train word embeddings on more recent corpora of text. To investigate this issue, we will measure gender bias in one of the largest openly available text corpora: Wikipedia.

Wagner et al. [2015] already showed the presence of gender bias in Wikipedia. The editors of Wikipedia have actively tried to reduce this bias since 2013 [Wikipedia contributors, 2020a]. Our research can be used to evaluate the effectiveness of these efforts, and may inspire new strategies to reduce bias further. Towards that end, we will answer the question: 'How does gender bias in word embeddings from Wikipedia develop over the years 2006-2020?'.

Contributions: 1. We extend the work of Jones et al. [2020] and Garg et al. [2018] by looking at more recent years and applying their methods to the corpus of Wikipedia.

2. Our work provides insight in how gender bias has developed in Wikipedia using four categories. So far, most research into this is static. Our research shows to what extent the efforts of Wikipedia editors were successful, while also providing possible improvements on their current strategy.

3. We illustrate that year of retrieval is important for gender bias in the word embeddings from Wikipedia. If gender neutrality w.r.t. a domain is important, our results suggest what year to use.

2 Gender Bias in Wikipedia

In 2011, a big survey on the demographics of Wikipedia editors showed that less than 15% of Wikipedia editors are female [Collier and Bear, 2012]. This led to further investigations into the impact on content of Wikipedia considering different dimensions of gender bias. Two important dimensions of gender bias as researched by Wagner et al. [2015] are coverage bias and lexical bias.

Coverage bias means that notable women are not covered as well as notable men. For example, a smaller percentage of notable women have their own Wikipedia page or these pages may be less extensive. Wagner et al. [2015] looked at three data sets of notable people and found no coverage bias.

However, later research by Wagner et al. [2016] did show a small glass ceiling effect. Google search trends were used to assess the notability of people covered on Wikipedia. Women on Wikipedia were found to be more notable than men on average, which suggests that women have to be more notable to be covered on Wikipedia. The efforts of Wikipedia editors have mostly focused on this coverage bias, specifically by making lists of missing notable women and creating articles for these women [Wikipedia contributors, 2020b]. In terms of gender associations in word embeddings, this may have caused words that are commonly used in these biographies to have become more female associated.

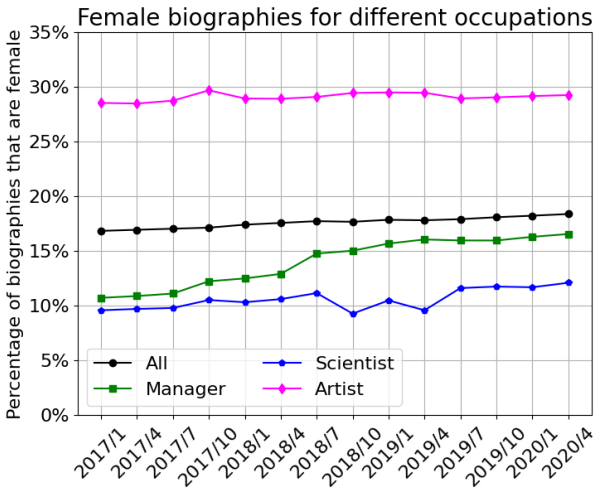


Figure 1: The percentage of biographies of women on Wikipedia for different occupations since 2017. Data from Envel Le Hir [2017-2020].

Lexical bias relates to the words used on pages written about women and men. Wagner et al. [2016] found two significant differences. Words related to family and relationships are more present in female articles compared to male articles. An article about a divorced person is 4.4 times more likely to be about a woman. The second difference is a stronger emphasis on gender. Articles about women contain more words that are gender-specific, such as 'female' or 'woman'. This can cause biases in the word embeddings. When biographies about women for example contain phrases as 'female scientist', whereas men are referred to as 'scientist', the word scientist would be more closely associated to female, despite there being both male and female scientists.

Besides this, there has also been research to the development of the gender proportion in the Wikipedia biographies. This has been recorded since 2014 and since 2017 this has also been measured by occupation (see Figure 1) [Konieczny and Klein, 2018].

The biggest change can be seen for the occupation 'manager', for which the percentage of female biographies increased with more than 5% in the last 3 years. However, this is still below average. The occupation artist has a female percentage far above average with almost 30%. Furthermore, the overall fraction of female biographies has increased steadily towards around 18% [Envel Le Hir, 2017-2020]. Thus matters are improving, but women are generally still less represented in Wikipedia.

3 Word Embedding Association Test

As proposed by Caliskan et al. [2017], we use the Word Embedding Association Test (WEAT) to quantify gender bias. This test uses four categories that are considered stereotypical towards gender: Arts, Science, Family and Career [Caliskan et al., 2017]. These categories have shown significant bias towards male or female words in embeddings from Google News corpora [Mikolov et al., 2013a], Google Books [Jones et al., 2020], as well as a 'Common Crawl' corpus [Caliskan et al., 2017]. Each category C has a set of eight words and there are two sets (M and F) of target words relating to male and female respectively (Table 7 in the Appendix). These words are based on an implicit association test also used in psychology [Caliskan et al., 2017].

The WEAT score is computed as follows: the association between a pair of words with vectors v_1 and v_2 is measured by the cosine similarity:

$$s(v_1, v_2) = \frac{v_1^T v_2}{\|v_1\|\|v_2\|}. \quad (1)$$

Let v_c denote a word from category C, v_m a male-specific word (e.g. "he" or "his") and v_f a female-specific word (e.g. "she" or "her"). First, the gender bias per word is calculated using equation 2.

$$\mathrm{b}(v_c) = \frac{1}{|M|}\sum_{v_m \in M} \mathrm{s}(v_c, v_m) - \frac{1}{|F|}\sum_{v_f \in F} \mathrm{s}(v_c, v_f). \quad (2)$$

Here, a negative value indicates the category word is female biased and a positive value indicates a male bias. This score is averaged over all words in the category C to get the bias score $\mathrm{b}(C)$,

$$\mathrm{b}(C) = \frac{1}{|C|}\sum_{v_c \in C} \mathrm{b}(v_c). \quad (3)$$

We chose to use WEAT since it is a popular way to measure bias in word embeddings and it allows us to compare our results to those of Jones et al. [2020]. This test will show whether these words contain differences in association with male and female, but how these differences relate to negative consequences in different applications is not precisely known. The results should be interpreted in this general sense, as it shows the existence of bias, but not how problematic the gender bias is.

4 Experimental Setup

All code and the models used for the experiments are made publicly available [1].

Data and preprocessing. We obtained full copies of all articles on Wikipedia in 2006, 2008 to 2010 and 2014 to 2020 from dumps.wikimedia.org and archive.org. To make a comparison between full Wikipedia backups and newly added articles, we created a second corpus by taking all articles for which the ID was not present on Wikipedia two years before. For example, to create a corpus for 2020, we removed all articles that were added before 2019. All articles were converted to tokens using the build-in functionality from the gensim library [Řehůřek and Sojka, 2010]. This tool removes all articles shorter than 50 words, next to all markup, comments and punctuation.

Training of word embeddings. The word2vec model was used to train word embeddings [Mikolov et al., 2013a]. This model uses Continuous-bag-of-words to obtain word vectors that represent the word semantics as well as possible [Mikolov et al., 2013a]. Vectors that are closer together in the vector space represent words that co-occur more often. We mostly used the default settings for word2vec as provided by gensim [Řehůřek and Sojka, 2010]. However, we did not remove the 5% most common words, because this would also remove the words 'he' and 'she'. To ensure that the training had sufficiently converged, we calculated the bias after training for one, ten and twenty iterations (epochs), besides the standard of five.

Quality of embeddings. We used the WordSim353 benchmark to assess the quality of word embeddings [Finkelstein et al., 2001]. This evaluation looks at the similarity of 353 word pairs and evaluates the correlation between the results of the embeddings and the true similarity as defined by

[1] https://gitlab.com/kschmahl/wikipedia-gender-bias-over-time

humans. We used this as a sanity check to assess whether the word embeddings reasonably embed true word semantics. These correlation scores can be found in Table 8 in the Appendix, they are all between .63 and .66. This is comparable to the correlations between .60 and .67 that were found using word2vec by Jatnika et al. [2019], which is already better than the model trained by Google they used as comparison [Mikolov et al., 2013b]. As may be expected with a smaller corpus, the scores for the data set of new articles are slightly lower (between .59 and .64), but still reasonable.

Significance of change in WEAT score. We performed a linear regression on the WEAT score versus time. We measured whether the change in WEAT score is significant by performing a t-test to compute whether the slope is significantly different from zero. To reduce the amount of false discoveries from multiple testing, we use a Benjamini-Hochberg correction with a False Discovery Rate (FDR) of 5% [Benjamini and Hochberg, 1995].

Significance against random words. A significant change in WEAT scores may not tell the whole story. It could be the case that, for some reason, all word vectors in the vocabulary become more similar to male or female words. To exclude this possibility, we also computed WEAT scores of random words, using a method proposed in the code from Jones et al. [2020]. We performed a regression on these WEAT scores for many different groups of random words to obtain a histogram of slopes. This histogram of slopes indicates the distribution of slopes for random words. We can then inspect how likely it is for a word category (such as Arts) to have the observed slope, and to see whether the slope is significantly different from slopes of random words. To this end, we used a sample of 1000 random word sets and counted how many of these slopes are at least as extreme as the observed one to determine a permutation p-value for the category word set. On these p-values we did another Benjamini-Hochberg correction with the same FDR of 5%.

Deviation of gender bias within a category. The WEAT score used to quantify the gender bias is a mean over several words in a category. It could be the case that one of the words of a word category influences the mean more than others (e.g. as an outlier). This could indicate either that a word in a word category is inappropriate, thus indicating a problem with the WEAT test. Alternatively, it can indicate where Wikipedia editors should focus their efforts on changing the language in the articles to reduce the measured gender bias. To investigate this, we also compute the deviation from the means of the different categories for 2008, 2014 and 2020. This will show if there are categories with words with large deviations. In case of large deviations, we look at the individual word scores to investigate which words have the largest influence on the bias.

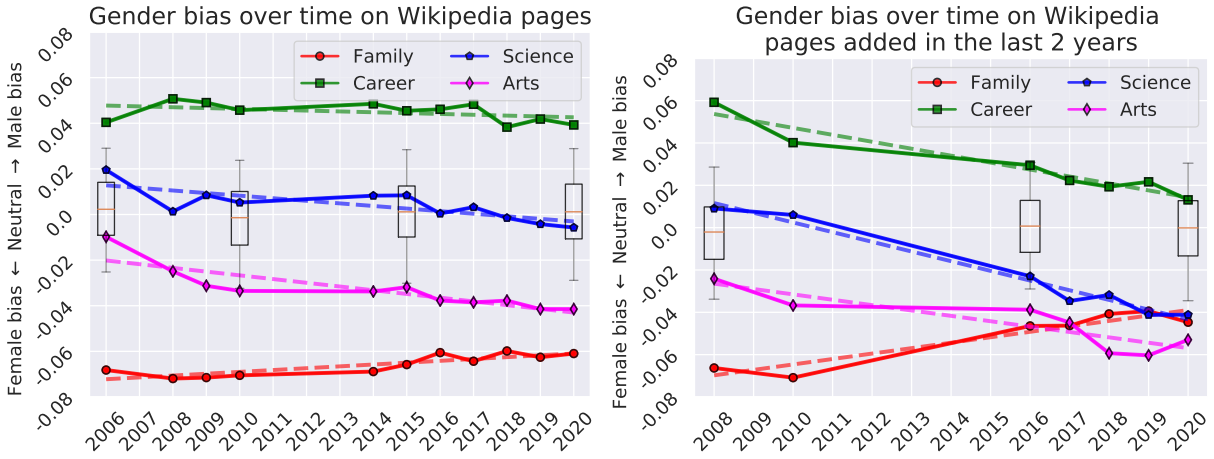
Number of articles per category. A further explanation of why gender bias has changed over time could be provided by looking at the categories of the articles on Wikipedia. We therefore counted the amount of articles which contained at least one of the words of the word categories for these three available time points.

5 Results

Gender bias scores over time. The gender biases for Wikipedia over time are shown in Figure 2a for the different word categories. The box plots indicate the distribution of WEAT scores for random words, which changes little over time and whose mean seems close to zero, indicating that random words are almost unbiased on average. Career, Arts and Family seem to have strong biases since they fall outside the box plots, while biases in Science seem milder, as its WEAT score is comparable to those of random sets of words.

Table 1 lists the p-values for whether a slope is significantly different from zero, corrected using the Benjamini-Hochberg method. Career has a strong association with male words that has not significantly changed over time. The category Science had a male bias in 2006, but this bias slowly changed over time, and is currently associated slightly more strongly with female words. This could be because the words in this category have been used in the same context as female words as opposed to male words more often since 2014. The words in the Family category have a significantly decreasing female bias, but in 2020 they are still strongly associated with female words. The Arts category is stereotypically female-associated and these words are becoming more biased towards female words, with a statistically significant slope.

Evaluation using only newly added articles. The gender bias over time for the articles added in the two years before the time point is shown in



(a) Bias scores for all articles

(b) Bias scores for new articles

Figure 2: The biases of word categories over time for Wikipedia from 2006 to 2020. Positive means the words are more associated with male, negative scores correspond to word sets more associated with female. Box plots show the distribution of biases for random word sets to put the amount of bias in perspective, the whiskers show the 5^{th} and 95^{th} percentiles. The years without box plots have a similar distribution and are hidden to improve clarity.

Figure 2b and the p-values for the slope tests in Table 1. It can be seen that the developments are similar, but steeper than when looking at all articles. The slope of the bias of all four categories is significantly different from zero in those articles. This suggests that new articles are especially less biased than older articles for the categories Career and Family. Arts and Science are more biased in recently added articles, so new articles do not seem to be better in all aspects of gender bias.

	p-value	
	All articles	New articles
Career	.207	$< .001$
Science	.007	$< .001$
Family	.001	$< .001$
Arts	.001	.010

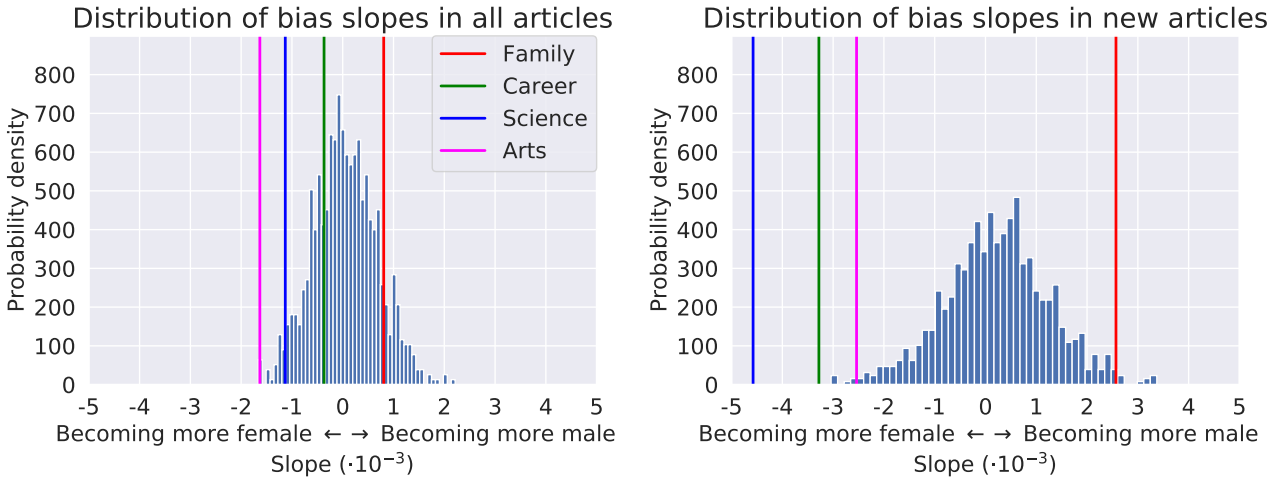
Table 1: The corrected *p*-values of t-test for the slope of the WEAT score over time. Considered significant $\leq .05$, values are corrected with a FDR of 5%.

WEAT scores of random words. The histograms of the slopes found from random word sets are given in Figure 3. The mean slope is $4.8 \cdot 10^{-5}$, with a standard deviation of $6.3 \cdot 10^{-4}$. We conclude that the whole vocabulary of Wikipedia has on average not become a lot more male or female biased over time. This is confirmed by the fact that the box plots in Figure 2a do not shift over time.

The slope for random words has a larger variance when looking at only the new articles. Random word sets have a mean slope of $2.3 \cdot 10^{-4}$ with a standard deviation of $1.0 \cdot 10^{-3}$ in the word embeddings from recent articles. This shows that the larger slopes seen in the category words for recent articles might be partly caused by larger changes seen in all word embeddings (see Figure 3b). Results of new articles are therefore less reliable, also due to a smaller corpus and less time points.

The p-values can be found in Table 2. Arts (.024) is the only category where the change is also significant compared to changes in random words for the complete Wikipedia corpus. All categories change significantly when considering only newly-added articles. The lower significance in comparison to random words means that despite the existence of slopes significantly different from 0, there may still be reason to doubt the effectiveness of the effort from Wikipedia. It also calls into question whether changes in bias in Table 1 were really significant.

Effect of number of word2vec iterations. We ran the training procedure of the word embeddings and computed the bias for each word category for one, five, ten and twenty iterations. The results are given in Table 3. Between one and five iterations the gender bias slope changes quite a bit. For example, the slope of Science changes from about $-3.1 \cdot 10^{-3}$ to $-1.1 \cdot 10^{-3}$ and the p-value of Arts varies between 0.05 and 0.01. However, most differences between five and ten iterations are smaller, including the slope values for Arts.



(a) All articles

(b) New articles

Figure 3: Probability density of the slopes of random word sets from the vocabulary.

	p-value	
	All articles	New articles
Career	.633	.024
Science	.115	$< .008$
Family	.255	.024
Arts	.024	.052

Table 2: The corrected p-values for the test of the slope of the WEAT score for categories as compared to the slopes of random words. Values $\leq .05$ are considered significant, they were corrected using a FDR of 5%. The p-value of $< .008$ is due to the finite amount of permutations (1000).

		1 epoch	5 epochs	10 epochs
C	slope ($\cdot 10^{-3}$)	-1.8	-0.37	-0.29
	p-value	.019	.554	.627
F	slope ($\cdot 10^{-3}$)	1.4	0.81	0.65
	p-value	.065	.191	.282
S	slope ($\cdot 10^{-3}$)	-3.1	-1.1	-1.3
	p-value	$< .001$	.072	.047
A	slope ($\cdot 10^{-3}$)	-1.5	-1.6	-1.3
	p-value	.054	.009	.045

Table 3: The bias scores of the categories Career (C), Family (F), Science (S) and Arts (A) from models trained with a different amount of iterations. The p-value is the computed probability comparing the category words to random words. Twenty epochs are not included since these models have much lower quality.

#Iterations	1	5	10	20
All articles	.63	.64	.64	.57
New articles	.57	.61	.62	.62

Table 4: Quality versus epochs, where quality is the average Pearson correlations of WordSim353.

The quality of the word embeddings also changed little after 5 iterations (see Table 4). This validates our choice of using the default value of 5 iterations. To further investigate if the slope and p-values were converged, we also tried 20 iterations. The resulting word embeddings had significantly lower quality scores (0.57 on average), with models trained on the most data (in more recent years) achieving scores as low as 0.52. We believe that this might be due to overtraining and therefore chose not to use these embeddings for measuring bias. We note that the number of iterations can influence the measured biases and should be varied to make certain the values have converged while models do not become overfitted.

Deviation within a word category. The means and standard deviations for the categories at three time points are given in Table 5. Family has a higher variance than the other categories. To understand why, we looked at the bias of each word in this category in 2020, see Table 6. The words 'wedding', 'marriage' and 'children' have a very strong female bias, whereas 'home', 'cousins' and 'family' are only slightly more female associated.

		F	C	S	A
2008	mean	−0.07	0.05	≈ 0.00	−0.03
	std	0.04	0.03	0.02	0.03
2014	mean	−0.07	0.05	0.01	−0.04
	std	0.04	0.02	0.03	0.02
2020	mean	−0.06	0.04	−0.01	−0.04
	std	0.04	0.02	0.02	0.02

Table 5: Means and variance within categories.

home	parents	children	family
−0.02	−0.07	−0.12	−0.02
cousins	marriage	wedding	relatives
≈ 0.00	−0.10	−0.10	−0.04

Table 6: Bias per word for the Family words in 2020.

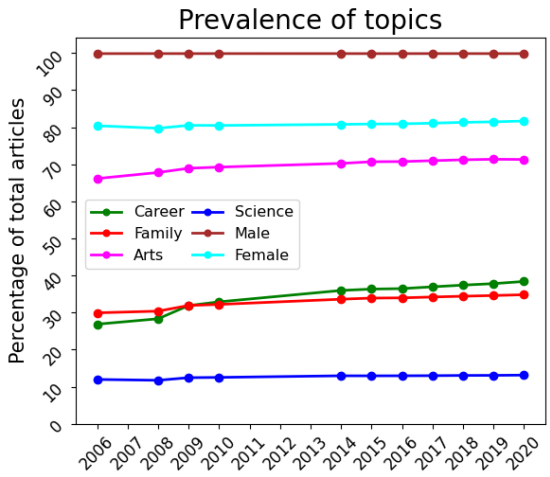


Figure 4: The percentage of articles in Wikipedia that contain at least one of the category words.

Number of articles per category. The percentage of articles which contained at least one of the words of the sets is given in Figure 4. Observe that the proportions have changed little over time, so this does not provide an explanation for the changes in bias over time. All periods thus have similar contribution to the category bias. Male words are present in more of the articles than female words.

6 Discussion and Future Work

Since societal gender bias is decreasing [Garg et al., 2018], we expected that using text written more recently would result in less gender biased word embeddings. We have shown that stereotypical gender bias in the categories Family and Science is indeed decreasing, but these changes are not significant in comparison to random word sets. Words related to Career did not seem to change since 2006. Bias in Arts has significantly increased, also in comparison to random words. Further research, maybe on a longer time period, is necessary to conclude what causes these changes and how significant the changes are.

The vast majority of biographies in Wikipedia are about men [Envel Le Hir, 2017-2020]. This discrepancy has decreased a little since 2017. This is confirmed by the fact that a lot more articles contain words from our male set than from our female set. However, we do not observe that random words are more associated with male words. This could also be seen in the fact that Science words are more female associated in 2020, despite less than 15% of the scientists with biographies being female. A possible reason for this is that articles about women contain more gender-specific words [Wagner et al., 2016], for example: 'female scientist'. The expected gender goes without saying, whereas the minority gender is explicitly specified [Pratto et al., 2007]. This causes words to become more female-associated than expected from the ratio of biographies. Wikipedia may inform its contributors about this skew in female biographies in the hope that this bias will be reduced.

To reduce gender bias in Family further, our results suggest that a focus on equal representation in the topics of marriage and children would be most beneficial. It is unclear why the Arts category is becoming more and more female biased.

When word embeddings are used in downstream tasks such as classification, our research shows it is important to consider the time of retrieval of a corpus. For example, if one wants to have a gender neutral word embedding related to Science, one may best use the corpus of 2018. Such effects may also occur in other corpora. More research is needed to further understand the quality of word embeddings as measured by performance in downstream tasks and unwanted biases in such tasks.

New articles are not gender neutral either. They have similar developments, but more strongly and also significant in comparison to random words. We could not completely determine if new articles are the cause for changes in gender bias, since we did not consider changes in existing articles. Little statistics are known relating to gender bias of Wikipedia. This makes it difficult to place our

results in a wider context. Since our work indicates biases are currently increasing further for some categories, current strategies to reduce bias may need to be changed. To further improve the editing strategies of Wikipedia, more automated measures of biases may provide necessary insights.

Compared to the historical embeddings (1800-2000) from the study of Jones et al. [2020], we find several differences but also agreements. In contrast, we find that Art related words are becoming more biased towards female. The bias of Family is decreasing in their study as well, however, they find less steep slopes. The decrease they found in the Career category was not found as clearly in our results, this may also be due to the shorter time span. It is hard to say where the differences stem from: perhaps due to different societal changes or because of a different platform?

One limitation of this research is the fact that no backups of Wikipedia were available between 2010 and 2014. Moreover, we did not look at what text was written exactly when. This information could provide more insight in the developments of gender bias. The current version of Wikipedia still contains text written in 2001, and thus biases in the full corpus of Wikipedia may not represent development of societal biases precisely. The analysis on only new articles may give a better estimate in that respect. However, due to the unreliability of using page ids, this still does not give a perfect representation.

The WEAT-score is not a perfect measure of gender bias of its underlying content. One of the problems is interpretability: where do the biases come from? To that end, Wikipedia's content should also be looked at in more detail. We tried to make this connection using word counts over all Wikipedia pages, but a more elaborate analysis is necessary to complement our analysis. Another option is to use the technique of Brunet et al. [2019] to find the most bias influencing articles. This will give further clues how to make Wikipedia more gender neutral.

Hamilton et al. [2016] discovered laws of semantic shift by looking at word embeddings over large time spans. These laws could explain some of our observed changes in gender bias. The most relevant law is the law of conformity: frequent words change embedding location more slowly. This might be taken to imply that the Arts category, whose words are most used on Wikipedia (see Figure 4), would change bias the least. However, the opposite is the case, as Arts has one of the steepest slopes. Sadly, we cannot compare our rates of change to those found by Hamilton et. al. since we cannot find the raw rates of change per year in their work. This could be used to place changes of WEAT-scores over time in context. We note, however, that the slopes of the categories are already (crudely) placed in context when they are compared against the slopes of random words. Here a further correction could be made with word frequencies to take the law of conformity into account. On the other hand, since our work focuses on a much shorter time scale, we can assume that such changes are negligible, especially for the WEAT words which are generally frequently used and therefore less likely to have major changes in meaning within 20 years.

Word embeddings were shown to be surprisingly unstable over restart with different random initialisation [Wendlandt et al., 2018]. In that work, stability was defined as the fraction of the 10 nearest neighbours of each word that are the same before and after the restart. Thus, this is a measure of local stability. The WEAT score is determined, however, over larger distances of word embeddings. Thus, local instability does not directly imply that WEAT scores would also be unstable. To mitigate this potential instability, we initialized each model with the same seed. While a more elaborate investigation of the stability of WEAT to multiple random restarts is out of the scope of this work, we think it is an important point to investigate in order to verify that our results and those of Jones et al. [2020] and Garg et al. [2018] are robust.

We considered the four default word sets as provided by the WEAT test, to allow comparison to Jones et al. [2020]. Remarkably, these word sets include two male names: Einstein and Shakespeare. Einstein is on average about 0.04 above the category mean of Science, and Shakespeare approximately 0.03 above the mean of Arts, influencing the category means positively, making them more male-biased. It is expected that the names Einstein and Shakespeare co-occur more with male words such as 'he' or 'him'. However, this may not be representative of the rest of Science or Arts words in general, and thus may overestimate male bias in these subjects. We realize that Einstein and Shakespeare were and still are very influential in the fields of science and arts respectively. However,

if our goal is that articles about more important individuals (which might be read by more people) have higher impact on the bias calculation we could weigh articles based on notability [Wagner et al., 2016] at the embedding learning stage. To further understand the (perhaps unwanted) effects of using these two words, we believe that more research in the choice of words of WEAT is necessary.

7 Conclusion

In this paper, we used word embeddings to estimate changes in gender bias in Wikipedia articles over time. We found evidence that gender bias is decreasing for Science and Family, while increasing for Arts. Biases in the male associated category Career seems constant. Further analysis of these results provides insights that can potentially lead to new practices to reduce gender bias in Wikipedia even more in the future.

Acknowledgments

We would like to thank the anonymous reviewers for their useful suggestions and comments. We would also like to thank Thijs Raymakers for his help coming up with the research plan.

References

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Eric Nalisnick, Bhaskar Mitra, Nick Craswell, and Rich Caruana. Improving document ranking with dual word embeddings. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 83–84, 04 2016. doi: 10.1145/2872518.2889361.

Yash Parikh, Abhinivesh Palusa, Shravankumar Kasthuri, Rupa Mehta, and Dipti Rana. Efficient word2vec vectors for sentiment analysis to improve commercial movie success. In *Advanced Computational and Communication Paradigms*, pages 269–279. Springer, 2018.

Beakcheol Jang, Inhwan Kim, and Jong Wook Kim. Word2vec convolutional neural networks for classification of news articles and tweets. *PloS one*, 14(8), 2019.

P Preethi Krishna and A Sharada. Word embeddings-skip gram model. In *International Conference on Intelligent Computing and Communication Technologies*, pages 133–139. Springer, 2019.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Jason J Jones, Mohammad Ruhul Amin, Jessica Kim, and Steven Skiena. Stereotypical gender associations in language have decreased over time. *Sociological Science*, 7:1–35, 2020.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey, 2018.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Quantifying and reducing stereotypes in word embeddings. *arXiv preprint arXiv:1606.06121*, 2016.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, pages 1–19, 2019.

Joel Escudé Font and Marta R Costa-Jussa. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116*, 2019.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 609–614, 2019.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Ninth international AAAI conference on web and social media*, 2015.

Wikipedia contributors. Gender bias on wikipedia — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Gender_bias_on_Wikipedia&oldid=952307164, 2020a. [Online; accessed 30-April-2020].

Envel Le Hir. Denelezh — gender gap in wikimedia projects. https://www.denelezh.org/, 2017-2020. [Online; accessed 25-May-2020].

Benjamin Collier and Julia Bear. Conflict, criticism, or confidence: An empirical examination of the gender gap in wikipedia contributions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, page 383–392, New York, NY, USA, 2012. Association for Computing Machinery. doi: 10.1145/2145204.2145265.

Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*, 5(1):5, 2016.

Wikipedia contributors. Wikipedia:wikiproject women in red — Wikipedia, the free encyclopedia, 2020b. URL https://en.wikipedia.org/w/index.php?title=Wikipedia:WikiProject_Women_in_Red&oldid=962959922. [Online; accessed 17-June-2020].

Piotr Konieczny and Maximilian Klein. Gender gap through time and space: A journey through wikipedia biographies via the wikidata human gender indicator. *New Media & Society*, 20(12):4608–4633, 2018.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, page 406–414, New York, NY, USA, 2001. Association for Computing Machinery. doi: 10.1145/371920.372094.

Derry Jatnika, Moch Arif Bijaksana, and Arie Ardiyanti Suryani. Word2vec model analysis for semantic similarities in english words. *Procedia Computer Science*, 157:160–167, 2019.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013b.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1): 289–300, 1995.

Felicia Pratto, Josephine D Korchmaros, and Peter Hegarty. When race and gender go without saying. *Social Cognition*, 25(2):221–247, 2007.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. Understanding the origins of bias in word embeddings. In *International Conference on Machine Learning*, pages 803–811, 2019.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*, 2016.

Laura Wendlandt, Jonathan K Kummerfeld, and Rada Mihalcea. Factors influencing the surprising instability of word embeddings. *arXiv preprint arXiv:1804.09692*, 2018.

A WEAT Categories

Table 7: The category and target words used to quantify biases in WEAT.

Topic	Words
Male	he, his, man, male, boy, son, brother, father, uncle, gentleman
Female	she, her, woman, female, girl, daughter, sister, mother, aunt, lady
Career (C)	executive, management, professional, corporation, salary, office, business, career
Family (F)	home, parents, children, family, cousins, marriage, wedding, relatives
Arts (A)	poetry, art, dance, literature, novel, symphony, drama, sculpture, shakespeare
Science (S)	science, technology, physics, chemistry, einstein, nasa, experiment, astronomy

B Quality per year (5 epochs)

Table 8: The Pearson correlation of the WordSim353 quality test for the word embeddings trained from Wikipedia.

	All articles	New articles
2006	.65	
2008	.64	.62
2009	.64	
2010	.64	.63
2014	.63	
2015	.64	
2016	.63	.61
2017	.63	.61
2018	.63	.62
2019	.63	.61
2020	.63	.60