Is Wikipedia succeeding in reducing gender bias? Assessing the development of gender bias in word embeddings from Wikipedia

Katja Schmahl[1] , **David Tax**[1] , **Marco Loog**[1] , **Stavros Makrodimitris**[1] , **Arman Nassiri**[1] , **Tom Viering**[1]

[1]Delft University of Technology

k.g.schmahl@student.tudelft.nl, d.m.j.tax@tudelft.nl, m.loog@tudelft.nl, s.makrodimitris@tudelft.nl, a.naserijahfari@tudelft.nl, t.j.viering@tudelft.nl

Abstract

Large text corpora used for creating word embeddings (vectors which represent word meanings) often contain a stereotypical gender bias. This unwanted bias is then also present in the word embeddings and in downstream applications in the field of natural language processing. To prevent and reduce this, more knowledge about the gender bias is necessary. This paper will contribute to this by showing how gender bias in word embeddings from Wikipedia develops over time. Quantifying the gender bias over time shows that words in Science and Arts have become more female biased. Family and Career have stereotypical biases towards respectively female and male words, which have steadily decreased since 2006. This provides new insight in what should be done to make Wikipedia more gender neutral and how important the time of writing can be when considering biases in training word embeddings from Wikipedia or from other text corpora.

1 Introduction

Word embeddings are vectors that represent the meaning of words and how words relate to each other. These representations are an important tool for natural language processing tasks. Word embeddings enable, among other things, to solve analogies, improve search results, analyze sentiment and classify documents [1–4]. These embeddings can be created with unsupervised learning from a large corpus of text [5].

Previous research uncovered that large corpora of text used for training word embeddings often contain a stereotypical gender bias [1, 6, 7]. When word embeddings are created from these corpora, they also contain this bias. Stereotypical words are more strongly associated with male words than female words or the other way around. For example, the word 'marriage' is a lot more closely associated with female words such as 'she' and 'woman' than male words such as 'he' and 'man'. This gender bias is representative of the bias present in society [6, 8]. Quantifying the biases in word embeddings can therefore also be used to evaluate the biases in society.

It is important to remove or reduce these biases from word embeddings in order to prevent unwanted consequences in usage. One example of a unwanted effect given by Bolukbasi et al. is that when improving search results, biased word embeddings can give biased results [9]. The bias could potentially cause scientific research with male names to be ranked higher since this name would have a higher association with the search words [9].

Another example of a downstream application with unwanted consequences in usage is machine translation. Google Translate uses word embeddings to improve translations [10]. However, these translations exhibit stereotypes. If you translate a sentence from a language with a gender neutral pronoun to English, a sentence about a nurse is translated with female pronouns while engineer is translated with a male pronoun [11]. These stereotypical translations can be improved by using more neutral embeddings [12].

To prevent these biases in downstream applications, more research on gender bias in word embeddings is necessary. Bolukbasi et al. have already proposed a method for debiasing [9]. However, research has shown that this debiasing covers up biases instead of removing them and there are still systematic biases present in the embeddinsgs [13]. This shows that it is very important to do more research into the presence of biases and how to measure different aspects of gender bias. This is necessary to be able to reduce the biases, either with a different debiasing method or by using less biased corpora.

Research shows that gender bias has decreased over time up to the year 2000 [7, 8]. For more recent years, how gender bias in word embeddings has progressed is still unknown. If the decreasing trend has continued in more recent years, this would mean that training algorithms on more recent data could already decrease gender bias without altering the word embeddings. To begin answering the question of whether this trend is continuing, we will research the gender bias in one of the large openly available text corpora: Wikipedia.

Research from Wagner et al. has already shown the presence of gender bias in Wikipedia [14], and the editors of Wikipedia have actively tried to reduce this since 2013 [15]. Our research could also be used to evaluate the effectiveness of these efforts and to adapt their strategies going forward.

To accomplish these two contributions, the research question 'How does gender bias in word embeddings from Wikipedia develop over time?' will be answered.

In short, it was found that words in the categories of Career and Family are respectively male and female associated, but this difference is steadily decreasing. The words that are related to the Science category used to be male biased, but have become more female than male associated over time. Arts is stereotypically female biased and this bias has increased over time since 2006.

2 Gender Bias in Wikipedia

In order to determine potential causes for the development in gender bias, it is important to review what is already known about the gender bias in Wikipedia. In 2011, a big survey was conducted to determine the demographics of Wikipedia editors. The responses of this survey showed that less than 15% of Wikipedia editors are female [16]. This resulted in research that established how this impacts the content of Wikipedia on different dimensions of gender bias. The two dimensions of gender bias that impact the word embeddings will be explained in more depth.

The first important bias to consider is the coverage bias. Coverage bias means that notable women are not covered as well as notable men are. This can be seen if a smaller percentage of the notable women has their own Wikipedia page or if these pages are less extensive. Research by Wagner et al. in 2015 looked at three data sets of notable people and no coverage bias was found. The percentage of notable women present on Wikipedia was not significantly lower than that of men [14].

However, research from 2016 showed a small glass ceiling effect was present in Wikipedia. Women on Wikipedia are on average more notable than men, which could be interpreted as proof that women have to be more notable to be covered on Wikipedia [17]. This is the bias that the efforts of Wikipedia have mostly focused on, specifically by making lists of notable missing women and creating articles for these women [18]. This could potentially cause words that are commonly used in these biographies to have become more female associated over time.

The second bias to consider is the lexical bias. Two differences in how women are represented on Wikipedia were found [17]. The first is that more words related to family and relationships are 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' [17]. 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', this would cause the word scientist to be more closely associated to female, despite there being both male and female scientists.

Little research has been done to show how Wikipedia has developed over time. Comparing the results from our research to other gender related developments on Wikipedia is therefore difficult. The only available measure is the gender bias by occupation since 2017 (see figure 1). Despite the short time period, it can be used to evaluate some of the occupations surrounding the stereotypical gender bias we are researching.

In general, the percentage of female biographies has increased steadily towards around 18%. This shows that the proportion of female biographies on Wikipedia is improving, but still very low. The biggest change can be seen for the stereotypically male occupation 'manager', for which the percentage of female biographies increased with more than 5% in this 3-year period. The occupation artist, which belongs to the stereotypically female category of Arts, has a female percentage far above average with almost 30%. However, this still means that the majority of artists on Wikipedia is male.

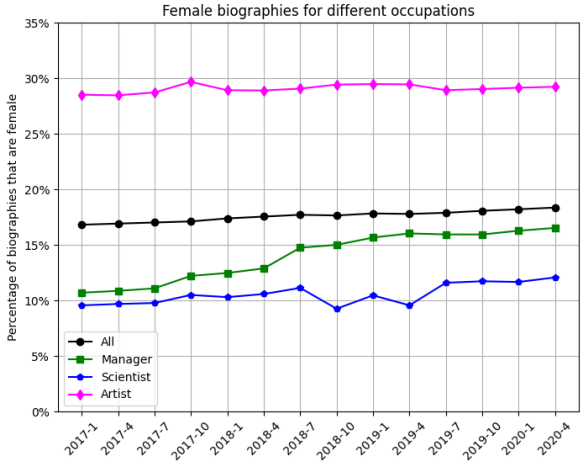

Figure 1: The percentage of biographies of women on Wikipedia for different occupations since 2014. Source: data from Denelezh [19]

3 Methodology

To answer the question of how the stereotypical gender bias in word embeddings from Wikipedia has changed over time, we first obtained the text corpora and trained word embeddings from this. Then, we used the WEAT test to quantify the biases.[1] Each of these steps will be explained in more detail in this section.

3.1 Data

For the first step in assessing the development of gender bias over time, we obtained full copies of all articles on Wikipedia in 2006, 2008 to 2010 and 2014 to 2020 from

[1]The code that was used for preprocessing, training and bias evaluation, including the used word embeddings, is available on https://gitlab.com/kschmahl/wikipedia-gender-bias-over-time.

dumps.wikimedia.org and archive.org. To make a comparison between full Wikipedia backups and newly added articles, we created a second data set by taking all articles for which the id was not present on Wikipedia two years before. The articles in these data sets were then converted to tokens. This means that all articles shorter than 50 words, all markup, comments and punctuation were removed.

3.2 Training

The gensim word2vec model was used to train word embeddings [1]. This model uses a combination of Skip-Grams (SG) and Continuous-bag-of-words (CBOW) to obtain word vectors that represent the word semantics as well as possible. This means that vectors that are closer together in the vector space represent words that are more similar. The linear structures allow for analogical reasoning [1]. For a precise explanation of word2vec, we refer to the paper of Mikolov et al. [1].

For the parameters of training, the standards of word2vec were mostly used. However, we did not remove the 5% most common words. It is important that 'he' and 'she' are not removed, since these words are relevant for assessing gender bias. We trained vectors of dimension 100 using one iteration, since more iterations did not impact the bias values (see appendix A).

3.3 Quality evaluation

Some standard quality evaluation using the common SIM353 evaluation was done. 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 humans. This was used to evaluate whether the models reasonably embed true word semantics, the embeddings have however not been optimized. We chose not to optimize the embeddings on quality metrics, due to the scope of the research and since we found no reason to believe it influences our research question.

3.4 Word Embedding Association Test

To allow more precise insight in the different areas of gender bias, we looked at four categories that are considered stereotypical towards gender: Arts, Science, Family and Career. These word sets have been proposed to compute stereotypical gender biases by Caliskan et al. [6]. Research to gender bias has shown that significant biases surrounding these category words are present in embeddings from Google News corpora [1], Google Books [7], as well as a 'Common Crawl' corpus [6]. For every category we used a set of eight category words and two sets of target words, male and female. All used word sets can be found in table 1.

For every category we formally define the bias using category set C, male set M and female set F with the word vectors from the words in table 1 as

$$bias(C, F, M) = mean_{\vec{c} \in C}(assoc(\vec{c}, M) - assoc(\vec{c}, F))$$

where:

$$assoc(\vec{w}, A) = mean_{\vec{a} \in A}(cos(\vec{w}, \vec{a}))$$

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	executive, management, professional, corporation, salary, office, business, career
Family	home, parents, children, family, cousins, marriage, wedding, relatives
Arts	poetry, art, dance, literature, novel, symphony, drama, sculpture, shakespeare
Science	science, technology, physics, chemistry, einstein, nasa, experiment, astronomy

Table 1: The category words used to quantify biases

A negative bias therefore implies that the category has a higher mean association with female words than with comparable male words and a positive bias implies that the category words are more associated with male words. We computed both the mean and the standard deviation to see how these associations are distributed over the category words. To determine the development over time of these biases, we looked at the linear regression of the bias scores from the available years.

3.5 Significance of change

The first hypothesis we are testing is whether there is a significant change in gender bias over time. This hypothesis was tested for each of the categories: Career, Science, Arts and Family. We tested this with the null hypothesis H_0 that the linear regression of the bias scores has a slope of 0. The alternative hypothesis H_1 is that the bias has changed over time. Formally defined, we computed the probability of the null hypothesis as:

$$p - value = Pr[slope(bias(C, F, M)) = 0]$$

3.6 Significance against random words

Besides establishing whether the gender bias of the categories has changed over the past 14 years, we also made a comparison with the developments of other words. This shows whether this change is stronger for the stereotypical words than it is for random words, for example due to the extra efforts of Wikipedia editors to reduce gender bias.

The second null hypothesis H_0 we are testing for every category is that the bias for these sets of words has changed no more than a set of eight random words from the vocabulary. The alternative hypothesis H_1 is that this stereotypical category words changed more than random words. We computed two probabilities to test this hypothesis.

For the first *p*-value, we computed the slope of the bias scores for 1000 random sets of words over time. We used the mean and standard deviation from this to compute the probability of our null hypothesis using a t-test. We approximate

the t-distribution to obtain a two-tailed p-value. This means that we computed the probability that the bias slope of a random word set is steeper than the slope of the category word set over time. Formally defined, let X be a random set of eight words. Then we compute a two-tailed p-value as

$$p - value = Pr[|slope_{regular}(C)| \leq |slope_{regular}(X)|$$

where:

$$slope_{regular}(S) = gradient(bias(S, F, M)$$

There is a possible weakness in this test, namely that it does not capture the absolute increase or decrease of the category biases. If all words in the vocabulary are generally becoming more biased or more neutral, this would not be captured by the distribution of the slope of the random words. The distribution of the slope would then still show some words becoming more female and some words becoming more male associated. The development of the stereotypical word categories could falsely be considered more or less significant.

Therefore we added a second test, which considers the slope of the absolute bias. This looks at whether the words are becoming more biased or more neutral instead of more female or more male associated. This test is added for all categories that can be accurately represented using absolute bias values. This results in the following probability for the null hypothesis:

$$p - value = Pr[|slope_{absolute}(C) \leq |slope_{absolute}(X)]$$

where:

$$slope_{absolute}(S) = gradient(|bias(S, F, M)|)$$

4 Contribution

This study contributes in two different aspects.
First of all, it expands on the work of Jones et al. and Garg et al. [7, 8] by looking at more recent years. This gives more knowledge about bias trends in recent years and also on a different, online, platform. Furthermore, a potential weakness in significance testing for biases in word embeddings is shown in this study. It adds the consideration to look at both absolute and normal biases. This means not just considering whether words are becoming more male or female associated, but taking the possibility of words becoming more biased or more neutral into account. An extra test of significance that takes these absolute changes into account is introduced in comparison to the work of Jones et al. [7].

Besides that, it gives more insight in the development of the systematic bias problem in Wikipedia. So far, most research into this is static and no research has been done to demonstrate how this bias has potentially changed. This research fills this gap and gives insight in whether the efforts of Wikipedia editors are successful and possibilities for how to adapt their strategy.

5 Results

The results given in this section were created using the methods described in section 3. The gender bias is quantified for different years in order to assess the development. This was done for both the full Wikipedia copies and for articles added in the previous 2 year period. Furthermore, the number of articles in which the category words are present was reviewed.

5.1 Gender bias of complete backups

The biases calculated for the available Wikipedia copies in the four categories are visualized over time in figure 2. The words in the category Career have a stronger association with male than with female words, but the difference is decreasing. The category Science had a male bias in 2006, but is currently associated more strongly with female words. This means that 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 decreasing female bias. The Arts category is stereotypically female associated and these words are becoming more biased towards the set of female words.

Significance of change

The first test we did is to see how likely it is that the category did not change in association with male and female. This resulted in p-values that are all lower than 0.001 (see table 2), which makes our null hypothesis of a slope of 0 for the bias values very unlikely. We therefore reject the null hypothesis for all categories in favour of the alternative hypothesis that the category words have made a significant change in gender bias. In all categories, the difference in male and female association has changed over the past 14 years.

Table 2: The p-values of the slope of the bias for all articles in Wikipedia, defined as the probability that the slopes of the bias scores is 0.

	All articles
Career	$4.37 \cdot 10^{-4}$
Science	$1.91 \cdot 10^{-6}$
Family	$1.43 \cdot 10^{-4}$
Arts	$2.42 \cdot 10^{-5}$

Significance of random words

To determine whether the development of the biases is significant in comparison to random words, we calculated the slope for 1000 sets of eight random words. We looked at the distribution of slopes of the regular bias values. The probability density functions we fitted from this can be found in figure 3. Random word sets have a mean slope of $1.29 \cdot 10^{-4}$, with a standard deviation of $7.35 \cdot 10^{-4}$. Since bias values are generally between -0.03 and 0.03, this is a small change. The vocabulary of Wikipedia has on average not become a lot more male or female biased. We also compared our categories with random word sets for absolute biases, this distribution is also visualized in figure 3. The mean increase of the absolute bias

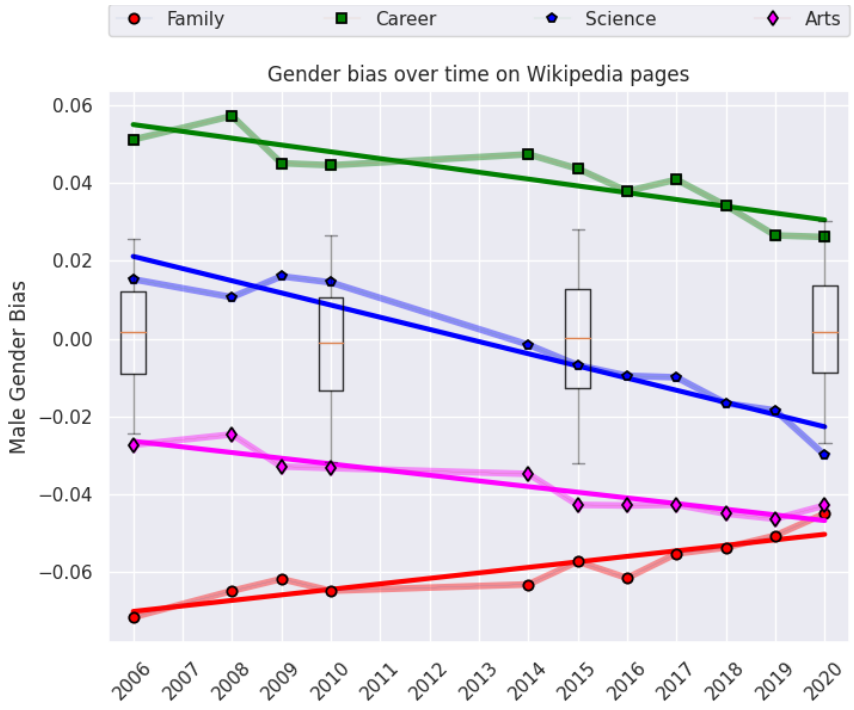

Figure 2: The bias over time in complete backups of Wikipedia for four categories since 2006. Positive means the words are more associated with male. 0 is equal male and female bias and negative gender bias means more associated with female. The boxplots show the distribution of biases for random word sets in that year to put the bias in perspective, the whiskers show the 5^{th} and 95^{th} percentiles. The years without boxplots have a similar distribution.

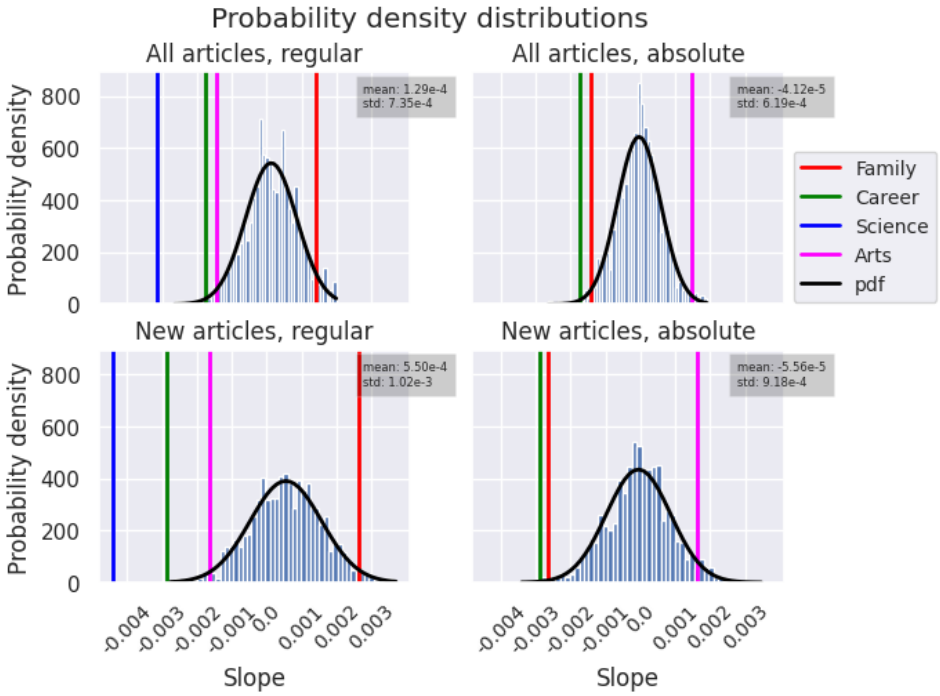

Figure 3: The probability density histograms and functions for the distributions of the random words. For the absolute bias values, positive means more biased and negative means more neutral. For the regular bias values, positive means more male and negative means more female.

is $-4.12 \cdot 10^{-5}$ with a standard deviation of $6.19 \cdot 10^{-4}$. This shows that random words have on average not become a lot more biased or more neutral.

These two tests result in the *p*-values in table 3. This shows that the null hypothesis for the slope of bias might not be confidently rejected for all categories. The probability that these words change gender association less than a random set of words is too large for the categories Family (0.0798) and Arts (0.0706). For the category of Career, it depends on the chosen significance level. Using a confidence level of 0.05, the highest *p*-value of 0.0273 is low enough to reject the null hypothesis. The *p*-value of Science is low enough to reject the null hypothesis for this category. Therefore, it can be stated that significant changes in the association of stereotypical words of the categories Career and Science are present in comparison to random words, but not for the categories Family and Arts.

The Science category has not been assessed using the absolute bias values, since this actually reverted from strong male bias to a strong female bias. Since it was first positive and then negative, using the absolute values does not result in a linear change. The linear regression of the absolute bias therefore does not represent the development. Looking at the development in comparison to the distributions of random word biases, shows that it goes from the tail of one side of biased values to the other tail. This, combined with the high slope, shows that the change in this category is significant in comparison to random words (see figure 2).

Table 3: The two-tailed *p*-values of the slope of the bias for either regular bias values or absolute bias values. This means the probability that the bias of a random word set is more extreme in either direction than the category word set.

	All articles	
	regular	absolute
Career	0.0273	0.00382
Science	$4.46 \cdot 10^{-5}$	n.a.
Family	0.0798	0.0187
Arts	0.0706	0.0156

Deviation within category

For three of the available years, we also considered how much this bias deviated between the category words. These results can be found in table 4 (appendix B). The bias of the category Family has a higher variance than the other categories. To understand why this is the case, we looked at the bias for the words in this category in 2020 (see table 5, appendix B). This gives the insight that the words 'wedding', 'marriage' and 'children' have a very strong female bias, whereas 'home', 'cousins' and 'relatives' are only a little bit more strongly associated with female than male words. To reduce the Family bias further, it is therefore important to focus on equal representation of men and women, when it comes to marriage and children. Career and Arts have deviations that decrease over time. When it comes to gender bias, the words in these categories are becoming more similar.

5.2 Comparison newly added articles

To have a better understanding of what causes the development of the gender bias, we also looked at the gender bias of newly added articles. We filtered this by looking only at pages with a page id that was not present on Wikipedia two years before. It can be seen that the developments are similar, but stronger than when looking at all articles (see figure 4). However, the data points have a stronger variance. This variance might be caused by the smaller corpus used with new articles. The slope for random words also has a stronger variance when looking at only the new articles (see figure 3). Therefore, the slope of the change for the categories can not be reliably tested on significance using only new articles (see table 3). Therefore, it is not possible to see whether the changes in gender bias are caused by new articles or rewriting existing articles.

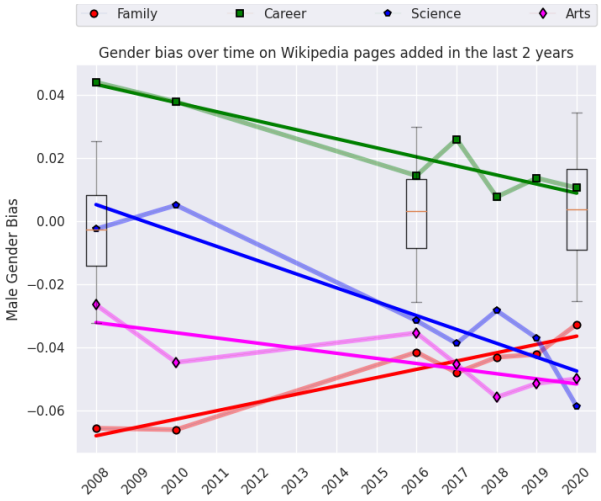

Figure 4: The bias over time in new articles in Wikipedia for four categories. The boxplots show the distribution of biases for sets of random words.

5.3 Category word counts

For three of the complete Wikipedia dumps, we evaluated the amount of articles which contained at least one of the words of the 6 categories (see figure 5). This shows the proportion of articles about the different topics over time. This proportion has changed little, so the different categories seem to not have become a lot more or less important for the content of Wikipedia. This is relevant to establish, since if the category word would be used more in a certain period, this period would also have the biggest contribution to the category bias. If words are used less, there will also be less development in the word embeddings.

This basic measurement already shows that there is a difference between male and female on Wikipedia. Almost every article on Wikipedia contains at least one of the words

from the set of male words, while only around 80% of the articles contains one of the female words. The article count also show that from the four categories, the Science category words are present in the least articles and the Arts words in the most.

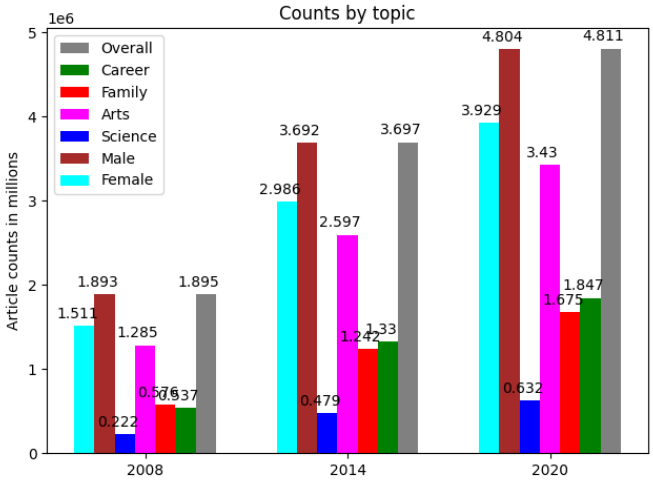

Figure 5: The amount of articles present in Wikipedia that contain at least one of the category words.

6 Responsible Research

All data and code is publicly available, so the research is easy to reproduce. This could also be done using data sets from different sources. Besides that, the word embeddings used are also made available. This makes it easier to research more categories and it makes it possible to further investigate the embeddings.

In this study, we were only able to include binary gender, i.e. male and female. We acknowledge that there are more gender identities, but due to technical limitations and the nature of Wikipedia articles, we made the choice to exclude those. In future research, it would be interesting to also incorporate more gender identities.

7 Discussion

It has been shown that the traditional stereotypical gender bias in the categories Family and Career is decreasing from being respectively female and male associated. The stereotypically male category of Science has reversed. It is currently biased towards female, and increasing. Arts words were already female biased and this bias is also increasing.

A comparison can be made between the results of evaluating gender bias in historical embeddings from the study of Jones et al. The biases of Career and Family are decreasing as they were in the results they found from literature from 1800 to 2000 [7]. The slopes are steeper in our shorter and more recent period, than they were in their research. It is not possible to determine from these results whether that is caused by stronger societal changes, a different platform or some other reason. When it comes to Science and Arts, the fact that they are increasingly biased in our Wikipedia results is different than it was in the research of Jones et al. [7]. It might be interesting to look into why this is different in future research.

The general societal gender bias is decreasing [8], therefore we expected that using text corpora written more recently would result in less gender biased word embeddings. When looking at the categories Career and Family, filtering text corpora on time would decrease the bias in word embeddings created. However, this is not the case for all categories and this shows that it would not only have beneficial effects when training from Wikipedia. The categories did however all show significant changes. Therefore we believe that time of writing should still be a considered factor when training word embeddings, especially when it comes to reducing or preventing biases.

The results as we found them can also be compared with what has already been established about Wikipedia's biases. The first thing that is known about Wikipedia is that the vast majority of biographies is about men [19]. This discrepancy has decreased a little since 2017. There are also a lot more articles containing one of our male words than articles containing the female words. However, this difference did not influence all words in the vocabulary. Random words are not a lot more male than female associated.

Another thing that was established in Wikipedia is that female biographies are more likely to contain gender specific words. This might be seen in the fact that Science words are more female, despite less than 15% of the scientists with biographies being female. This shows that it might be relevant for more efforts to decrease this difference in representation between men and women.

7.1 Limitations

First of all, this research is limited by the fact that no backups of Wikipedia were available for some of the years in the period we are researching. This caused a gap between 2010 and 2014, during which we are unsure how the bias developed. Besides that, it was not possible within this research to look at what text was written exactly when. Adding this could provide more insight in the developments. The current version of Wikipedia still contains text written in 2001, so it might not represent development of societal biases as precisely. The comparison that was made with new articles also does not give this precision, since this does not use text added to other articles. Besides that, page ids can also have changed, which will cause it to be falsely considered new.

Another limitation is in the significance testing. This is complicated by the difference between absolute and regular bias. In order to reject the null hypothesis that the stereotypical categories change less than or equal to the random words, it is necessary to ensure that there is no clear pattern in how random words change. If all words are becoming more male/more female or all words are becoming more biased/more neutral, this should be considered when comparing with the category words. We found no common standard

of which method to use for determining the significance of bias developments and argue that this is something that requires more research. It can have beneficial consequences if the development of biases becomes something more common to evaluate and standards for how to do this would allow better comparison.

Thirdly, we only considered four categories of gender bias. These category words have been used more often for quantifying biases, however it includes two male names (Einstein and Shakespeare). They are both more male biased than their category average, which makes the category more male associated. However, the bias for this word does not have to be problematic, since it is a male person. Furthermore, stereotypical gender associations are likely to be present in a lot more categories. The categories of bias might be changing, so research to which categories present biases and how to measure these should be continuously done.

Besides that, it has been shown that more common words in a language change less in meaning [20]. This means that comparing to random words, might not be a completely fair comparison. However, the words used to represent the categories are not very uncommon, so this does not explain the significance of the development. Besides that, the counts of the articles containing words in the category has shown that these categories have not become a lot more or less common in Wikipedia, so we believe this statistical law did not impact our results a lot.

Lastly, little is known about the development in Wikipedia. Data about gender of biography articles has only been tracked since 2014. We have found nothing else that is used to assess progress in reducing the systematic gender bias. This makes it difficult to place our results into a wider context. Based on our limited results, we believe the development of biases in Wikipedia is cause for concern. Therefore, we argue that is important that Wikipedia starts using more measurements and keeps track of the changes in gender bias, also in how people are represented.

8 Conclusions and Future Work

In this paper, we looked at word embeddings from Wikipedia to answer the main research question: 'How does gender bias in word embeddings from Wikipedia develop over time?'. We answer this question using four categories of stereotypical gender bias. A male biased category is a category where the words are on average more strongly associated with a set of male words than with female words. The Career category has the expected male bias and the Family category has the expected female bias. These biases have both steadily decreased since 2006. The stereotypically male category Science has changed from male biased towards female biased and this female bias has also increased up to 2020. The last category is Arts, which is stereotypically female and has also become more female biased over the past 14 years.

The Science category shows an unexpected development. The stereotypical male bias for Science words present in 2006 has reversed and is now female and increasing. This shows that despite the fact that only 12% of scientists with Wikipedia articles are female, science words are more associated with female. A possible reason for this might be that articles about women contain more gender-specific words [17]. This is known as the principle of a stereotypical 'default gender'. The expected gender goes without saying, whereas the minority gender is explicitly specified [21]. When gender-specific words are present more in female biographies, this causes words to become more female than expected from the ratio of biographies. Wikipedia aims to provide an objective point of view, so this is something that they might have to adapt their strategy for reducing gender bias to.

All categories have made a change in the difference of association with male and female words over the last 14 years, so time of writing makes a difference for gender bias. This shows that looking into the bias over time can provide more insight in the stereotypical gender bias and that it should be more common to do this for large text corpora. For Wikipedia, not all categories has a decrease in bias, so using recent text might not be an improvement for the gender bias. Therefore, the effect of using more recent text corpora requires more research, also for other corpora. This further research should also incorporate additionally the quality consequences filtering on time of writing would have.

Looking at new articles shows that also articles added recently are not gender neutral, however this did not allow to draw any conclusion to whether the developments in gender bias are caused by new articles or by rewriting older articles. Family and Career are making good progress, but Science went from male to strongly female biased and the Arts category is only becoming more biased. Since some categories have actually become more biased over time, the current way of reducing biases is not going towards a more gender neutral Wikipedia in every aspect. This shows how important it is to measure these developments using more measures than only the percentage of female biographies.

References

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

[2] Eric Nalisnick, Bhaskar Mitra, Nick Craswell, and Rich Caruana. Improving document ranking with dual word embeddings. pages 83–84, 04 2016. doi: 10.1145/2872518.2889361.

[3] 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.

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[11] 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.

[12] 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.

[13] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *CoRR*, abs/1903.03862, 2019. URL http://arxiv.org/abs/1903.03862.

[14] 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.

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

[16] 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. ISBN 9781450310864. doi: 10.1145/2145204.2145265. URL https://doi.org/10.1145/2145204.2145265.

[17] 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.

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

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

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

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

A Bias over iterations

To determine how the amount of iterations affects gender bias, we computed the biases after every iteration for 10 iterations (see figure 6). The gender bias does not change a lot, so we did the research using one iteration.

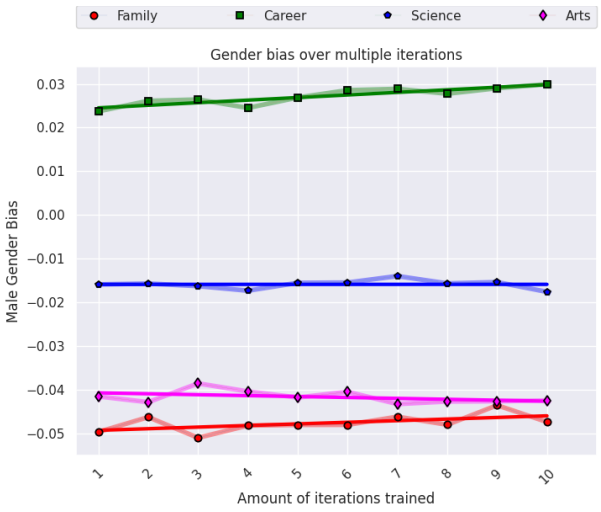

Figure 6: The development of the gender bias as over multiple iterations, trained from the full Wikipedia corpus from 2020.

B Category bias deviations

To know more about the coherence of the categories, we looked into the deviation of bias within the category words. This shows that Family words deviate the most (see table 4, this can be explained by looking at the bias per word in table 5. It shows that especially 'marriage', 'wedding' and 'children' are very female biased. 'home', 'cousins' and 'relatives' are only a little biased.

		Family	**Career**	**Science**	**Arts**
2008	*mean*	-0.0548	0.0508	0.0127	-0.0441
	std	0.0416	0.0244	0.0202	0.0281
2014	*mean*	-0.0495	0.0517	0.0032	-0.0458
	std	0.0410	0.0180	0.0150	0.0203
2020	*mean*	-0.0372	0.0289	-0.0193	-0.0582
	std	0.0443	0.0152	0.0202	0.0125

Table 4: The mean and standard deviation of the bias for the different categories.

word	**bias**
home	-0.0124
parents	-0.0590
children	-0.1126
family	-0.0149
cousins	0.0131
marriage	-0.0856
wedding	-0.0759
relatives	-0.0125

Table 5: The mean difference in association with male and female words for the Family words in 2020.