Analysis of Gender Bias in Wikipedia Biographies

Jiemin Tang
Department of Computer Science
University of Southern California
Los Angeles, CA
jieminta@usc.edu

Zixin Zhang
Department of Computer Science
University of Southern California
Los Angeles, CA
zzhang09@usc.edu

Abstract

The purpose of this project is to evaluate the text contents of Wikipedia Biographies and quantify the extent of gender bias if such bias does exist. Two main approaches are used to evaluate gender bias among different categories, with the first one being Word2Vec model with Word Embedding Association Test evaluation, and the second one being BERT classification model to predict the gender described given sentence context. The results of the above models substantiate the existence of gender bias in Wikipedia biographies, indicating that females are more likely associated with art and literature related subjects while males more likely associated with work, math, and science related subjects. Details of the findings will be discussed in length in the following sections.

1 Introduction

With the extensive application of Natural Language Processing algorithms in the field of Artificial Intelligence, potential bias and unfairness targeting a specific gender are becoming more and more noticeable. Since the algorithms are designed, implemented, and trained on real data produced by humans, such bias and unfairness could often result from the data itself. Taking text data as an example, there are numerous biographies of celebrities on Wikipedia from various backgrounds and areas of expertise. Apart from the achievements and related work, the biographies also record many aspects of the private life of each celebrity. People could be depicted very differently depending on their gender, and the emphasis on specific parts of their life may also vary. Thus, the goal of this project is to perform a comprehensive analysis of Biographies published on Wikipedia and determine whether gender bias exists. [1]

By comparing the adjectives used to describe males and females, if a notable difference is found in the categories of words associated with each gender, then the existence of gender bias can be concluded. This will be evaluated by the Word2Vec model and Word Embedding Association Test (WEAT) in the first part of the analysis, trained on all of the gender-related sentences in the data set. Another approach to analyze gender bias in Wikipedia biographies would be using the BERT classification model to predict the masked gender pronouns given the sentence context. Ideally, if there does not exist gender bias in the biographies, then the BERT model should predict males and females with equal probability regardless of the categorical words present in the text. Combining BERT prediction results with Word2Vec cosine similarities and WEAT scores from the first approach, the extent of gender bias can thus be quantified with numerical values. [2]

[1] Link for data set: https://drive.google.com/drive/folders/1hVp6b9CBizCfFvTDbxgtdQrtIPTqy4qU?usp=sharing

[2] Link for project code: https://github.com/zitanshu/Wiki_Bios_Gender_Bias

2 Literature Review

There have been a lot of studies conducted on gender bias in Wikipedia biographies. Those aspects include but are not limited to the length and coverage of female biographies, the style of writing regarding women's pages, and the likelihood of women's pages being deleted on Wikipedia. Reagle and Rhue found that women are more likely neglected when citing the work, and the coverage of their contents is less comprehensive [4]. Multiple studies have mentioned that female pages are more likely deleted because their pages are less connected than male pages, meaning those pages are less likely to be read and edited [9 & 10]. In addition, Tripodi and Graells-Garrido et al. discussed that female pages are more likely to contain gendered language, resulting from the assumption that the page is about a notable male unless otherwise indicated [5 & 8].

All the related work above argued that there exists substantial bias towards females' biographies, and such conclusions were mostly supported by machine learning models or NLP techniques. For example, Reagle and Rhue performed logistic regression on the relationship between gender and biography coverage, and the coefficient for males in Wikipedia is significant, proving that gender does contribute to the degree of completeness in biographies [4]. Field et al. performed pivot-slope TF-IDF matching to examine the potential bias among different gender and racial groups, and the statistics suggested that women do face discrimination, and failure of considering demographics would neglect the fact that some women face more discrimination than the others [6].

While the authors unanimously agreed that there is a certain degree of bias or unfairness towards females in Wikipedia biographies, few had attempted to quantify the extent of such bias among the two genders. Although Wagner et al.'s study indicated a stronger association between females and family/romantic relationships [7], they did not include the results for work or more diverse subjects. Therefore, the goal of this project is to include a more comprehensive range of categories and try to quantify gender bias in Wikipedia biographies with numerical values.

3 Data Set

The data set used for the project is found on:

`https://github.com/DavidGrangier/wikipedia-biography-dataset`

There are a total of 728,321 biographies in the above data set stored in 16 zipped folders. In each folder, the biographies had been separated into individual sentences. The team is going to filter out sentences that contain gender pronouns, e.g., he/she, and words from predefined categories for further processing. The sentences would then be used to train Word2Vec and BERT models to compare whether the adjectives are significantly different for males and females.

3.1 Word Stimuli Table

The word stimuli table 1 below is carefully generated from the paper below after adjusting the precision, comprehensiveness, and frequency of word occurrence in order to adequately represent each category of word. This table is further used to evaluate the gender stereotype phenomena on more than 65 million words extracted from both adult and children's conversations, books, movies, and TV throughout the time.

`https://www.benedekkurdi.com/files/Charlesworth_PS_2021.pdf`

Similar word categories and word stimuli will be used in this project to identify potential gender bias in wiki biography. Considering the size and wide variety of sentences in the biographies, the team has decided to include more words in each category. With a larger word stimuli table, the models would be able to produce more accurately by using a larger proportion of the original data set.

3.2 Data Preparation

As the purpose of the project is to evaluate and quantify the extent of gender bias in biographies, only the sentences that include gender-specific pronouns and words from predefined categories are reserved for modeling. The team has decided to evaluate gender bias in two aspects, one being

Table 1: Word Stimuli Used to Represent Each Category and Attribute

	Part
Category	Word Stimuli
Female	she, her, hers, mommy, mom, girl, mother, lady, sister, mama, momma, sis, grandma herself, woman, women, female
Male	he, his, him, daddy, dad, boy, father, guy, brother, dada, papa, bro, grandpa, himself, man men, male
Good	happiness, happy, fun, fantastic, lovable, magical, delight, joy, relaxing, honest, excited laughter, cheerful, great, moral, kindly, upright, worthy, outstanding, admirable, gracious valuable, excellent, exceptional, superb, wonderful
Bad	torture, murder, abuse, wreck, die, disaster, mourning, killer, nightmare, stress, kill, death deficient, horrible, wrong, unsatisfactory, dreadful, flawed, useless, worthless, immoral nefarious, sinful, unethical, harmful, hostile
Home	baby, house, home, homes, wedding, kid, family, marry, domestic, household, households chore, chores, families, kitchen, infant, babysit, childcare, parenting, garden, laundry clean, cook
Work	work, works, worked, labor, worker, workers, economy, office, job, jobs, business businesses, trade, trades, company, companies, industry, industries, pay, pays, paid working, salary, salaries, wage, wages, activity, act, money
Art	art, artist, artwork, dance, dancing, dancer, sing, singing, singer, paint, painting, painter song, draw, drawing, craft, handcraft, handicraft, music, sculpture, design, photography
Science	science, scientist, chemistry, physics, engineer, space, spaceship, astronaut, chemical microscope, theory, physics, medicine, engineering, biology, technology, astrophysics biochemistry, geology, laboratory, radiology
Literature	book, books, novel, read, write, story, word, writing, reading, tale, history, poetry, poem poet, literacy, written, prose, fiction, essay, author
Math	puzzle, number, count, math, counting, calculator, subtraction, addition, arithmetic calculation, calculus, ciphering, computation, mathematics, geometry, algebra

the differences in associating each gender to general categories Good and Bad, the other being the differences in associating gender to detailed categories such as art, science, etc., as indicated in Table 1. For the Word2Vec model, all of the filtered sentences in the Wikipedia Biography data set will be used for training. For the BERT classification model, the filtered sentences are split into the train, validation, and test sets with an 80:10:10 ratio. More specifically, prior to further processing and cleaning, the train, validation, and test set each has 2395103, 299273, and 298900 sentences respectively.

After filtering the sentences for each aspect, the number of sentences in train, validation, and test set for Good/Bad categories are 37009, 4606, and 4657 respectively, and 270052, 33506, and 33585 sentences for subject categories respectively. Although the number of data points is largely reduced compared to the original data set, the size is still significant enough for a generalizable evaluation of gender differences.

3.3 Data Cleaning and Data Preprocessing

Prior to training the models, a few more steps of NLP techniques are performed on the data set. Firstly, the stop words and punctuation are removed from sentences, as they do not have many contributions to the meaning of the sentences, nor would they provide valuable information to the models. In addition, removing stop words and punctuation would reduce the size of the sentences, leaving only the significant words, thus reducing the training time and potentially improving the accuracy of the model. Lemmatization is also performed on the data set to standardize the words into their base forms. This guarantees that the meaningful words will be correctly matched to each gender without having a lot of variants to disrupt or confuse the model.

3.4 Data Imbalance

After the step of filtering out sentences with gender-specific pronouns, it was noticed that there exists the problem of class imbalance. There are a total of 1,103,899 male-related sentences while there are only 223,523 female-related sentences. This proportion is about the same after filtering for general categories as well as specific categories mentioned above. Since this reflects the actual frequency of females and males being mentioned in Wikipedia biographies, and the proportion of each category should be the same for each gender in the absence of bias regardless of sample size, the team decided not to adopt any technique to modify the data set. Moreover, down-sampling the number of male-related sentences would negatively affect the prediction of the BERT classification model, the sample size is thus kept unchanged.

4 Methods

4.1 Word Embedding Model

After combining the train, validation, and test set together, all gender-related pronouns are replaced by {female, male} labels in the sentences, so are the other categorical words replaced by their corresponding categories as stated in table 1. All sentences after combination and replacement are further divided into two main categories: Good vs. Bad categories, and other categories (Home, Work, Art, Science, Literature, Math) to train two different word embedding models using Word2Vec and make a better comparison within each main category. Upon constructing the data set and Word2Vec models, the cosine similarity score between each gender and each category is calculated to investigate the potential existence of gender bias. Ideally, in the gender bias-free situation, the cosine similarity in the same category should be equal or close between the female and male groups. Detailed results and findings will be discussed in the following section.

4.2 Word Embedding Association Test (WEAT)

WEAT serves as the evaluation metrics for the Word2Vec model regarding the two genders. It measures the degree to which the model associates sets of target words (female vs male) with different sets of attribute words (e.g. Good vs Bad) used for training the Word2Vec model. The score of the test ranges from -2.0 to 2.0, with a large numerical value closer to 2 indicating a strong negative or positive relation between the first target set and the first attribute set. If there is no gender bias in the biographies, WEAT scores should be close to 0 for any pair of attribute sets.

4.3 Bert Classification

For the BERT classification model, the train, validation, and test set remained in their proportion and each sentence after the data preparation process gets assigned a gender label and a unique category label defined in Table 1. All gender-related words are masked and the two BERT classification models, one for Good vs. Bad categories, and one for other categories (Home, Work, Art, Science, Literature, Math), are trained and validated to learn how to classify each gender-word-free sentence into the correct gender label. If the text is unbiased, for each sentence with a pre-defined category, the BERT model should assign it to either gender with the same probability. One of the main goals of the project is to compare the gender classification results in each of the categories and check for potential bias.

5 Results and Evaluation

The cosine similarity scores between each gender and category are displayed in Table 2. In the main division of Good vs. Bad, the male group has a higher score for both good and bad categories, which may be caused by class imbalance since the amount of text about males is more extensive. As for the other categories, the female group only has a higher association with the Art category, and the computed scores of the two genders are almost equal in the Literature category. In all other categories (Home, Work, Science, Math), the results show a higher association with the male group, which corresponds to the typical gender stereotypes and suggests the existence of gender bias in the wiki-biography context.

Table 2: Similarities Between Gender and Categories

Category / Gender	Male	Female
Good	0.18	0.06
Bad	0.13	0.06
Home	0.20	0.16
Work	0.33	0.15
Art	0.28	0.35
Science	0.25	0.11
literature	0.28	0.27
Math	0.25	0.11

Table 3: Word Embedding Association Test Scores

Female and Male w.r.t.	Scores
Good vs Bad	0.206
Home vs Work	1.191
Art vs Science	0.912
Art vs Literature	0.032
Art vs Math	1.346
Science vs Math	0.433
Science vs Literature	-0.881
Literature vs Math	1.314

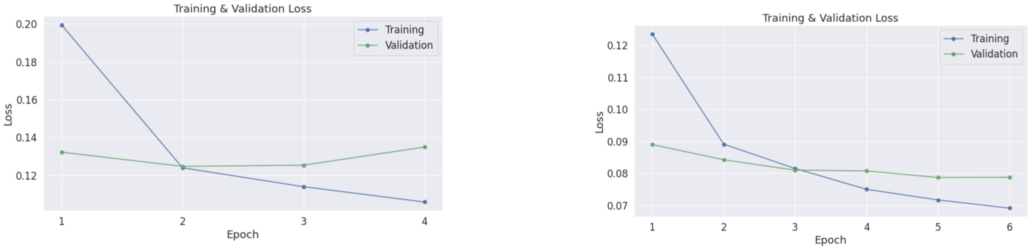
As indicated in the table 3, for general categories Good/Bad, WEAT does not associate the attributes to either gender. However, in the specific categories, it is clear that females are highly associated with home, art, and literature (indicated by high positive scores), while males are highly associated with work, science, and math (indicated by high negative scores). It is worth noting that WEAT scores only measure the association between females and the first attribute set relative to the second attribute set compared to males, not against all other attributes as a whole. For example, the score of 1.191 for Home vs Work indicates that females are more likely associated with home than work, without taking other categories into account. Consistent associations were also found in the Word2Vec cosine similarities.

Table 4: BERT Evaluation

	Accuracy	Precision	Recall
Good vs Bad	0.952	0.944	0.998
Other Categories	0.968	0.975	0.994

Aside from Word2Vec cosine similarity and WEAT scores, the team further applied the BERT model to classify the sentences into gender labels. As the train and validation loss value shown in Figure 1a and Figure 1b, the BERT model at epoch 2 and epoch 5 are chosen to be the final classification model for each main division since the validation losses are minimized, and the accuracy, precision and recall on each selected model are all satisfactory as demonstrated in Table 4.

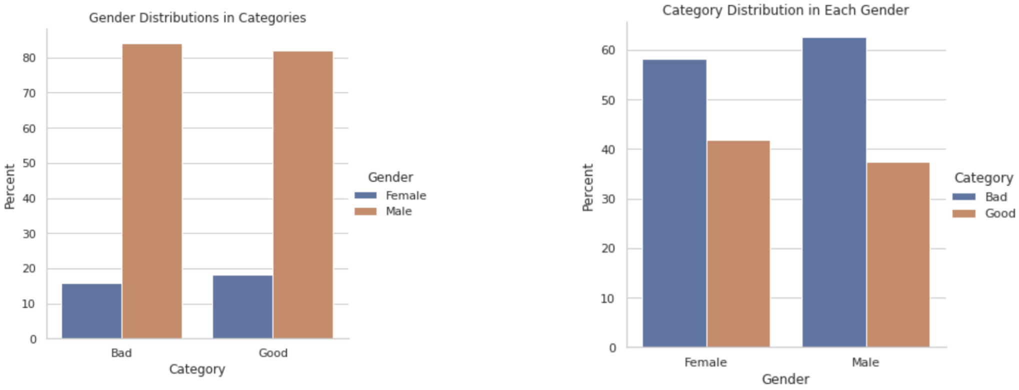
For each main division, the corresponding classification results are generated based on the 10% test set. In the Good vs. Bad division, according to Figure 2a and Figure 2b, the predicted label for males is significantly higher than females in both good and bad categories which is coherent with the previous results. As for the other categories (Home, Work, Science, Math, Art, Literature), the gender distribution of the male group again dominates in every sub-category as shown in Figure 3a. Although this indicates the BERT model is affected by the issue of class imbalance, removing it causes the model to lose a significant amount of information and reduces the accuracy by over 40%. Thus, considering the adverse impacts on models and this imbalance reflects the true gender ratio in wiki-bios, it is not eliminated.



(a) Good vs Bad categories

(b) Other Categories

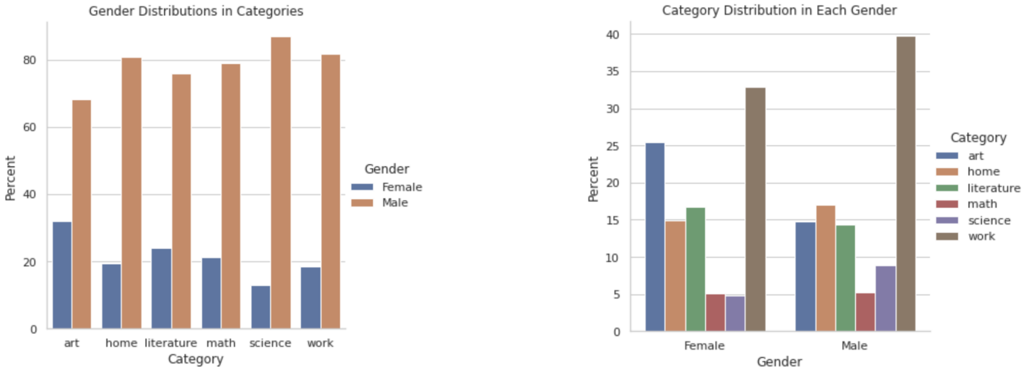
Figure 1: BERT Train & Validation Loss



(a) Gender Distribution in Each Category

(b) Category Distribution in Each Gender

Figure 2: BERT Results on Good and Bad Categories



(a) Gender Distribution in Each Category

(b) Category Distribution in Each Gender

Figure 3: BERT Results on Other Categories

However, the audiences can still get the demonstration of removing the impact of class imbalance without losing a considerable amount of information by computing the category proportions within each gender as in Figure 3b. By comparing the relative proportion of each sub-category across different gender, one can get a consistent outcome with previous methods. The classified female group contains a higher proportion of art and literature categories than the classified male group, whereas the proportion of all other categories (home, math, science, work) are all subordinate.

6 Conclusion and Discussion

In general, the project enhances and takes a step further compares to the previous works in several ways. First, it takes advantage of the large data size by incorporating all gender-related sentences from wiki bios which provides a solid and comprehensive foundation for the models. Then, the size of the word stimuli table is enlarged to conduct a maximum coverage of sentences that includes gender and categorical words. From the technical perspective, multiple NLP techniques are applied to better quantify and prove the existence of gender bias. The cosine similarities from Word2Vec indicate the association between each gender and each sub-category, which are further justified by computing the WEAT scores in relative levels and the proportion of each sub-category in different classified gender.

Upon considering all the findings in Word2Vec cosine similarity, WEAT, and BERT classification proportion results, gender bias is discovered in wiki biographies. Such result also falls into the general subject stereotypes, in which sentences related to males tend to associate more with terms in work, science, and math categories but sentences related to females tend to associate more with terms in art and literature categories. The conclusion is also consistent with many literary works as mentioned in section 2.

The current methodologies choose not to modify the data because the imbalance in data collected from all available wiki-bio text can be considered as another evidence of gender bias since it reflects the true gender ratio, and retains the original size of the data set can ensure the accuracy of BERT models. However, the project can become more comprehensive if more methods are used to adjust the impacts of class imbalance in the future, such as using SMOTE to up-sample the minority data. If the two classes can become relatively balanced, then the analysis can focus more exclusively on the terms that are used to describe both genders.

References

[1] Chaloner, K., & Maldonado, A. (2019, Augustus). Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 25–32. doi:10.18653/v1/W19-3804

[2] Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corpora of More Than 65 Million Words. *Psychological Science*, 32(2), 218–240. https://doi.org/10.1177/0956797620963619

[3] Lebret, R., Grangier, D., & Auli, M. (2016). Generating Text from Structured Data with Application to the Biography Domain. *CoRR, abs/1603.07771*. Opgehaal van http://arxiv.org/abs/1603.07771

[4] Reagle, J. & Rhue, L. (2011). Gender Bias in Wikipedia and Britannica. *International Journal of Communication 5*, (2011), 1138–1158. https://ijoc.org/index.php/ijoc/article/viewFile/777/631

[5] Tripodi, F. (2021). Ms.Categorized: Gender, notability, and inequality on Wikipedia. *New Media Society*. https://journals.sagepub.com/doi/10.1177/14614448211023772

[6] Field, A., Lin, K. Z., Park, C. Y., & Tsvetkov, Y. Controlled Analyses of Social Biases in Wikipedia Bios. *arXiv:2101.00078v4*. https://arxiv.org/pdf/2101.00078.pdf

[7] Wagner, C., Garcia, D., Jadidi, M., Strohmaier, M. (2021). It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. Proceedings of the International AAAI Conference on Web and Social Media, 9(1), 454-463. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14628

[8] Graells-Garrido, E., Lalmas, M., Menczer, F. (2015). First Women, Second Sex: Gender Bias in Wikipedia. CoRR, abs/1502.02341. Opgehaal van http://arxiv.org/abs/1502.02341

[9] Adams, Kimberly; Alvardo, Jesus (27 July 2021). "Why it's so hard for biographies about women to stay on Wikipedia". Marketplace. Retrieved 3 August 2021

[10] Amber G. Young, Ariel D. Wigdor Gerald C. Kane (2020) The Gender Bias Tug-of-War in a Co-creation Community: Core-Periphery Tension on Wikipedia, Journal of Management Information Systems, 37:4, 1047-1072, DOI: 10.1080/07421222.2020.1831773