12-12-2019

Overrepresentation of the Underrepresented: Gender Bias in Wikipedia

Anna Marinina

Overrepresentation of the Underrepresented: Gender Bias in Wikipedia

Anna Marinina

Computer Science

Advisor: Yegin Genc

Presentation: December 12, 2019

Graduation: December 20, 2019

Abstract

The goal of our research is to determine if gender bias exists in Wikipedia. Wikipedia is a very large dataset that has been used to train artificial intelligence models. If a dataset that is being used for this purpose is biased, then the artificial intelligence model that was trained with it will be biased as well, therefore making biased decisions. For this reason, it is important to explore large datasets for any potential biases before they are used in machine learning. Since Wikipedia is ontologically structured, we used graph theory to create a network of all of the website's categories in order to look at the relationships between men-related categories and women-related categories with measures of shortest paths, successor intersections, and average betweenness centrality. We found there is an overexposure of categories that relate to men as they are far more central in Wikipedia and easier to get to than categories that relate to women. However, although women-related categories are not as central, there are about six times more categories that mention women in the title than men, which we consider to be overrepresentation. This is most likely due to women being considered an exception in many fields while men are considered the norm.

Our methods can be used to either periodically study gender bias in Wikipedia as its data changes relatively frequently or our methods can be used to study other biases in either Wikipedia or other network-like datasets.

Table of Contents

List of Figures

Introduction

Artificial intelligence (AI) has become an essential tool in decision making. Whether it is used for sifting through resumes to highlight key candidates or for anticipating customer behavior in financial services, these machine-made decisions are meant to augment human-made decisions by replacing systemic biases inherent in human judgment—such as gender or race biases—with objective, data-driven decisions. However, recent studies have found AI-based decisions to show signs of racial bias in aiding criminal defense decisions [1], gender bias in supporting hiring decisions [2], and a mixture of biases in aiding policy-making [3]. Many of these issues stem from inherent biases in the large datasets that these AI models learned from. These biases can be accounted for if AI modelers can detect them. Therefore, the purpose of our study is to focus on detecting a particular systemic bias, i.e. gender bias, within a large textual data source that has been widely used to train AI models: Wikipedia content.

Our general method for detecting bias is using graph theory to analyze relationships in Wikipedia data. This method is fairly novel as it appears to have been used to study gender bias only once before. While many researchers have studied gender bias in Wikipedia in various ways that we will discuss in the next section, findings have differed over the years as Wikipedia evolves. Therefore, it is important to continue looking at and drawing attention to potential biases in the data so AI developers can work to minimize them as much as they can. Our research builds on top of past work by adding a layer to the bigger picture and sets a foundation for others to pick up where we leave off by using our methodology to either analyze any kind of systematic biases in either Wikipedia or other network-like structures.

Review of Literature

Bias in Computers and Artificial Intelligence

Bias in computers and artificial intelligence has been studied for many years across many contexts such as race, gender, and even page ranking in Google searches. There is an extensive amount of material on the topic dating as far back as the 1980s when it was claimed that two reservation systems that matched customers with flights that fit all of their criteria or came close were biased against international customers who rarely take internal U.S. flights and against domestic customers who rarely fly internationally [4]. However, while bias in computer systems was noted and briefly discussed as far back as thirty years ago, it appears that Batya Friedman and Helen Nissenbaum [4] were the first to focus exclusively on bias in computers and study it on a deeper level, as they claim themselves in their 1996 paper, "Bias in Computer Systems."

Friedman and Nissenbaum define and outline three overarching types of bias they found after examining seventeen computer systems from various fields: pre-existing bias, technical bias, and emergent bias. Pre-existing bias is described as bias that is already instilled in social institutions and has its roots in society or culture at large [4]. This pre-existing bias is translated into a computer system primarily by developers or other individuals who were heavily involved in the system design, even if they were making conscious efforts to prevent this from happening [4]. Technical bias, on the other hand, arises from various technical constraints. Examples of this would be if there are imperfections in a random number generator which may cause it to not be truly random, or when qualitative variables are attempted to be quantified [4]. Lastly, emergent bias of a design takes some time to come to light after a change in societal knowledge or cultural values and is mostly related to user interface designs. One population may find a specific user

interface to be very intuitive and user-friendly if it properly reflects their character and habits, but if there is a change of context, this design may prove to be unsuitable for a new population [4]. The given example of this is an ATM that has an interface that displays only written instructions. If this ATM is installed in a predominantly nonliterate area, this interface is only useful for a small group of people and therefore unintentionally biased in favor of the literate.

While Freidman and Nissenbaum focused on bias in computer systems, researchers Eishvak Sengupta et al. concentrated on bias specifically found in artificial intelligence [5]. In their 2018 paper, "Techniques to Eliminate Human Bias in Machine Learning," they outline three types of cognitive human biases that artificial intelligence is prone to as well: interaction bias, latent bias, and selection bias. All three of these biases stem from the dataset that was used to train the machine. Interaction bias occurs if a dataset primarily has entries of one particular type instead of a more well-rounded representation [5]. For instance, if a machine was asked to recognize telephones from a series of images but the dataset it was trained with contained primarily images of smartphones, the machine would have difficulty recognizing rotary phones. Next, latent bias occurs when a dataset has some elements that have a prominent common characteristic [5]. If there are enough of these kinds of elements, the data entries that do not contain the common characteristic may be overlooked. As an example, this paper mentions a 2018 study [6] done on the accuracy of various facial recognition algorithms made by IBM, Microsoft, and Megvii, which found that the algorithms most accurately classified gender when white men were shown while having difficulty classifying darker-skinned faces [5]. The dataset used to test these facial recognition algorithms contained 1,270 faces of politicians from different countries, including a high percentage of women. Because these facial recognition algorithms

succeeded mainly with white male faces when tested with a diverse input, it can be said that the algorithms contained a latent bias. Lastly, selection bias arises in an algorithm when the selection of data to be used to train the algorithm cannot be properly randomized [5]. Continuing with the facial recognition example, this kind of algorithm could have a selection bias in addition to other potential biases because the dataset cannot contain every type of face that exists due to many various kinds of possible facial structures. The more useful data that is missing from a dataset, the less random the dataset becomes, which in turn leads to a higher chance of selection bias [5].

While both of these papers describe different kinds of biases across different levels of computing, they also prescribe potential ways to reduce the aforementioned biases. Friedman and Nissenbaum state that it is important to first identify bias in any system [4]. Next, methods need to be developed to avoid bias or to correct it once it has been discovered. Friedman and Nissenbaum offer advice specific to minimizing each of the three biases they had described. First, to minimize pre-existing bias, they suggest scrutinizing design with a good understanding of biases that exist in society at large. This needs to be done in the early stages of the design process and must be discussed with the client who is requesting the program. Minimizing bias in a system is much easier before it is actually built. Next, to minimize technical bias, designers must think critically about the way their system is making decisions to make sure it is not going against some moral values [4]. Lastly, to minimize emergent bias, it is important for designers to consider all of the diverse social contexts in which their program or system will be used. To be more specific, designers should anticipate probable contexts of system use and base their design off of them [4]. If it is not possible to design the system for some specific context, there should be constraints in place for the contexts that are appropriate for use of the system.

While Friedman and Nissenbaum offer specific solutions for minimization of every bias they identified, researchers Eishvak Sengupta et al. suggest more general solutions for minimizing bias as a whole, whatever kind it may be. Because artificial intelligence contains the same biases that humans do, they suggest methods to reduce human bias first. These methods are finding comprehensive data, experimenting with various datasets, increased diversity in the technology workforce, external validity testing, and auditing [5]. While "Bias in Computer Systems" and "Techniques to Eliminate Human Bias in Machine Learning" describe different kinds of biases across various levels of computing and offer different solutions, the conclusion that can be made is that it is up to the designers and developers to first identify any bias that exists, or think about how bias could arise from their design, and weigh various design options to choose the one that would best minimize any unfairness.

Studies of bias in computers mainly focus on the potential ways to avoid biases or minimize them once they are detected. Especially in the field of AI, the detection processes require making sense of complex prediction models, as well as large and unstructured data. Somewhat paradoxically, we need computational approaches to accomplish them. In this research, we address the bias detection problem by examining biases in large collaborative knowledge domains. The implications of successful bias detection in such domains are two-fold: first, since they are collectively curated, they can help us understand "human bias" as explained above; secondly, since such knowledge domains are commonly used for training AI models, they help model builders minimize such biases during their development efforts. Particularly, our research explores the ontological structure of Wikipedia to see if gender bias exists within the way the data is organized. While we are not suggesting any specific ways to minimize bias,

pointing out its existence and specific manifestations is still an important step towards minimizing any unfairness. In fact, this is directly contributing to a suggestion that Friedman and Nissenbaum made in regard to remedying bias: we first need to be able to identify that bias exists in a given system. Our research is also directly concerned with examining a potentially biased dataset before it may be used to train artificial intelligence, which is something that Sengupta et al. discussed in depth.

Gender Bias in Artificial Intelligence

There have been many studies that looked into gender bias in artificial intelligence and datasets but one of the most well-known studies looked at this in terms of word embeddings. Word embeddings are used as a kind of dictionary for computer programs that need to utilize word meaning [7]. Even though word embeddings have been researched extensively, researchers Tolga Bolukbasi et al. were the first to point out that word embeddings were highly sexist [7]. They came to this conclusion by looking at various analogies and word counts. Since word embeddings are represented as vectors, finding analogies is done by looking at distances between these vectors. First, Bolukbasi et al. tested the system to make sure it was reasonable by finding proper analogies like "sister-brother," "waitress-waiter," and "mother-father" [7]. However, when given the input, "Man is to computer programmer as woman is to *x*," the system output was *x* = "homemaker." As another example, when given the input, "Father is to doctor as mother is to *x*," the system output was *x* = "nurse" [7]. Aside from analogies, the researchers also looked at word counts such as how many times "male nurse" appears versus "female nurse." It turns out that the phrase "male nurse" is much more frequent, similar to the result that the phrase "female quarterback" was much more frequent than "male quarterback" [7].

Aside from demonstrating various examples of gender bias in word embeddings, Bolukbasi et al.'s true focus was developing some algorithms to debias them. The first step involves identifying gender subspace where the direction of the embedding that contains bias is identified [7]. The second step has two options: hard debiasing or soft debiasing. Hard debiasing consists of "neutralizing" and "equalizing." The process of neutralizing makes sure that gender neutral words have a value of zero and the process of equalizing makes sure that every gender neutral word is equidistant to all words in the "equality set" [7]. For example, if {*mother*, *father*} was an equality set, then after equalization, the word {*cooking*} would be the same distance away from *mother* as it is to *father*. However, equalization holds a disadvantage where it eliminates certain distinctions that may be useful in some contexts and applications [7]. This is where the process of soft debiasing comes in. The "soften" algorithm is able to reduce the distinctions between sets while still holding as much similarity to the original word embeddings as possible [7]. Even though this means the bias is not completely gone, it is still minimized.

Bolukbasi et al.'s study inspired many other researchers to look into gender bias, both within the realm of word embeddings and in other contexts as well. A study by Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan [8] regarding word embeddings directly stems from Bolukbasi et al.'s work. Caliskan et al.'s paper complements Bolukbasi et al.'s work by focusing on demonstrating human-like biases in word embeddings, some of which are specifically related to gender. To measure bias, the researchers used the Implicit Association Test (IAT), which was first introduced by Greenwald et al. [9]. The test works by measuring the association of two concepts with an attribute. The exact measurement is found by the difference between two means of log-transformed latencies in milliseconds, divided by the standard deviation, which is also

known as Cohen's *d* [8]. To start, the IAT in Greenwald et al.'s work yielded results where the categories *flower* and *pleasant* had a high level of association whereas the categories *insect* and *unpleasant* also had a high level of association. Caliskan et al. take this concept and technique to demonstrate that machine learning absorbs various stereotyped biases [8]. To focus specifically on gender bias, Caliskan et al. were able to replicate a prior finding that female names are more associated with family related words than career related words than male names are [10]. They were able to replicate this finding using far fewer keywords than the original study, which proves the strength and persistence of this bias. Furthermore, Caliskan et al. replicated other findings using the IAT that showed female words such as "woman" or "girl" are more related to arts than mathematics [10] or science [11].

While many studies are able to prove gender bias exists in various computer systems, others also mention that the gender bias that exists in society is not only reflected by these systems but amplified as well [7]. This concept of bias amplification has been mentioned in a study by Matthew Kay, Cynthia Matuszek, and Sean A. Munson where they looked at Google image search results of forty-five different occupations and found unequal gender representation and gender stereotypes in the images [12]. More specifically, they found that images of a person who matched the majority gender of their occupation tended to be more professional and less inappropriate than images of a person who went against a gender stereotype of the occupation they were portraying [12]. For example, images of female construction workers were highly sexualized and, in general, there were many images of various professions where women were much more sexualized than men. As for bias amplification, it was found that occupations that are stereotypically male-dominated preferred image results of men and occupations that are

stereotypically female-dominated preferred image results of women [12]. This is, again, another example of artificial intelligence reflecting a larger societal bias.

In addition to bias amplification, this study found instances of both systematic underrepresentation and overrepresentation. For example, it was found that an occupation with 50% women can expect to have 45% women in the image search results for that occupation. However, when people look at these image search results for various professions, they prefer seeing a gender that is stereotypically associated with the job [12].

Gender Bias in Wikipedia

The idea of systematic underrepresentation, specifically of women, has been examined on a deeper level in the context of Wikipedia by various researchers. In fact, gender bias in the form of underrepresentation is what is most commonly looked at in the realm of Wikipedia. One example of this is a study done by Menno H. Schellekens, Floris Holstege, and Taha Yasseri where they analyzed the probability of being listed as a scientist on Wikipedia for both male and female scholars among similar levels of academic achievement [13]. To do this, the researchers used data from Google Scholar and checked it against Wikipedia entries. More specifically, they looked at citation count and gender of each scholar to determine the probability of appearing on Wikipedia. The results showed that women are less likely to be recognized than their male counterparts across all levels of achievement and they need to achieve more for equal recognition [13]. The researchers found that out of men and women of the same h-index--therefore of the same level of accomplishment--men have a higher probability of appearing on Wikipedia. Furthermore, women need to have a higher h-index than men in their field to have the same level of recognition [13].

Another study that focused primarily on gender representation in Wikipedia was done by Joseph Reagle and Lauren Rhue where they looked at women's biographies in terms of coverage, gender representation, and article length in both Wikipedia and Britannica to see if one was more representative of women than the other [14]. To gather their data, the researchers used a Python program to find and compare Web pages related to biographical subjects that were chosen from sources such as *The Atlantic*'s 100 most influential figures in American history and American National Biography Online. Then, they used a Google API for language queries to find the top four article results for a given person on Wikipedia and Britannica. The results were that Wikipedia has greater coverage of women's biographies than Britannica [14]. However, it was found that Wikipedia covered men's biographies more comprehensively than women's. While both notable men and women were missing some articles, women were overall less represented.

Other studies have looked at gender bias in Wikipedia beyond primarily underrepresentation. Claudia Wagner et al.'s study investigated four types of bias in Wikipedia related to gender: coverage bias in terms of underrepresentation, structural bias in terms of how articles of notable people tend to be linked, lexical bias in terms of inequalities in the words used to describe notable men and women in their biographies, and visibility bias in terms of which gender is more likely to make it to the front page of Wikipedia [15]. This study, which was done four years after Reagle's and Rhue's, actually found that men and women were covered equally well with a slight overrepresentation of women [15]. Additionally, Wagner et al. did not find any bias in visibility, meaning men and women are equally as likely to be on the front page of Wikipedia. However, they did find that the way women are portrayed lexically differs greatly from how men are portrayed and articles about women are more likely to be linked to articles

about men than vice versa [15]. Wagner et al. utilized Freebase and Wikipedia's API to gather their data and used Pantheon to analyze and visualize their data.

While all of these studies look at gender bias in Wikipedia, the work that is most comparable to ours was done by Graells-Garrido et al. where they looked at gender bias in Wikipedia in terms of meta-data, language, and network structure [16]. More specifically, in terms of network structure, the researchers built a directed network graph of biographies from links between articles in the *Person* DBPedia class, a structured version of Wikipedia, and compared it to various null graphs in order to find structural differences between genders by looking at measures of node centrality [16]. Their key findings were that women's biographies contained more marriage-related content and more sex-related content than men's biographies while sports-related content was more related to men than women. This reinforces the results of Wagner et al.'s study where they found a lexical difference in the way men and women are portrayed on Wikipedia. Aside from this, Graells-Garrido et al. found evidence of a strong bias in the way pages are linked which results in articles about men being more central than articles about women. The researchers used DBPedia and the Wikipedia English Dump of October 2014 to gather and analyze their data.

Our research is similar to all mentioned Wikipedia studies in the sense that we are looking at differences in representation of men and women. As exemplified, these findings have changed over time so it is still worth investigating. However, our research is primarily concerned with structure bias, much like Graells-Garrido et al.'s [16] and Wagner et al.'s [15] studies and less with any lexical differences in pages related to men and women. Out of all mentioned studies, Graells-Garrido et al.'s research is most similar to ours because we share the focus of

looking at the network structure and various network properties of linked Wikipedia pages. Even though the focus is the same, their study was published four years ago and Wikipedia has changed. We are building on top of past research to paint a bigger picture and demonstrate the potential evolution of gender bias in Wikipedia.

Graells-Garrido et al.'s methods of data collection and analysis are most similar to ours as well, which is unsurprising because of the common focus on network structure. Our specific methodology is to be discussed in the next section.

Methodology

Our research explores the ontological structure of Wikipedia using graph theory to detect any gender bias that may exist in the data. Therefore, the first step was obtaining our dataset from a Wikipedia data dump [17]. We used a Jupyter notebook on a Pace University server to store our data and analyze it. The Wikipedia dataset is vast and therefore required a powerful machine for the most efficient analysis. Once we had our Wikipedia data, we created a directed graph of it using NetworkX, a Python library used for visualization and analysis of large networks. Each category on Wikipedia became a node in the graph and each link between two categories became an edge. For instance, the *Physical Exercise* category can directly take you to the *Dance* category and vice versa. So, *Physical Exercise* and *Dance* are nodes with an edge between them. Using a graph to analyze Wikipedia data is very useful for looking at paths between categories to see how they are linked to each other and therefore gauge relatedness, which is key when searching for bias.

Since we were specifically looking for gender bias in the data, we shifted our focus to Wikipedia categories that contained "Women," "Women's," "Men," and "Men's" in their titles

and created another directed graph from those categories. We began by gathering some quantitative data such as the number of categories that contain “Women” or “Women’s” in the title versus the number of categories that contain “Men” or “Men’s” in the title. We then divided our data further by looking at different topics such as location, warfare, sports, politics, the arts, and STEM to see how many “Men” and “Women” categories contain keywords related to each of those topics. To start, we did this by identifying the highest parent of relevant categories to be counted. For example, if we were finding the number of “Women” categories that are related to different locations, we would look at the category *Women in Places* and count that category’s predecessors. When we consider a Wikipedia category as a node in the network of categories that are connected to each with parent and sub-category links, we can apply principles of graph theory to make sense of this categorical structure. For example, a subcategory in Wikipedia can be theorized as the predecessor of a parent node (parent category). In the case of our earlier example of the *Physical Exercise* and *Dance* categories, *Dance,* the subcategory*,* is a predecessor of *Physical Exercise,* the parent category. Finding predecessors of a category is then a simple graph search task that can be achieved with NetworkX, as the library contains a built-in function for this. For topics like “sports” that contained a much larger amount of relevant nodes, we had to recursively look at the predecessors of predecessors, as many levels of relevant nodes as there were. However, sometimes the predecessors of a relevant node contained some irrelevant nodes that should not be counted. This was especially the case when looking at “Men” and “Women” nodes related to location. In this case, string matching was required to find categories that contained any country or state in the title, such as categories like *Men in the United States* and *Women in the United States* and excluding categories that did not. Similarly, in the case of

finding nodes related to politics, we used string matching to extract categories that contained both “politics” and either “Men” or “Women” in the title and found relevant predecessors stemming from them. By dividing “Men” and “Women” categories further into the different topics they relate to, this helps us look more closely into the data to see which specific areas, if any, contain gender bias as either underrepresentation or overrepresentation.

In addition to dividing the data by various topics, we looked at intersections between the successors of every “women” node and the successors of every “men” node. A successor of a category in NetworkX is a supercategory. In the case of our earlier example of *Physical Exercise* and *Dance*, *Physical Exercise* is a successor of *Dance*. Similarly to finding predecessors of a node, NetworkX contains a built-in function to find successors of a node. Once we had a list of successors of all of our “Women” nodes and a list of successors of all of our “Men” nodes, we used a built-in Python function to return a list of the intersections. Finding the intersections between “Men” and“Women” nodes allows us to see where they converge or where they diverge if there are some nodes that have no intersections. This enables us to see what is “common” between men-related and women-related content on Wikipedia and what is not common. Furthermore, this helps us determine which nodes are mirrored. We define mirrored nodes to be nodes that exist for both men and women. For example, if there exists a *Women in Basketball* category and there exists a *Men in Basketball* category, these nodes are mirrored. If one of those nodes exists and the other does not, they are not mirrored. Looking at these kinds of nodes, or perhaps the absence of them, is important in determining potential bias and in what genres specifically.

The next step is the heart of this research paper: analyzing the shortest paths between categories. In graph theory, a shortest path between two nodes is a measure of the distance between them and in our case, can be used to measure the "relatedness" of two categories. Due to Wikipedia's ontological structure, a user can be taken directly from one category to another by continuously clicking on one of each category's predecessors. For example, one path from the category *Plant* to the category *Rose* is *Plant* → *Pollination* → *Rose*. If a user started at the *Plant* category on the Wikipedia website, there would be a list of its direct successors and predecessors where the user could then click *Pollination*, and from *Pollination,* they could click *Rose.* This path is two hops long. Since there are no paths between *Plant* and *Rose* that are shorter than two hops, we can say the length of the shortest paths between these categories is two. However, there may be other paths of length two between *Plant* and *Rose* which would be considered shortest paths as well.

Following this principle, we picked 107 different categories in Wikipedia and looked at the shortest paths from each of those categories to both the *Men* category and the *Women* category, comparing the number of shortest paths between categories and how long any shortest path was, as every shortest path was the same length. We chose categories that we expected to have the highest chances of being skewed towards either men or women, such as *Housewives*, *Computer Scientists,* and *Sports.* In general, the categories we chose can be grouped by occupations and fields of study, the arts, various human qualities such as *Emotional Intelligence* and *Altruism*, family roles, mental health, religion, and some general phrases like *Gender Roles* or *Bias*.

Once again, the NetworkX library made data collection very simple as it contains a built-in function that provides all of the shortest paths between two categories. To further simplify data collection, we wrote our own Python function that takes the index of a category of our choosing and, using NetworkX's all_shortest_paths() function, returns the number of shortest paths and the length of these paths between the input category and *Men,* and between the input category and *Women.* It is important to note that the all_shortest_paths() function accepts a "source node" and a "target node," and this input order matters. Specifically, the source node is considered to be the predecessor and the target node is considered to be the successor. Therefore, if we were looking at the shortest paths between *Sports* and *Men*, we needed to check *Sports* → *Men* as well as *Men* → *Sports* because those two inputs yield different results. The difference lies in which category the user would start from to work their way down the line of predecessors before reaching the target category.

We used our quantitative shortest paths data to determine the "relatedness" of two categories. The more shortest paths that existed and the shorter they were in length, the more related the two categories were considered to be. This is because having more paths means there is a higher chance for users to reach one category from the other since there are more options, and having a shorter path means this will happen much quicker. Sticking with the *Sports* category as an example, if there is a greater number of shortest paths between *Sports* → *Men* than *Sports* → *Women*, and if the length of any shortest path between *Sports* → *Men* is shorter than the length of any shortest path between *Sports* → *Women*, we can say that *Sports* is more related to *Men* than it is to *Women.*

Since many of the 107 categories we chose to test relatedness to either *Men* or *Women* can be grouped into general topics, we calculated the average number and average length of shortest paths from all 107 categories to both *Men* and *Women*, as well as the average number and average length of shortest paths from various groups within the 107 categories to *Men* and *Women.* The former helps us gauge if either *Men* or *Women* is more central to all other Wikipedia categories and the latter helps us gauge if *Men* is more central to certain groups of categories than *Women* or vice versa. Therefore, finding especially high discrepancies in average shortest paths data between categories as they relate to men and women is another way of determining if there is a gender bias in Wikipedia data, "high discrepancies" being defined as either *Men* or *Women* having at least double the amount of shortest paths to the same category than the other.

While looking at shortest paths data for *Men* and *Women* can help us make some conclusions about whether or not one of the two is more central to all other Wikipedia categories, we needed specific measurements of betweenness centrality, a measure of centrality in a graph based on shortest paths, to confirm our hypothesis. This was done by another built-in function in NetworkX that returns the betweenness centrality of every node in a graph. Since we had one Wikipedia sub-graph for women and another for men, we checked betweenness centrality of all nodes in each graph and then found the average for each graph using a built-in Python function. If the betweenness centrality of either the men or women graph is higher than the other, we can say that whichever graph has higher betweenness centrality has more central nodes, or categories, in Wikipedia, leading to further evidence of potential gender bias.

However, this quantitative data is not fully sufficient without some qualitative data to paint a better picture of how categories are connected. As *Plant* → *Pollination* → *Rose* is a specific shortest path from *Plant* to *Rose*, the nature of our shortest paths and which categories they contain are very important. If there is a high discrepancy between *Men* and *Women* for a category, the reason or explanation for it could lie in the actual chains of categories that make up the shortest paths. Therefore, this was something we looked into after gathering our quantitative data before drawing any final conclusions.

Results and Discussion

After obtaining our dataset of Wikipedia and creating a directed graph of it using NetworkX, our graph contained 1,782,726 nodes and 5,131,621 edges between them. After extracting all of the categories that contained "Women" or "Women's" and "Men" or "Men's," and creating separate graphs of this data, we found there are 5,539 nodes relating to women and 954 nodes relating to men. This means there are nearly six times more categories that mention women than men in Wikipedia. Visualizing this data as a standard graph is not effective because the large number of nodes in each graph cluster together and make it impossible to discern anything. Therefore, we used Datashader, a Python library, to visualize differences in the graphs in another way.

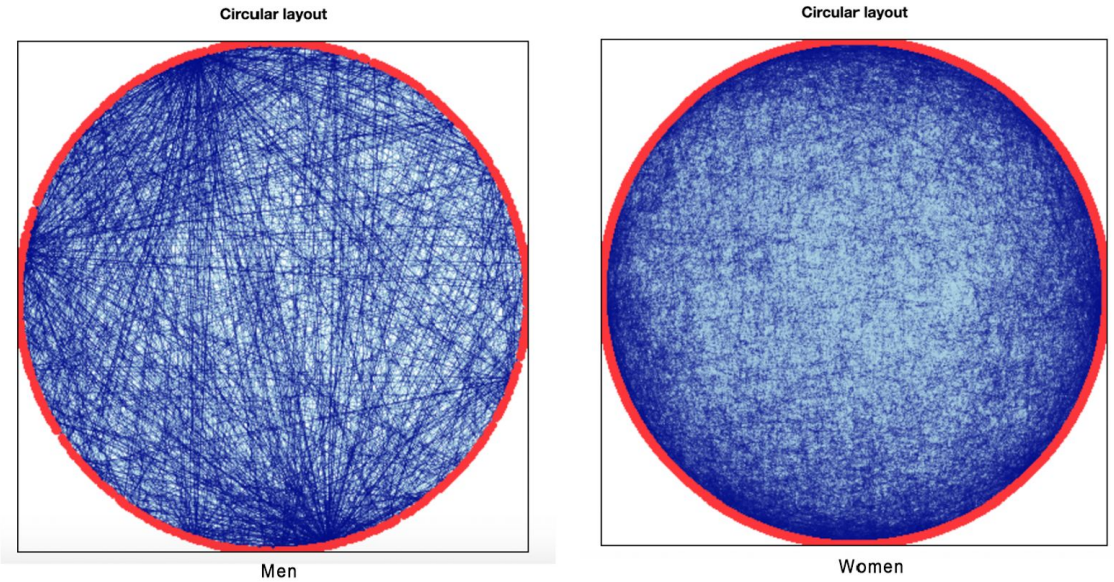

Figure 1

This circular layout arranges all nodes (in red) along the outside of the circle and draws lines across (in blue) to represent edges. According to our data, the men graph contains 4,981 edges while the women graph contains 28,536. This difference is seen in Figure 1 by the fuzziness of the women graph as opposed to the men graph that has more clarity due to the smaller number of edges.

To go a bit further, we extracted all of the categories that contain the phrase "Women In" and all of the categories that contain the phrase "Men In." We found that there are 639 categories that contain "Women In" and only 259 categories that contain "Men In." Immediately, the high disparity in numbers stood out as hinting towards overrepresentation of women.

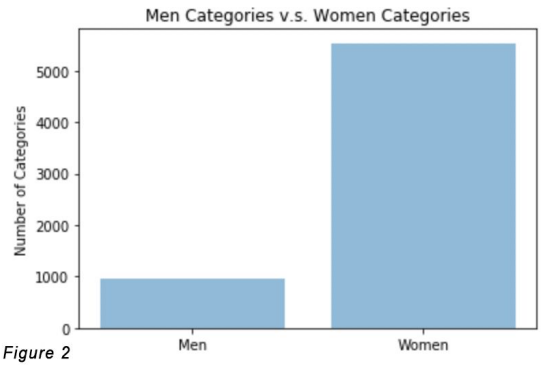

Figure 2

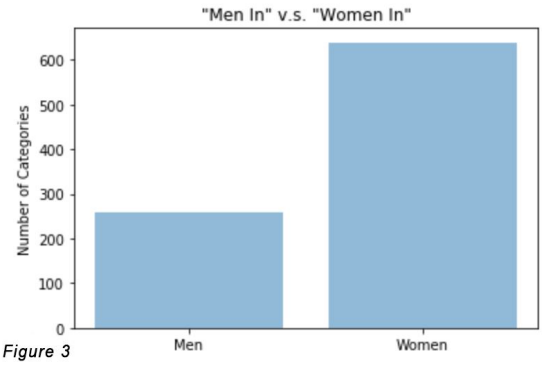

Figure 3

To analyze where this disparity may lie, we divided the women nodes and men nodes by genre, such as location, sports, warfare, politics, the arts, and STEM to see how many categories that contain "Women" or "Men" also contain words related to each of those genres. We found that the number of women categories for each genre outweighed the number of men categories, with the exception of the "sports" genre. In general, the sports genre contained much more men-related and women-related categories than any other genre so it needed to be graphed separately.

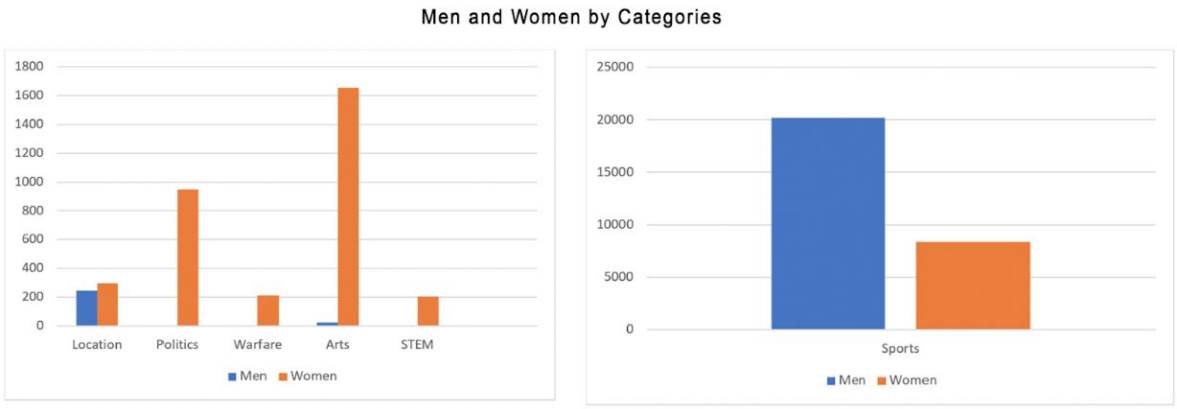

Figure 4

Based on Figure 4, the "location" genre contains the least disparity between men and women categories while the "sports" genre contains the greatest disparity by far. By "location", we mean categories that contain "Men In" or "Women In" followed by a country, continent, or state. It is interesting to note that "Women In" contains a category for every state in the United States, whereas "Men In" does not. For example, there are categories titled "Women in Arkansas" and "Women in Delaware" but there are no categories titled "Men in Arkansas" or "Men in Delaware." Some of the more striking results are that men have no categories specifically mentioning them with words related to politics, warfare, or STEM fields. So, for Wikipedia categories that contain titles relating to politics, war or STEM, gender is only mentioned when the category is related to women. For example, there is a category titled, "Women in war," but there is no such category titled, "Men in war."

In addition to dividing our data by genre, we looked at intersections between the successors of every "women" node and the successors of every "men" node. After creating a list of successors for both genders, we found that in total, there were 4,981 successors for the "men" nodes and 28,536 successors for the "women" nodes. We then took both lists and used a Python

function to find 856 intersections between all of the successors. We found that most of these intersections were categories primarily related to sports and gender in various places, as well as some general categories such as “Health” or “People by nationality and occupation.”

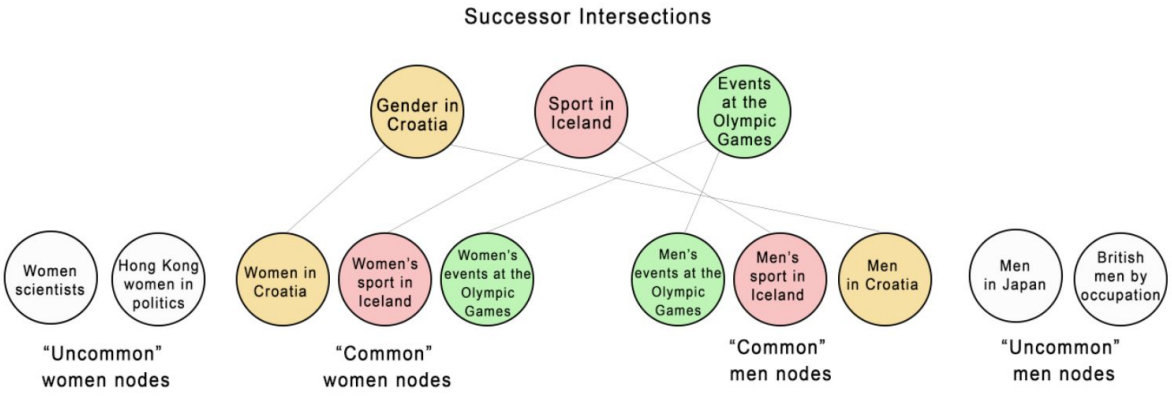

Figure 5

While it is important to see the intersections of successors, it is equally as important to see which categories were not part of the intersections. We found there to be 863 men-related categories and 8,320 women-related categories that were not part of the successor intersections. When looking at what these “uncommon” categories are, we found there to be a difference for men and women, aside from sheer quantity. The list of uncommon categories for men contained primarily sports-related and location-related titles while the list of uncommon categories for women contained these as well, but with the addition of many categories related to politics, STEM, warfare, and various general occupations. It might make sense for some successors to not be intersected if the category title specifically mentions either men or women, but what is striking is that many of these categories do not. So, many categories with general titles about politics and various occupations are successors of categories that specifically mention women but not of those that specifically mention men. For example, categories such as *Kansas*

politicians and *Video game programmers* are successors of women categories but there are no such kinds of successors of men categories.

Next, we picked 107 different categories in Wikipedia and looked at the shortest paths from each of those categories to both the *Men* category and the *Women* category, comparing the number of shortest paths and the length of any shortest path. As previously explained, it was important to look at paths bidirectionally from *Men/Women* → *Category* and *Category* → *Men/Women* as these produce different results. The more shortest paths exist between two categories and the shorter they are, the more "related" these two categories are considered to be.

With this in mind, we found that the average number of shortest paths from *Men* → any category is 3.26 and the average length of these paths is 8.15. The average number of shortest paths from *Women* → any category is 2.80 and the average length of these paths is 8.44. While the number of paths for men is greater and the length of these paths is shorter than women, these numbers are fairly close. These numbers are very close to each other because the paths from *Men/Women* to any category tend to be mirrored or extremely similar. Therefore, when analyzing shortest paths, we will mainly be focusing on the opposite direction as it is more revealing of differences. When looking at this opposite direction, we found that the average number of shortest paths from any category → *Men* is 13.36 and the average length of these paths is 11.67, while the average number of shortest paths from any category → *Women* is 5.36 and the average length of these paths is 14.21. There is a striking difference here, with men having nearly 250% more paths to 107 different categories than women, with paths that are about 120% shorter. This means that even though women have a much greater number of categories

than men, women have fewer links to other categories and it is much easier for Wikipedia users to travel from any category to *Men* than to *Women*.

While there are many categories that have a very high disparity in the number of shortest paths and their length, some of the ones that really stand out are *Housewives*, *Politics*, and *Surgeons.* There are 30 paths of length 14 from *Housewives* → *Men* and only one path of length three from *Housewives* → *Women.* This is highly unexpected since the word *Housewives* is gendered towards women in the name. Although it is striking that there are 30 times more paths from *Housewives* to *Men* than *Housewives* to *Women,* it is important to note that the shortest path from *Housewives* to *Women* only contains three hops, as opposed to fourteen, meaning *Housewives* is about 4.67 times closer to *Women* than *Men.*

As for *Politics*, there are 86 paths of length 14 from *Politics* → *Men* but only 6 paths of length 14 from *Politics* → *Women.* Since the length of the paths are the same, we can simply take the number of paths into consideration. According to this data, men are 14.3 times more related to politics than women are.

Next, looking at *Surgeons*, there were no paths from *Men* → *Surgeons* or from *Women* → *Surgeons* but there were 36 paths of length 16 from *Surgeons* → *Men* and only six paths of length 18 from *Surgeons* → *Women.* In this case, men have six times more paths to *Surgeons* and the length of these paths is shorter so we can say that men are considered to be more closely related to *Surgeons* than women.

At this point, we would like to bring attention to a relationship between our shortest paths data and the data shown in Figure 4. Based on Figure 4, there are no categories that mention "Men" and "Politics" together. However, based on shortest paths, men are considered to be 14.3

times more related to politics than women are. Additionally, there are no categories that mention "Men" and "Warfare" together. However, there are 30 paths from *Warfare* → *Men* of length 13 while there are only 2 paths from *Warfare* → *Women* of length 16 which suggests men are about 15 times more closely related to "Warfare" than women are. This same trend can be seen with categories relating to the arts. Looking at the arts genre, there are once again more categories that mention women than categories that mention men. To compare this to shortest paths, we combined the following categories to look at the average number of paths and the average length: *Artists, Filmmakers, Musicians, Music, Fashion, Fashion Designers, Creativity, Comedians, Stand-up Comedy, Stand-up Comedians, Comedy, Theatre,* and *Writers*. Out of these categories, the average number of paths to *Men* is 25.38 with an average length of 12.38 while the average number of paths to *Women* is 6.31 with an average length of 16.85. Next, looking at the sports genre, we combined the following categories: *Sports, Athletes, Bodybuilding, Baseball, Baseball players, Basketball, Basketball players, Football, Cricket, Golf, Golfers, Swimming, Swimmers, Running,* and *Track and Field*. Out of these categories, the average number of paths to *Men* is 5.5 with an average length of 12.75 while the average number of paths to *Women* is 5.38 with an average length of 16.13. Lastly, to look at the STEM genre, we combined the following categories: *Nurses, Medicine, Computer Science, Computer Scientists, Mathematics, Mathematicians, Science, Scientists, Biologists, Chemists, Physicists, Logic, Physicians, Surgeons, Psychology,* and *Psychologists*. The average number of paths to *Men* is 11.38 with a length of 13.56 while the average number of paths to *Women* is 5.25 with a length of 17.19.

We notice that, in general, women have more categories mentioning them but are less related or central to the topics at hand while the opposite is true for men. Therefore, we can say there are fewer categories that mention two concepts together if the concepts are considered to be closely related. The reason women have a much greater number of categories than men is most likely because they are not the norm in many fields. Wikipedia titles do not mention men when talking about warfare or politics, for example, because historically, these fields have been exclusively populated by men.

Since on average there are more shortest paths from any category to *Men* with shorter lengths, we can see evidence that the general *Men* category may be more central to all other Wikipedia categories than *Women*. As confirmation, we found the measurements of average betweenness centrality for every node in our men graph and women graph. We found that the average betweenness centrality in our men graph is 2.1287004536940604e-06 while the average betweenness centrality in our women graph is 1.8500865865636283e-07. This is an extremely high discrepancy as the men graph's average betweenness centrality is about 11.5 times more than that of the women graph's. This means categories that specifically relate to or mention men are 11.5 times more central in Wikipedia than categories that specifically relate to or mention women.

In addition to this quantitative data, it was very important to look at the nature of these shortest paths. By this we mean it was necessary to look at the actual chain of categories that made up each shortest path. There are two key observations that came from this: first, many of the shortest paths are variations of each other. This means the paths contained the same exact categories in a different order or the paths contained mostly the same categories with the

exception of one or two. Secondly, there were common subpaths that could be found in every shortest path sequence, regardless of what categories the paths were connecting. Shortest paths relating *Men* to any category tended to have the following subpath:

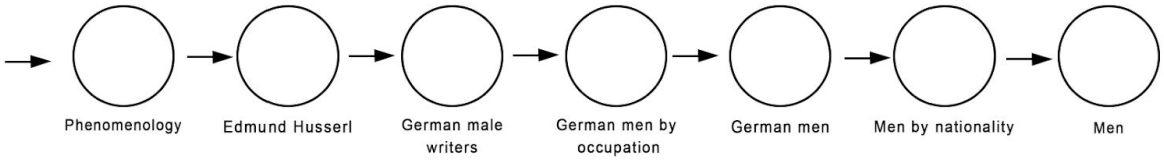

Figure 6 *Most common subpath from any category to Men*

Meanwhile, shortest paths relating *Women* to any category tended to have any of the three following subpaths:

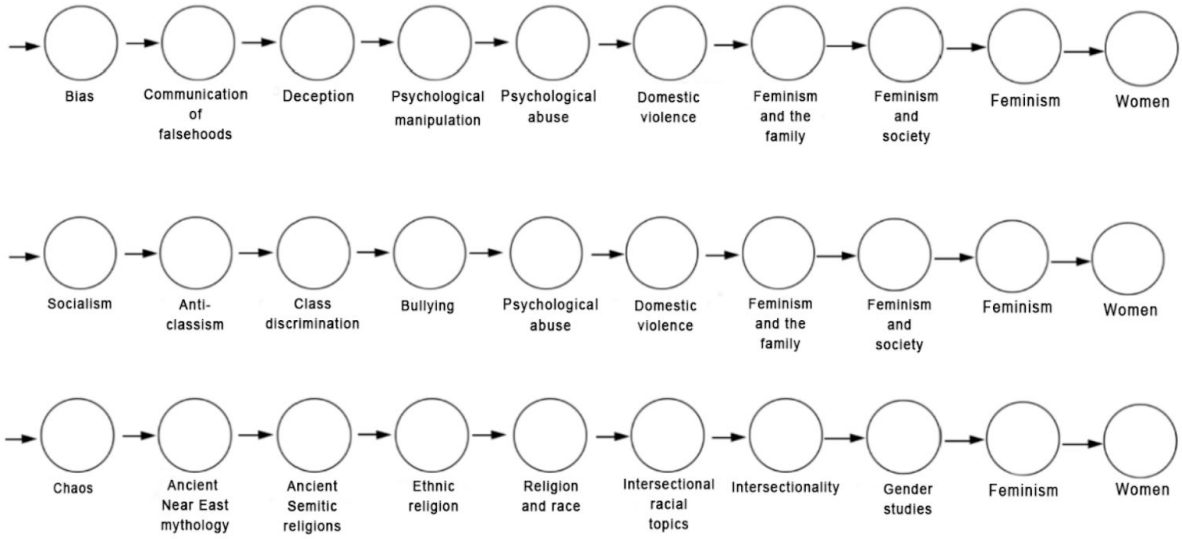

Figure 7 ***Most common subpaths to Women***

The difference is quite striking. Any shortest paths relating *Women* to any category contain some kind of negative categories within them, specifically ones mentioning abuse and violence. Shortest paths relating *Men* to any category contain no such topics, and instead tend to

mention notable men, specifically Edmund Husserl and occasionally, Karl Marx. We found there are no shortest paths relating *Women* to any category that mentions notable women. It is interesting to note that even when looking at shortest paths between *Men/Women* and a positive category such as *Peace*, the paths relating *Women* to this category still contain dark, violent topics. Since most paths relating *Women* to any category contain topics about abuse, violence, or bias, it can be said that these topics are closely related to women and their history while the opposite is true for men.

To summarize, there is overrepresentation of women in Wikipedia as evidenced by the sheer number of categories related to them. This largely comes from many categories that contain "Women in" followed by various topics in the title. This is most likely due to men being the norm in many more fields than women. Since women are treated as an exception, they tend to be explicitly mentioned more than men in the topics they are being connected to. However, despite there being more categories for women, they are not as closely related to these categories as men are based on the number of shortest paths and the length of these paths. This again supports the idea that since men are the norm, they are not explicitly mentioned. Since there are many more paths between *Men* and any category than *Women* and any category, and the paths are shorter, it is easier and more likely for Wikipedia users to travel from any category to *Men*. In this case, this can be considered underrepresentation of women. This also suggests that the *Men* category is more central in Wikipedia than *Women*. Our measurements of betweenness centrality confirm that, actually, any nodes that specifically mention or are related to men are 11.5 times more central in Wikipedia than any nodes that specifically mention or are related to women. Lastly, when looking at the nature of our shortest paths, it seems that if users want to travel from

any category to *Women*, they *must* travel through categories such as *Bias, Psychological Abuse, Deception,* and *Intersectionality.* If users want to travel from any category to *Men,* the most common categories they must travel through are *Edmund Husserl, Karl Marx,* or *German male writers*. The neutrality of the categories in the shortest paths to *Men* is very different from the negativity of the categories in the shortest paths to *Women*. It can be said, therefore, that women are much more associated with being the target of bias and violence while men are not.

With all of this in mind, we can conclude that there exists a gender bias in Wikipedia. More specifically, the gender bias in Wikipedia appears to be a reflection of the bias present in our society. This is especially likely because Wikipedia is created and edited by volunteers, so anyone can contribute to the knowledge that is there. In a way, having gender bias in Wikipedia is a form of accuracy in the provided information since it is reflective of society. However, it is problematic since the way Wikipedia data is organized links more categories to men than women and puts men-related categories in the forefront. Potential ways to minimize gender bias in Wikipedia is not in the scope of this paper, but some steps should be taken to work towards this if this dataset is to be used to develop any artificial intelligence models.

Conclusion

After studying the ontological structure of Wikipedia, we can conclude there exists a gender bias in Wikipedia due to the overrepresentation of women through the large number of categories that mention them, and due to the underrepresentation of women through the number of shortest paths and their lengths from any category to women versus the number of shortest paths and their lengths from any category to men. Although this data is considered to be biased, it is a reflection of the bias present in our society, especially since Wikipedia is created and

edited by anyone who wishes to contribute. This data is reflective of the bias in our society due to the probable reason behind the overrepresentation of women in content and underrepresentation of women in relation to various categories: men are considered to be the norm while women are considered to be the exception. There are many fewer categories that explicitly mention "men" in the title and many more categories that lead to the *Men* category through more shortest paths for this reason. In fact, generally, any categories that mention or are related to men are statistically much more central in Wikipedia than any categories that mention or are related to women.

The purpose of this research is to bring attention to the gender bias that exists in Wikipedia since this dataset has already been used to train artificial intelligence models. If artificial intelligence models are trained with this data, it is highly likely that the artificial intelligence models will contain gender bias as well, therefore making biased decisions. While discussing potential ways to minimize gender bias in Wikipedia is not within the scope of this paper, our research is the first step and lays some groundwork towards this. As mentioned in a previous section, gender bias in Wikipedia has been studied for many years but the findings of each study are different. Since Wikipedia changes, it is important to periodically study its data to see any potential differences or progress. Our research adds onto and expands what others have done by finding different kinds of gender bias in the data, but finding gender bias nonetheless. This leaves a lot of room for others to build upon our research and continue the story. Our methods can be used to study additional other types of bias in Wikipedia as well.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine bias: There's software used across the country to predict future criminals, and it's biased against blacks". ProPublica (May 23, 2016).

[2] "*Amazon scraps secret AI recruiting tool that showed bias against women*," Oct. 9, 2018. Accessed on: Nov. 2, 2019. [Online]. Available:https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[3] Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice, New York University Law Review Online, Forthcoming (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423)

[3] R. Richardson, J. Schultz, K. Crawford, "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice," *New York University Law Review Online, Forthcoming,* Feb. 2019.

[4] B. Friedman, H. Nissenbaum, "Bias in Computer Systems," *ACM Transactions on Information Systems*, Vol. 14, No. 3, pp. 330 –347, Jul. 1996.

[5] E. Sengupta, D. Garg, T. Choudhury, A. Aggarwal, "Techniques to Eliminate Human Bias," *IEEE Conference*, Nov. 2018.

[6] Joy Buolamwini, "Facial recognition software is biased towards white men, researcher finds," *MIT Media Labs*, Feb. 11, 2018.

[7] T. Bolukbasi, K. Chang, J. Zou, V. Saligrama, A. Kalai, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," *Advances in Neural Information Processing Systems*, pp. 4349–4357, 2016.

[8] A. Caliskan, J. Bryson, A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, Apr. 2017.

[9] A. Greenwald, D. McGhee, J. Schwartz, "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test," *Journal of Personality and Social Psychology*, Vol. 74, No. 6, pp. 1464-1480, 1998.

[10] B. A. Nosek, M. Banaji, A. G. Greenwald, "Harvesting implicit group attitudes and beliefs from a demonstration web site.", *Group Dynamics: Theory, Research, and Practice*, Vol. 6, No. 1, pp. 101-115, 2002.

[11] B. A. Nosek, M. R. Banaji, A. G. Greenwald, "Math=Male, Me=Female, Therefore Math≠Me.", *Journal of Personality and Social Psychology*, Vol. 83, No. 1, pp. 44-59, 2002.

[12] M. Kay, C. Matuszek, S. Munson, "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations," *CHI '15*, pp. 3819-3828, Apr. 2015.

[13] M. Schellekens, F. Holstege, T. Yasseri, "Female scholars need to achieve more for equal public representation," *ArXiv*, 2019.

[14] J. Reagle, L. Rhue, "Gender Bias in Wikipedia and Britannica," *International Journal of Communication*, pp. 1138-1158, 2011.

[15] C. Wagner, D. Garcia, M. Jadidi, M. Strohmaier, "It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia," *ArXiv*, 2015.

[16] E. Graells-Garrido, M. Lalmas, F. Menczer, “First Women, Second Sex: Gender Bias in Wikipedia,” *ArXiv*, Jun. 2015.

[17] https://dumps.wikimedia.org/enwiki/latest/ Visited at April 29th 2019