Investigating the Gender Pronoun Gap in Wikipedia

Mehrdad Yazdani

Qualcomm Institute, University of California San Diego | USA

Contact: myazdani@gmail.com

Abstract

In recent years there have been many studies investigating gender biases in the content and editorial processes of Wikipedia. In addition to creating a distorted account of knowledge, biases in Wikipedia and similar corpora have especially harmful downstream effects as they are often used in Artificial Intelligence and Machine Learning applications. As a result, many of the algorithms that are deployed in production “learn” the same biases inherent in the data that they churned. It is the therefore increasingly important to develop quantitative metrics to measure bias. In this study we propose a simple metric, the Gender Pronoun Gap, that measures the ratio of the occurrences of the pronoun “he” versus the pronoun “she.” We use this metric to investigate the distribution of the Gender Pronoun Gap in two Wikipedia corpora prepared by Machine Learning companies for developing and benchmarking algorithms. Our results suggests that the way these datasets have been produced introduce different types of gender biases that can potentially distort the learning process for Machine Learning algorithms. We stress that while a single metric is not sufficient to completely capture the rich nuances of bias, the Gender Pronoun Gap can be used as one of many metrics.

Keywords: Wikipedia, bias, gender, gender pronoun gap

Introduction

For over 15 years, Wikipedia contributors and editors have produced millions of articles in hundreds of languages. In addition to being a free source of knowledge for millions of citizens around the globe, Wikipedia has also been used as a data source for developing Artificial Intelligence and Ma chine Learning algorithms. For example, Denoyer & Gallinari (2007) is as an example of an early Wikipedia corpus used in the Information Retrieval research community. Wikipedia is an ideal data source for researchers and developers since it is large, covers a vast array of topics, and is free. Large data sources such as Wikipedia have enabled significant progress and advances in algorithm development.

However, as critics of Big Data driven algorithms note, Artificial Intelligence and Machine Learning algorithms are not immune to the biases inherent in data. As a result, data driven algorithms are at risk of having undesirable biases that institutions wish to minimize (Calders & Ž liobaite, 2013; Kay, Matuszek, & Munson, 2015; Pasquale, 2015). Determining how to minimize the introduction of biases from data sources into learning is an active research area (Zliobaite, 2015; Zemel, Wu, Swersky, Pitassi, & Dwork, 2013; Romei & Ruggieri, 2014).

Similar to other data sources, Wikipedia also invariably suffers from bias that has undesirable effects on representing knowledge to users in a fair manner (Holloway, Bozicevic, & Börner, 2007; Callahan & Herring, 2011; Bar-Ilan, Keenoy, Levene, & Yaari, 2009). In particular, numerous studies have noted and investigated significant gender bias in Wikipedia (Collier & Bear, 2012; Reagle & Rhue, 2011; Wagner, Garcia, Jadidi, & Strohmaier, 2015; Hill & Shaw, 2013). In addition to having undesirable effects for knowledge dissemination, these biases are particularly harmful by creating echo chambers and stereotype reinforcements in algorithms used by billions of people, since Wikipedia is free and used for developing Artificial Intelligence and Machine Learning algorithms. To address gender bias, many activists and organizations have devoted to creating communities and “Edit-a-thon” marathons to correct and mitigate bias (Boboltz, 2015). Such efforts have been greeted with great enthusiasm as they have formed communities to address a systematic problem.

But how do we know if such efforts are effective? And how do we know which sub-community and subtopic of Wikipedia suffers the most from gender bias? For example, while there have been considerable efforts in addressing gender inequality in the arts section of Wikipedia (“Meetup”), what other sections of Wikipedia need urgent oversight? To address such questions, we propose the need for developing quantifiable metrics that measure gender bias in Wikipedia. As done in economics, having metrics on gender bias and inequality allows to compare different groups, occupations and track changes over time (Kilbourne, England, Farkas, Beron, & Weir, 1994; Estevez-Abe, 2005; Becker & Toutkoushian, 2003). While a single metric will not measure the full degree of bias variations, having a set of metrics will allow us to gain a better understanding of the full picture and focus our efforts.

In this work we propose a simple metric that we refer to as the “Gender Pronoun Gap” for English Wikipedia. This metric computes the ratio of the number of “he” pronouns against the number of “she” pronouns for each article. If an article contains more “he” (or “she”) words, then we say that the article is biased towards the “he” (or “she”) pronoun. If an article contains roughly the same amount of “he” words and “she” words (or none at all) then the metric assigns no gender bias.

Armed with this metric, we investigate two Wikipedia corpora prepared for Artificial Intelligence and Machine Learning research: “The WikiText Long Term Dependency Lan- guage Modeling Dataset” by MetaMind (Merity, Xiong, Bradbury, &

Socher, 2016) and “The Unknown Perils of Mining Wikipedia” by Lateral (Wilson, 2017). Both corpora are for the English Wikipedia. The corpus provided by MetaMind is over twenty-five thousand articles that have been verified as “Good” or “Featured” articles on Wikipedia. This is an indication of additional scrutiny by editors and gives an article high exposure since many of these articles can be featured on the home page. The corpus provided by Lateral, on the other hand, consists of all English articles that get at least 20 page views per day (only about 12% of Wikipedia articles are viewed more than 20 times per day).

Using our proposed Gender Pronoun Gap metric and basic techniques from Natural Language Processing (NLP), we investigate these two corpora. While we would hope that the Gender Pronoun Gap would be mostly neutral in Wikipedia, we find that there is a heavy bias towards “he” words. In addition, we see that the bias increases (that is worsens) as we change our corpus from general popular Wikipedia articles (captured by Lateral) to the “Good” and ”Featured” articles captured by MetaMind. This further suggests that the editor process of an article to be “Good” or “Featured” introduces additional bias. Using topic modeling, we also identify topics where our proposed metric suggests strong discrepancies in gender representation. Specifically, articles about athletics are heavily biased towards “he” words, while articles about battleships are heavily biased towards “she” words.

Methods

We begin this section by giving an overview of the corpora we used in our study. We then show our proposed metric for computing the Gender Pronoun Gap, followed by the bag-of- words features that we use for performing Latent Semantic Indexing.

Corpora overview

We use two Wikipedia corpora that have been prepared by two different Machine Learning companies. According to both companies, the primary motivation behind preparing and releasing such datasets is for the benefit of researchers to have a consistent dataset to benchmark and systemically evaluate various algorithms. As recent progress in Artificial Intelligence has shown, having high quality and well-established datasets is crucial for advancement of the field. The number of articles for the two corpora are summarized in Table 1. “The WikiText Long Term Dependency Language Modeling Dataset” by MetaMind corporation (Merity, Xiong, Bradbury, & Socher, 2016) consists of over twenty-five thousand articles that have been verified as “Good” or “Featured” articles on Wikipedia. These articles have been elevated to this status by additional editorial scrutiny and are featured on the homepage of English Wikipedia. As a result of this exposure, these articles get high volumes of traffic and readership. We refer to this corpus as the “Featured Wikipedia” corpus.

We also consider a corpus prepared by the Lateral corporation that tries to capture the most “useful” or topic relevant English Wikipedia articles (Wilson, 2017). Many English Wikipedia articles are not written by humans and instead by bots that are essentially “stubs” about cities, rivers and so on. A quick heuristic for getting topically relevant articles is to consider articles that get consistent page views. The corpus provided by Lateral consists of all English articles that get at least 20 or more page views per day.

Table 1

Corpus	**Number of Articles**
All Wikipedia	463,820
Featured Wikipedia	25,951

Number of articles studied for each corpus. “All Wikipedia” refers to English Wikipedia articles with 20 or more daily page views. “Featured Wikipedia” refers to English Wikipedia articles that were either “Featured” or deemed “Good” by the Wikipedia community.

Gender Pronoun Gap metric and text features

Using these two corpora, we compute the gender pronoun gap ratio for each issue as the he-to-she ratio $r_{i,c}$ per article *I* from text corpus c ∈ {MetaMind, Lateral}:

$$r_{i,c} = \frac{1 + \text{"he" count}}{1 + \text{"she" count}}$$

where the “he” and “she” counts are found for each article *i* in corpus *c*. To find the occurrences of the gender pronouns in the Wikipedia corpora, we changed all words in the articles to lower case and search for the strings “he” and “she”. In our study, we do not consider other forms of pronouns such as possessives and contractions. That is, words such as “his” and “he’s” are not considered in our pronoun counts. We used this metric with great success on a student newspapers corpus in an earlier study (Yazdani

& Glass, n.d.). When $r_{i,c}$ = 1.0, the number of "he" pronouns and "she" pronouns is the same (both can also be zero).

When $r_{i,c}$ > 1.0, then there is bias to wards having more "he" pronouns and we refer to this as a "He-heavy" article. Similarly, when $r_{i,c}$ < 1.0, then there is bias towards having more "she" pronouns and we refer to this as a "She-heavy" article. As a concrete example, the sentence "There are offices which she, and only she, can perform" will have a count of 2 "she" words and a count of 0 of "he" words. This example sentence would then yield the Gender Pronoun Gap ratio of $r_{i,c}$ = 1/3.

In the bag-of-words approach to Natural Language Processing (NLP) in each document we count the number of occurrences of the terms in our dictionary (fixed-size vocabulary), creating a count vector that we can use as a feature vector for machine learning tasks. The idea here is that high occurrences of specific terms reflect the content or subject of the document. These raw counts of terms are referred to as "term frequencies." Since the raw counts of the terms may inadvertently weight so-called "stop words" (such as "the") that do not reflect the subject of a document, we normalize the raw counts of each term to diminish bias from commonly occurring words. A common normalization is to count the number of documents that contain each of the terms in our dictionary, referred to as "document frequency." The motivation with this normalization is that if a word is common then it should appear in most of the documents in the corpus that we are studying and will have a high document frequency count. We use the document frequencies to normalize the term frequencies to obtain the TF-IDF features.

In our application, we define each of the terms (also referred to as word or token) that occur in Wikipedia articles as t and each of the articles as d (also referred to as a document in NLP). We compute the TF-IDF features for each article as

$$\mathrm{tf} - \mathrm{idf}(t, d) = \mathrm{tf}(t, d) \cdot (\mathrm{idf}(t, d) + 1)$$

where tf(*t*,*d*) is the number of occurrences (that is, the frequency) of term *t* in article *d* and idf(*t*,*d*) is the normalization of this count with respect to the number of occurrences of *t* in all the other articles. This normalization idf(*t*, *d*) is computed as:

$$\mathrm{idf}(t, d) = \log\frac{1 + N}{1 + \mathrm{df}(d, t)}$$

where *N* is the total number of articles, and df(*d*, *t*) is the number of articles with the term *t*. We also explicitly drop English stop words so their statistics and features are not included. We use the Scikit-learn stop words list taken originally from the Glasgow Information Retrieval Group (Pedregosa et al., 2011; Glasgow, n.d.). We further limit the terms by restrict our vocabulary to terms that have a document frequency of more than 10. That is, if a word appears in only 10 articles or less, we do not include it in our vocabulary.

Computing TF-IDF features allows us to use machine learning models to explore corpora at scale. While there are numerous topic and text modeling methods available, here we use Latent Semantic Indexing (LSI, as computed by the Singular Value Decomposition) to project the TF-IDF features to a lower dimensional subspace. This allows us to explore the distribution of the documents in a 2-dimensional plane and see how the distribution of documents varies as a function of the Gender Pronoun Metric defined by Equation 1. Note that since we removed stop words before computing the TF-IDF features, gender pronouns such as “he” and “she” are removed from the analysis.

Method limitations

The Gender Pronoun Gap metric that we propose has several limitations. For one, the metric only applies to English Wikipedia and potentially to languages with a simple binary gender pronoun structure. Additionally, gender bias is far more complex than a single pronoun so to better understand gender bias we suggest using multiple metrics instead of only one. Finally, since the TF-IDF features that we use are a bag-of-words approach that loses the structure of an article, we are unable to capture higher-level semantic in- formation. More sophisticated and modern topic modelling methods may also yield a richer set of topics, allowing us to better understand the nuances of how the gender bias changes from different topics. However, despite the limitations of the methods, we still hope that we are able to find new hypotheses for further exploration.

Results

Figure 1 shows the percentage of neutral ($r_{i,c}$ = 1.0, labeled as Equal), "he"-biased ($r_{i,c}$ > 1.0, labeled as "He-heavy") and "she"-biased ($r_{i,c}$ < 1.0, labeled as "She-heavy") articles in the "All Wikipedia" and "Featured Wikipedia" corpora. Here "Equal" refers to Wikipedia articles that have an equal number of "he" and "she" pronouns (that is, Equation 1 is equal to 1.0). As discussed in methods, He-heavy articles are article with more "he" pronouns com- pared to "she" pronouns, while She-heavy articles have more "she" pronouns.

The distributions in Figure 1 show that regardless of which corpora we are considering, there is a heavy bias towards the "he" pronoun. More troubling, the "Featured Articles" have higher percentage of bias. This suggests that the selection process in Wikipedia to be a “Good” or “Featured” article introduces even more bias than is already inherent in the data. As a result, both corpuses are at risk of introducing bias into developing Artificial Intelligence and Machine Learning algorithms.

It may seem counter-intuitive that the percentage of “Equal” articles has reduced significantly in the Featured articles compared to All articles. As we defined the gender

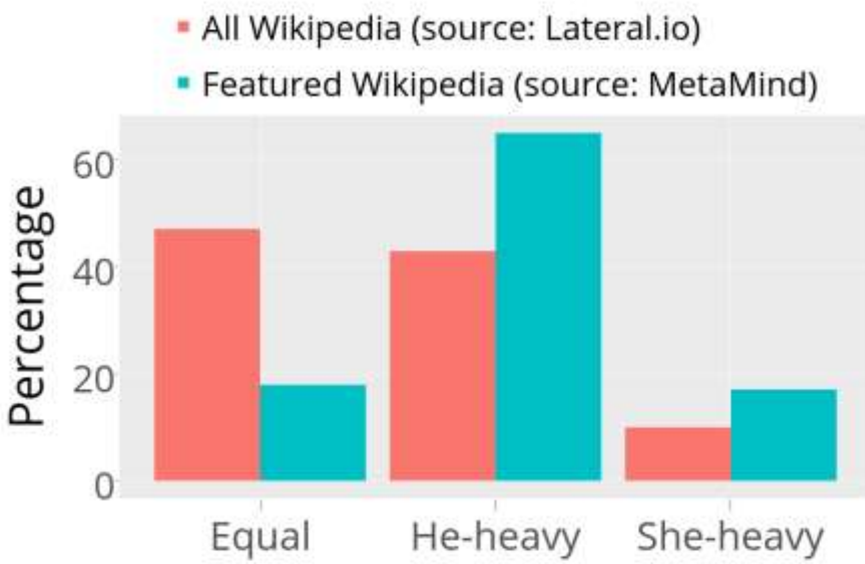

Figure 1. Percentage of neutral (Equal) “he”-biased (labelled as “He-heavy”) and “she”-biased (labelled as “She-heavy” articles in the two different corpora. The bias increases when articles are “Featured” or deemed “Good” by the Wikipedia selection process.

pronoun gap $r_{i,c}$ in Equation 1, articles that have either an equal number of “he” and “she” words (including zero for both) will have an $r_{i,c}$ = 1.0. One explanation for the dramatic de- crease in the percentage of “Equal” articles in the Featured articles may be that there are far fewer featured articles that have to do with individuals or groups of people. For example, there are many articles in mathematics and the sciences addressing specific ideas which do not give historical background on the people behind the topics at hand. Consider one such article: the continuous uniform distribution ("Uniform"). This article describes the continuous uniform dis-tribution from probability theory. While the article discusses the various ideas and properties of the uniform distribution, it does not give any historical background or discussion of the people behind its development. As a result, there are no "he" words and "she" words in this article yielding a $r_{i,c} = 1.0$ and therefore we classify this as an "Equal" or "Neutral" article. Incidentally the article on the uniform distribution has never been featured. We therefore suspect that a great deal of articles have topics that do not mention any individuals and therefore use no pronouns. These articles may be more technical and not approachable for a wider audience and therefore are less likely to reach featured status in Wikipedia.

To better understand the topical variety of the articles in the corpora, we perform Latent Semantic Indexing (LSI) on the TF-IDF features that we defined in the methods section.

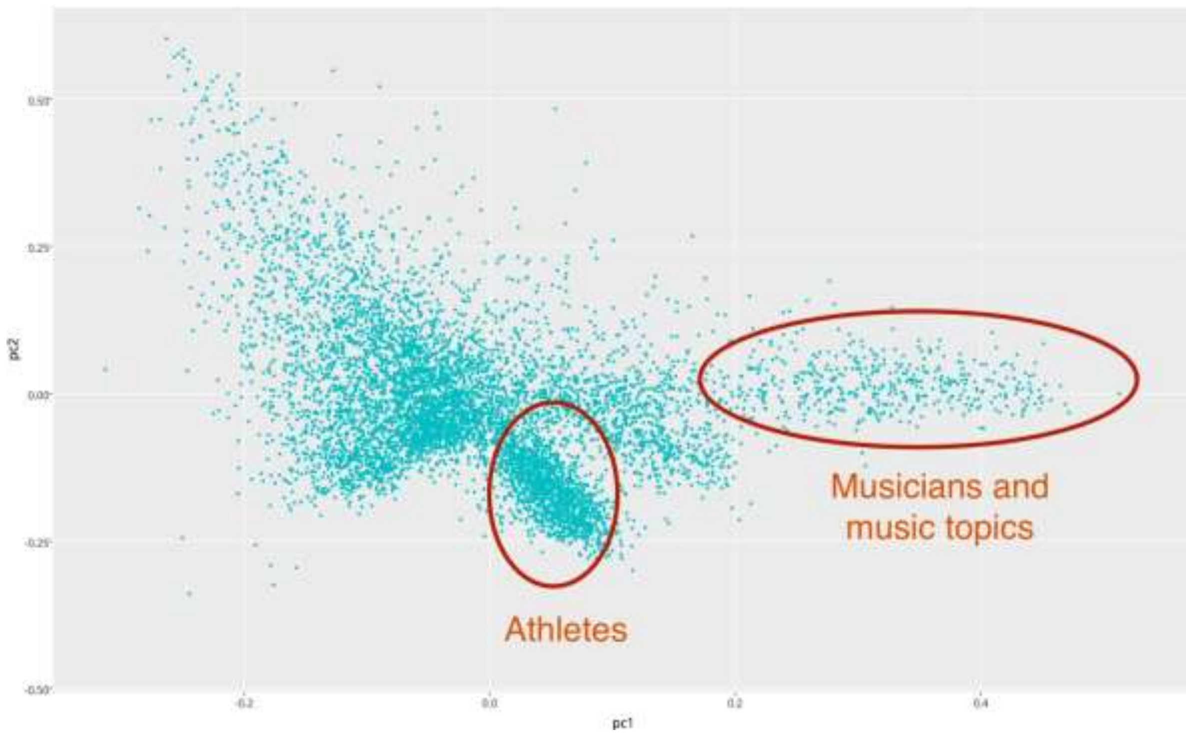

Figure 2. The top 2 Principal Components of LSI on TFI-DF features on the "Featured Wikipedia" corpus prepared by MetaMind. In this Figure we are only showing articles that are "He-heavy" ("She-heavy" follows below). We note that LSI correctly places several articles that are in related topics in a semantically meaningful place. For example, we have highlighted a cluster of articles that belong to music related topics and another cluster for topics related to athletes. For the full and interactive figure, see: https://plot.ly/~crude2refined/2034.embed

Figure 2 shows the top 2 Principal Components for just the He-heavy articles and Figure 3 shows the same for the She- heavy articles.

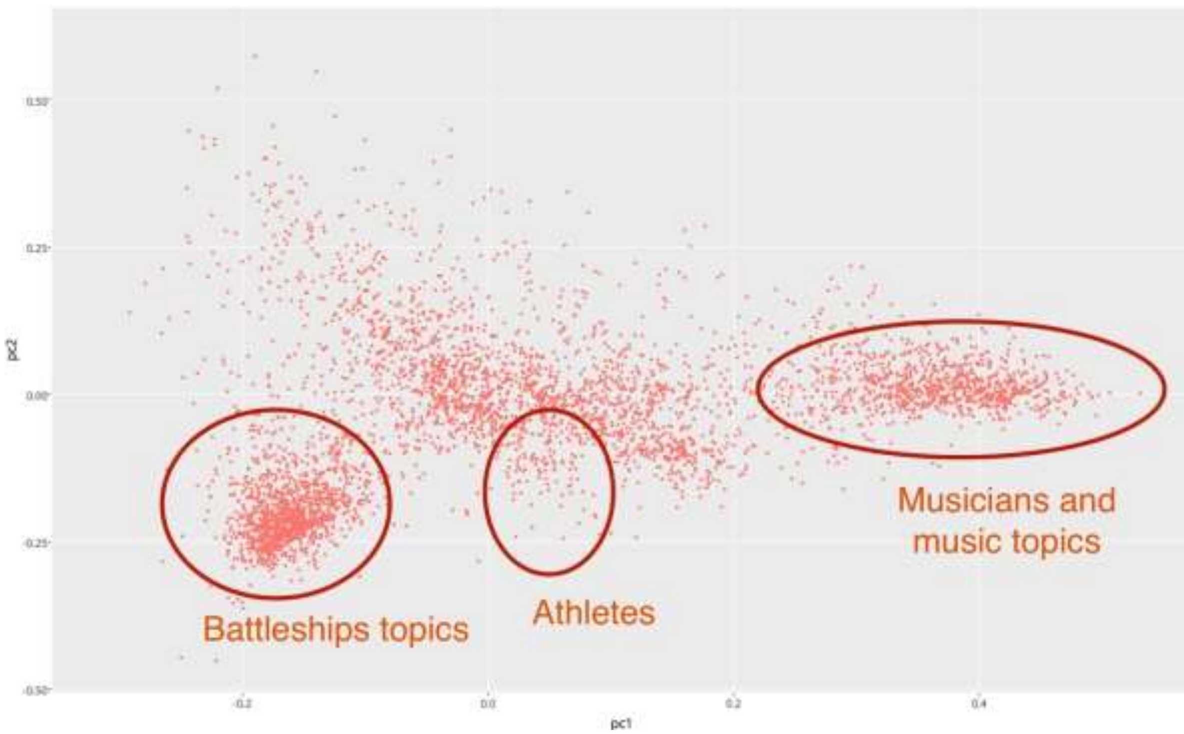

Figure 3. The top 2 Principal Components of LSI on TFI-DF features on the "Featured Wikipedia" corpus prepared by MetaMind. In this Figure we are only showing articles that are "She-heavy." Compare with the previous Figure where we show articles that are "He-heavy." Again, we see that LSI organizes articles in a semantically meaningful way. However, we note that the number of articles in the Athletes cluster is significantly reduced. We also note the introduction of a new cluster that is not present in the He-heavy figure. This cluster corresponds to topics related to battleships. It is interesting to note that such topics are assigned as "She-heavy" articles when in fact they are not about people but rather battleships. For the full and interactive figure, see:https://plot.ly/~crude2refined/2034.embed

Note that each dot in these scatter plots corresponds to a single article. These projections attempt to allow us to visualize how articles in a corpus are topically organized and distributed based on the TF-IDF feature vectors derived for each. In other words, the aim of such projections is to have an overview what articles are related to each other. When a pair of articles is placed close to each other in such a projection, then the suggestion is that the pair are semantically related. We stress that in this LSI experiment, stop words (including pronouns such as "he" and "she") were removed from the documents.

Investigating Figures 2 and 3 reveals that LSI places semantically similar articles close to each other and those that are unrelated are far apart (see the interactive figure to better explore and see this: https://plot.ly/~crude2refined/2034.embed). Here we have highlighted several interesting clusters. In the far right, articles that relate to musicians and musical topics (such as songs, albums, bands, etc) are placed in the far right. We see that both in He-heavy and She-heavy (corresponding to Figures 2 and 3 respectively) articles that have musical topics are placed in the same semantic space. This suggests that semantically Wikipedia articles that describe musicians and musical topics are very similar. In other words, while there may be a discrepancy in the number of articles that are He-heavy versus She-heavy, when it comes to musicians and musical topics, Wikipedia articles are very similar. These results suggest that Wikipedia contributors and editors do not significantly change their language or vocabulary when writing topics related to music and musicians for He-heavy articles or She- heavy articles. For example, the LSI analysis suggests that Wikipedia articles describing a female musician (an article that is She-heavy) uses similar language for articles that de- scribe male musicians (He-heavy articles).

In the bottom center we see a cluster that corresponds to articles about athletes. We note that the biggest difference between the He-heavy and She-heavy articles (corresponding to Figures 2 and 3 respectively) for articles on athletes is that there are far more articles with the he pronoun bias. This suggests that there are far fewer articles about female athletes than male athletes. We therefore suggest that better efforts need to be placed to improve the representation of female athletes as part of “Good” or “Featured” Wikipedia articles.

Perhaps the most disturbing revelation of the She-heavy articles in Figure 3 is the presence of a cluster on the left that is not present in He-heavy articles in Figure 2. This cluster corresponds to articles related to various battleships from old and new eras. At first blush this may appear to be an error by LSI, but note that LSI is not using the Gender Pronoun Gap as a feature. In fact, LSI is correctly placing articles related to battleship close to each other since they are discussing the same topic. It is common that that in naval vernacular battleships are often referred to as “she” instead of the typical “it” pronoun reserved for objects. We therefore find a cluster that should not have significant she-bias, but because of the culture of the battleship article sub-community this set of outliers arises. In other words, by using LSI to explore the different topics in the Featured articles corpus, we were able to identify a topic (battleships) that had nothing to do with peoples or individuals, but has a significant proportion of She-heavy articles. We let the reader decide whether the vernacular of such articles is desirable or not.

Conclusions

In this work we propose using the Gender Pronoun Gap as a possible metric to measure gender bias in Wikipedia. We are in particular concerned with such biases in corpora used in the Artificial Intelligence and Machine Learning community for developing algorithms. If such corpora have bias, then the algorithms that are learning from the data will inevitably learn undesirable biases as well. We investigated two particular corpora prepared by two companies.

Our analysis reveals that “Featured” and “Good” articles that were selected by Wikipedia editors appear to increase the Gender Pronoun Gap (thus potentially increasing gender bias). We further showed that “Featured” and “Good” articles in Wikipedia also have a cluster of articles that are related to battleships that have an unusually high bias towards She-heavy (that is, bias towards the use of the she pronoun).

Acknowledgements

We acknowledge support from Calit2 and its Qualcomm Institute. We also thank the Institute for Pure and Applied Mathematics (IPAM), which is supported by the National Science Foundation for support.

About the Author

Mehrdad Yazdani

Mehrdad Yazdani is a research scientist at the Qualcomm Institute in the University of California, San Diego and a Senior Data Scientist at ID Analytics, a Symantec Company. His primary research interests are in developing machine learning algorithms in a variety of applications including biological and microbiome analysis, satellite image analysis, and anomaly/fraud detection.

References

Bar-Ilan, J., Keenoy, K., Levene, M., & Yaari, E. (2009). Presentation bias is significant in determining user preference for search results -- a user study. *Journal of the Association for Information Science and Technology*, *60*(1), 135–149.

Becker, W. E., & Toutkoushian, R. K. (2003). Measuring gender bias in the salaries of tenured faculty members. *New Directions for Institutional Research*, *2003*(117), 5–20.

Boboltz,S. (2015, April 15). Editors are trying to fix Wikipedia's gender and racial bias problem. *Huffington Post*. Retrieved from http://www.huffingtonpost.com/2015/04/15/wikipedia-gender-racial-bias_n_7054550.html.

Calders,T.,& ž liobaitė , I. (2013). Why unbiased computational processes can lead to discriminative decision procedures. In Custers, B. (Ed.) *Discrimination and privacy in the information society: Data mining and profiling in large databases.* (pp. 43–57). Berlin, Germany: Springer.

Callahan, E. S., & Herring, S. C. (2011). Cultural bias in wikipedia content on famous persons. *Journal of the American society for information science and technology*, *62*(10), 1899–1915.

Collier, B., & Bear, J. (2012). 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.* (pp. 383–392). doi: https://doi.org/10.1145/2145204.2145265

Uniform distribution (continuous) (n.d.).In *Wikipedia*. Retrieved from http://en.wikipedia.org/w/index.php?title=Uniform\%20distribution\%20(contin uous)&oldid=776358129.

Denoyer L., Gallinari P. (2007) The Wikipedia XML Corpus. In: Fuhr N., Lalmas M., Trotman A. (Eds) *Comparative Evaluation of XML Information Retrieval Systems. INEX 2006. Lecture Notes in Computer Science*, vol 4518. (pp. 12–19). Berlin, Germany: Springer. https://doi.org/10.1007/978-3-540-73888-6_2.

Estevez-Abe, M. (2005). Gender bias in skills and social policies: The varieties of capitalism perspective on sex segregation. *Social Politics: International Studies in Gender, State & Society, 12*(2), 180–215.

Glasgow Information Retrieval Group, G. I. R. (n.d.). Retrieved from https://www.gla.ac.uk/schools/computing/research/researchoverview/informationretrieval/

Hill, B. M., & Shaw, A. (2013). The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation. *PLoS ONE, 8*(6), e65782. http://doi.org/10.1371/journal.pone.0065782.

Holloway, T., Bozicevic, M., & Börner, K. (2007). Analyzing and visualizing the semantic coverage of wikipedia and its authors. *Complexity, 12*(3), 30–40.

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems.* (pp. 3819–3828). doi: 10.1145/2702123.2702520

Kilbourne, B. S., England, P., Farkas, G., Beron, K., & Weir, D. (1994). Returns to skill, compensating differentials, and gender bias: Effects of occupational characteristics on the wages of white women and men. *American Journal of Sociology, 100*(3), 689–719.

Meetup/artandfeminism. (n.d.). Meetup/artandfeminism. In *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:Meetup/ArtAndFeminism.

Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). Pointer sentinel mixture models. arXiv:1609.07843.v1 [cs.CL] 26 Sep 2016.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Cambridge, MA: Harvard University Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. https://github.com/scikit-learn/scikit-learn/blob/4d9a12d175a38f2bcb720389ad2213f71a3d7697/sklearn/feature_extraction/stop_words.py. GitHub.

Reagle, J., & Rhue, L. (2011). Gender bias in Wikipedia and Britannica*. International Journal of Communication, 5* (21), 1138-1158.

Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review, 29*(05), 582–638. https://doi.org/10.1017/S0269888913000039

Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. arXiv:1501.06307v2 [cs.CY].

Wilson, B. (2015, June 2). The unknown perils of mining Wikipedia. Retrieved from https://blog.lateral.io/2015/06/the-unknown-perils-of-mining-wikipedia/

Yazdani, M., & Glass, E. (n.d.). Using exploratory data analysis to investigate the gender pronoun gap in student newspapers.

Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013).

Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, *(3),* 28. (pp. 325–333).

Zliobaite, I. (2015). A survey on measuring indirect discrimination in machine learning. arXiv Advance online publication. arXiv:1511.00148