Bias and Context: Using Word Embeddings to Compare Word Use Across Contexts

Eric Schnell

Ariyo Vahdat

Abstract

Word embeddings are a powerful tool which can pick up on the relationship between words. In the past, word embeddings have been used to track bias. We build on this previous work by also contextualizing the texts used to build our word embeddings. It has been shown that context can itself influence the language people use and whether or not they are biased in their word choice. We have categorized our word embeddings as representing casual or formal contexts for speech, where our casual word embeddings are built from posts to social media platforms and our formal word embeddings are built from news articles. We then test the levels of bias in each of word embeddings to see if there is in fact a difference across contexts. This work shows the interconnectedness of all human processes. Issues such as bias do not exist alone in individuals but rather across cultural as a whole. By changing the cultural context in which an individual is speaking, the bias that individual uses is also changed.

Keywords: Bias; Cultural Context; Gender Stereotypes; Ethnic Stereotypes; Word Embedding

Introduction

Stereotypes, prejudice, and bias are a very real and often very prevalent part of many of our lives. The now dated thought that sports are for boys and arts are for girls was a very real part of the culture many of us grew up with. Recent cultural movements have made efforts to eliminate these biases from our culture, especially from being casually used in conversation. In addition to bias being ever present, it is also not equal in all contexts, and the context that people are speaking in has an affect on how biased they are (van Dijk, 2008). Our research was interested in exploring this. The purpose of this experiment was to compare bias in language use in between different contexts. Specifically, we were interested in how bias would differ between formal and casual contexts. We collected data from the social media website Reddit to serve as our casual context, and data from CNN and the Daily Mail to serve as our formal contexts. We were interested in seeing if in a casual context such as Reddit, where users are mostly anonymous and for the most part cannot be accounted for their comments, people would be more likely to allow bias to slip into their writings. We also believed that sources like CNN and the Daily Mail, which are expected to be formal and have a public image to uphold, would be more strict in their speech and word choice, allowing less bias to find it's way into their writings.Thus, the hypothesis for this study was that more bias would be observed in the casual contexts.

The method we selected to forego this experiment with, and the primary source of inspiration for pursuing this study, was word embeddings. Word embedding models take in a collection of text and assign a high-dimensional numeric vector to each word in the text source. These vectors act as coordinates for each word in the vector space. Vectors are assigned to words based on their context words, thus the position of these vectors often can often give insight into the meaning of the words they are assigned to. Words of similar meaning are often located near each other in vector space, as these words often share similar context words. For this reason, word embeddings have been shown to be highly effective in studies of analyzing bias in language use (Garg, Schiebinger, Jurafsky, & Zou, 2018; Kozlowski, Taddy, & Evans, 2019).

Previous studies in bias analysis using word embeddings have also developed methods for quantifying bias in word embedding models. It has been shown that different kinds of bias can be captured and represented by a direction in the word embedding model (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016). There is also the Word Embedding Association Test (WEAT), which measures whether or not two sets of target words are equally similar to a set of attribute words. For our research, WEAT was chosen as the method for quantifying and comparing bias; the procedure returns a p-value as evidence for whether or not we have enough evidence to conclude that there is any bias present in the model for the given trial, as well an effect size for directly comparing a numeric quantity of the bias between contexts.

Previous studies have looked into comparing how bias in language used changed overtime using word embeddings. Such studies made use of word embeddings and geometric bias quantification methods and looked at how gender and ethnic bias changed over the course of the 20th and 21st century (Garg et al., 2018) and also how class bias changed overtime (Kozlowski et al., 2019). Though there are various kinds of bias that can be measured using word embeddings, we decided to limit the scope of our study to focus on comparing gender and racial bias.

Methods[1]

Word Embeddings

The study utilized 250-dimensional word embeddings, trained using the word2vec algorithm operationalized by the gensim package for Python 3. They were trained using gensim's default parameters. Notably the window size for training was 5, each word needed to be more than one letter long, and each word needed to appear at least 5 times to be included in the model. We trained six word embedding models using datasets of various contexts and discussing various topics.

Analysis

The method of analysis used to quantify and compare bias between the different models is the Word Embeddings Association Test, or WEAT. WEAT is a permutation test that takes as inputs two sets of target words (for example, one set comprised of words related to "science" and the other set of words related to "arts") and two sets of attribute words (one set being male pronouns and the other being female pronouns). The null hypothesis for the test is that both sets of target words are equally similar from both sets of attribute words. Similarity of words is calculated by finding the cosine similarity of their word embeddings in the given model. First, a test statistic is calculated. Let A, B be two sets of attribute words of equal size, and let X, Y be two sets of target words of equal size. Let $\cos(a,b)$ denote the cosine similarity between two vectors a and b. The test statistic is calculated as the following:

$$s(X,Y,A,B) = \sum_{x \in X} s(x,A,B) - \sum_{y \in Y} s(y,A,B)$$

where

$$s(w,A,B) = \text{mean}_{a \in A} \cos(w,a) - \text{mean}_{a \in A} \cos(w,a)$$

$s(w,A,B)$ calculates which set of attribute words a given target word w is more similar to, then $s(X,Y,A,B)$ repeats this for each word in each target set, sums the result for each target set, and returns the difference.

The p-value is calculated using the following:

$$P[s(X_i,Y_i,A,B) > s(X,Y,A,B)]$$

where $(X_i,Y_i)_i$ are all possible equal sized sets of words selected from $X \cup Y$.

Finally, the effect size is calculated as follows:

$$\frac{\text{mean}_{x \in X} s(x,A,B) - \text{mean}_{y \in Y} s(y,A,B)}{\text{std-dev}_{w \in X \cup Y} s(w,A,B)}$$

(Caliskan, Bryson, & Narayanan, 2017)

The implementation of the WEAT functions was provided by Hod (2018).

[1]Materials can be found at: https://github.com/rio-v/Bias-and-Context-COG403

Trials

There were six sets of trials performed on each of the word embedding models. Three tests for measuring gender bias and three for measuring racial bias. The terms used to test gender bias come from the Caliskan et al. (2017) study where they introduced WEAT. Lists of words were shortened to account for computational demands of running the WEAT. Certain words were also removed to account for words that weren't in any given model. The word lists for comparing gender bias were the following:

Career vs. Family
Career terms: executive, management, professional, corporation, salary, office, business, career
Family terms: home, parents, children, family, cousins, marriage, wedding, relatives
Male Names: John, Paul, Mike, Kevin, Steve, Greg, Jeff, Bill
Female Names: Amy, Joan, Lisa, Sarah, Diana, Kate, Ann, Donna

Math vs. Arts
Math Terms: math, algebra, geometry, calculus, equations, fraction, numbers, addition
Art Terms: poetry, art, dance, literature, novel, symphony, drama, sculpture
Male Pronouns: male, man, boy, brother, he, him, his, son
Female Pronouns: female, woman, girl, sister, she, her, hers, daughter

Science vs. Arts
Science Terms: science, technology, physics, chemistry, einstein, NASA, experiment, astronomy
Art Terms: poetry, art, Shakespeare, dance, literature, novel, symphony, drama
Male Pronouns: brother, father, uncle, grandfather, son, he, his, him
Female Pronouns: sister, mother, aunt, grandmother, daughter, she, hers, her

The terms used to test racial bias come from the Garg et. at.'s research. The names for each ethnic group are last names found within the United States, and the occupations are ones that were found to be most associated with each ethnicity according to their research. As with the gender bias trials, lists of words were shortened to account for computational demands of running the WEAT. Certain words were also removed to account for words that were not in any given model. The word lists for comparing gender bias were the following:

White Last names: Harris, Nelson, Robinson, Thompson, Moore, Wright, Anderson, Clark, Jackson
Asian Last names: Cho, Wong, Tang, Hong, Kim, Chen, Ng, Wu, Liu
Hispanic Last names: Cruz, Castro, Garcia, Torres, Martinez, Gonzalez, Sanchez, Lopez, Rodriguez

Rated White Occupations: smith, blacksmith, surveyor, sheriff, weaver, administrator, statistician, clergy
Rated Asian Occupations: physicist, scientist, chemist, accountant, professor, official, secretary, conductor
Rated Hispanic Occupations: housekeeper, artist, janitor, dancer, mechanic, photographer, baker, cashier

Data

In this study we used six datasets, four of which came from Reddit, one from CNN, and one from the Daily Mail. These datasets represent different contexts and topics which vary in terms of their formality. For a casual context we used Reddit, a social media site where anyone can post topics of discussion and have people comment their replies. The Reddit datasets we used are made up of comments rather than posts, as comments are more discussion oriented, whereas posts are often short descriptions or stories. The news datasets, which represented a formal context, are CNN and the Daily Mail. To ensure that each model is representative of the same amount of information, we made each corpus from roughly 35 million words. This number was selected to stay within the file size limit on github. We also ensured each Reddit comment was at least ten words long so it could be representative of actual discussion and the first line of each news article was omitted for data cleanup purposes.

We have also separated our datasets based on formality of topic being discussed. Two of our Reddit datasets pulled comments from the entire site, and thus mostly comprised of casual topics. Two of our Reddit datasets only used comments on posts concerning news and politics, thus being discussion on more formal or serious topics. Specifically, these formal topics were from comments on posts found in the communities r/politics, which focuses on American politics, r/news, which focuses on American news, and r/worldnews, which focuses on news from around the world. For the news datasets, the CNN was considered to be discussing formal topics and the Daily Mail was considered to be discussing casual topics. The reason for this is that CNN typically focuses more on traditional news and politics, whereas the Daily Mail is a tabloid and focuses a lot on celebrity news and gossip. It must be noted that in all these cases the lines between casual and formal are not so clear as presented here. For instance, CNN will also discuss celebrities and the Daily Mail will also discuss politics, but the more common topics will be the ones with a larger influence on the word embeddings and so for the sake of the paper these categories will be used.

The datasets used for the news sites comes from Hermann et al. (2015) for their paper "Teaching Machines to Read and Comprehend". They gathered CNN articles from 2007 until 2015 and Daily Mail articles from 2010 until 2015. The researchers posted their program for the set of articles. These datasets were then hosted online by Kyunghyun Cho at New York University,[2] which is the dataset used in this study. For the Reddit data, we used a dataset made and hosted by pushshift.io.[3] This group gathers all Reddit posts and comments, with metadata as provided by Reddit, for a given month and then stores them on their website. The months used in this study are from September 2019 and January 2014. The reason for having two months is that we wanted the most recent data available (2019) and a dataset from the same time frame as the news datasets (2014). We do not expect there to be a significant difference between the separate Reddit datasets.

Table 1 lays out the relevant statistics for each dataset, where "All xxxx" represents the dataset of all Reddit comments from the relevant year and "News xxxx" represents the dataset of Reddit comments pertaining to news and politics of the relevant year.

Table 1: Dataset Statistics.

Dataset	# of Entries	# of Total Words	# of Unique Words
All 2019	820 587	35 000 247	81 008
All 2014	845 925	35 000 004	72 218
News 2019	760 393	35 000 039	49 897
News 2014	664 873	34 422 118	52 258
CNN	54 403	35 000 478	75 810
Daily Mail	58 031	35 000 381	79 159

Results

The following tables contain the results from each set of trials. Tables 2, 3, and 4 contain the results from the gender bias trials, and tables 5, 6, and 7 contain the results from our racial bias trials. The significance level for our p-values were 0.05 and are signified with a *.

Gender Bias Results

For the Career vs. Family trial set, all six models were found to be biased, with the All Reddit 2019 and Daily Mail models having the highest effect sizes (Table 2).

Table 2: Career vs. Family

Dataset	Test Score	Effect Size	P-value
All Reddit 2014	1.241	1.820	<0.0001*
All Reddit 2019	1.044	1.902	<0.0001*
News Reddit 2014	0.828	1.617	0.0002*
News Reddit 2019	0.541	1.116	0.013*
CNN	1.962	1.849	<0.0001*
Daily Mail	1.826	1.897	<0.0001*

For the Math vs. Arts trial set, both the All Reddit 2019 and the Daily Mail models were found to be biased (Table 3).

[2]https://cs.nyu.edu/ kcho/DMQA/

[3]https://files.pushshift.io/reddit/comments/

Table 3: Math vs. Arts

Dataset	Test Score	Effect Size	P-value
All Reddit 2014	0.081	0.292	0.297
All Reddit 2019	0.332	1.051	0.019*
News Reddit 2014	-0.173	-0.845	0.950
News Reddit 2019	-0.012	-0.057	0.538
CNN	0.065	0.383	0.235
Daily Mail	0.161	1.035	0.019*

For the Science vs. Arts trial set, none of the models showed significant evidence of being biased (Table 4).

Table 4: Science vs. Arts

Dataset	Test Score	Effect Size	P-value
All Reddit 2014	0.222	0.691	0.094
All Reddit 2019	0.286	0.745	0.078
News Reddit 2014	-0.042	-0.155	0.612
News Reddit 2019	-0.251	-0.833	0.945
CNN	0.1477	0.820	0.056
Daily Mail	0.134	0.654	0.106

Race Bias Results

For the White vs. Asian trial set, both the CNN and Daily Mail models were found to be biased (Table 5).

Table 5: White vs. Asian

Dataset	Test Score	Effect Size	P-value
All Reddit 2014	-0.304	-0.310	0.716
All Reddit 2019	-0.162	-0.172	0.620
News Reddit 2014	0.206	0.328	0.270
News Reddit 2019	0.268	0.401	0.224
CNN	0.837	1.130	0.010*
Daily Mail	0.708	0.969	0.021*

For the White vs. Hispanic trial set, none of the models showed significant evidence of being biased (Table 6). The CNN and Daily Mail did have the lowest P-values in the trial, as well as the highest effect sizes.

Table 6: White vs. Hispanic

Dataset	Test Score	Effect Size	P-value
All Reddit 2014	-0.371	-1.035	0.979
All Reddit 2019	-0.066	-0.149	0.597
News Reddit 2014	-0.346	-0.693	0.907
News Reddit 2019	0.019	0.042	0.471
CNN	0.303	0.757	0.073
Daily Mail	0.579	0.77291626	0.070

For the Hispanic vs. Asian trial set, the CNN model was found to be biased (Table 7).

Table 7: Hispanic vs. Asian

Dataset	Test Score	Effect Size	P-value
All Reddit 2014	-0.128	-0.264	0.686
All Reddit 2019	-0.011	-0.028	0.520
News Reddit 2014	0.095	0.191	0.361
News Reddit 2019	-0.030	-0.109	0.581
CNN	0.580	0.910	0.038*
Daily Mail	0.064	0.162	0.380

The results show the existence of a bias for certain tests. Notably the career versus family gender bias test, where all six models had a significant bias. However, no other test found such overwhelming results. The math versus arts test also uncovered a gender bias in the model for all Reddit comments 2019 and the Daily Mail model. That is one casual context model and one formal context model. As for racial bias, this bias only appears in formal context models. These being both the CNN and Daily Mail models for the White versus Asian test and CNN model for the Hispanic versus Asian test.

Overall, the level of bias in our models, as calculated using the WEAT test, is lower than expected. Seeing as all these tests were shown to unearth bias in the papers they were originally taken from, we expected there to be consistent levels of bias across the board. As a general trend, context does appear to induce bias, however not in the direction we hypothesised. Overall, the formal context had a higher level of bias. This can be seen not only in the fact it was significantly biased more frequently than the casual contexts, but also in comparing the average test scores across contexts as shown in Table 8.

Table 8: Average WEAT Score.

Test	Casual Context	Formal Context
Career vs. Family	0.914	1.894
Math vs. Arts	0.057	0.113
Science vs. Arts	0.054	0.141
White vs. Asian	0.083	0.773
White vs. Hispanic	-0.158	0.441
Hispanic vs. Asian	-0.019	0.322

Discussion

First, as stated above, our results found generally lower levels of bias than expected. Context did appear to play a role, but not in the direction expected, as formal contexts had higher levels of bias than casual contexts. An indicator of this is that there were more instances where the formal context was significantly biased (6) than there were where the casual context

was significantly biased (5), even though there were twice as many casual models. Also, all but one instance of bias in a casual model came from the career versus family gender bias test, which uncovered high levels of bias in each model. Another indicator of higher bias in the formal context is shown in Table 8, where the WEAT score was on average higher in the formal context than the casual one for every single test. Overall it appears that bias is more prevalent in a formal context rather than a casual one, which is the opposite of our predicted hypothesis.

It must also be noted that there is a conflicting effect of formality of topic on bias. In the case of gender bias, where there was significant bias, the casual topic Reddit data had more bias than the formal topic Reddit data. Whereas the formal topic news data had more bias in the career versus family test, but less bias in the math versus arts test. With respect to racial bias the formal news topic had more bias than the casual news topic. So, there are conflicting effects of topic formality on bias and thus this topic formality does not appear to play a role in bias.

To understand why the results go against our hypothesis, we must look at the most significant issue with our word embeddings, that being the size of the corpus with which we trained our word embeddings. According to Lai, Liu, He, and Zhao (2016), when working with corpuses which focus on a specific domain, the larger the corpus the better. Notably they found that corpuses built from over 1 billion words performed much better than ones of 100 million words. In our case we had a significantly smaller corpus of just roughly 35 million words each. Another issue may have been the dimensionality of our corpus. In the relevant literature concerning bias in word embeddings, word embeddings are typically either 200 (Kozlowski et al., 2019) or 300 (Caliskan et al., 2017) dimensional vectors. Our thinking was to split the difference and use 250-dimensional vectors. However, these papers used much larger corpuses to build their embeddings, and thus a higher dimensionality was required. It could be that the dimensionality we used was too large given our corpus size, and that we should have rather made 200-dimensional or even 100-dimensional word embeddings. However, the clearer issue is that of our corpus size.

This issue of corpus size not only suggests that our results are not representative of true bias, but it may also suggest why we found more bias in the formal context. News sites follow major stories over large periods of time and write many articles about the same story. They also cover many similar stories and each story is considerably longer than the average Reddit comment. As such the corpus is much more focused for news sites than for Reddit, which in turn means that the word embedding will more accurately reflect trends in the data and can more easily pick up on any bias in the data. Seeing as Reddit will have a more varied focus, especially when taking comments from all subreddits, it is more difficult for our word embeddings to accurately characterize the corpus. This issue becomes even more severe considering the small corpus sizes used. Comparatively the focused corpus of news sites leads to more accurate word embeddings which make bias more noticeable.

Another possibility is that Reddit is simply less biased than news sites. The lack of barrier to entry on Reddit may make it more diverse and representative of varied viewpoints. It could mean that there is less bias overall as biased individuals are countered by ones who hold opposite viewpoints. On the other hand, articles on the Daily Mail and CNN are all written by journalists. Seeing as these organisations do not release such details about their journalists, it is impossible to know for sure the backgrounds and educations of all the journalists, but it can be safely assumed that they are mostly college-educated. This means they are bringing specific backgrounds, not shared by all, into their writing. As such journalists may reinforce biases innate to the craft that may not be found on a more open platform such as Reddit. In this sense then we are not comparing whether people are more likely to use biased language in a casual rather than a formal context, but rather we are comparing the bias amongst two specific populations. Rather we would want to compare bias across context for the same group of speakers. Reddit may simply be less biased than the news, but this does not answer whether bias is stronger in a casual or formal context.

On top of the issues with corpus size, the corpus selection could also have impacted the results. Seeing as we built word embeddings for each separate corpus, there may have been other factors built into these corpuses that impact bias but have nothing to do with context. The most obvious of these is political alignment. Three of our four types of datasets discuss politics and the news and so the political alignment found in each of these datasets will be significant in how it is discussed. CNN and the popular political subreddits are considered left-leaning, whereas the Daily Mail is considered right-leaning. These political leanings could have impacted the bias of the corpus while circumnavigating the context of the discussion.

The tests used may have swayed our results. It seems clear that our models do have some level of bias which can be found, seeing as for the career versus family test this bias was overwhelming. It may be that the other tests were incorrectly chosen, and that is why we could not find high levels of bias elsewhere. We selected attribute words and target words as found in Caliskan et al. (2017) and Garg et al. (2018), and so we had assumed that this meant that these were reliable tests. Of our tests, only the career versus family gender bias test had results similar to that found in the original paper. For this test, Caliskan et al. (2017) had a distance score of 1.89, which is similar to some of the results we had, notably for the Daily Mail and CNN datasets at 1.96 and 1.82 respectively. In our other gender bias tests that found a significant bias, the scores were quite distant from that in Caliskan et al. (2017). As for racial bias, the paper by Garg et al. (2018) from which we took our attribute words did not use the WEAT test so we cannot directly compare our results to theirs. The ques-

tion now becomes, how come we got such similar results to one test, but for none others? It could be that we overestimated the accuracy of the original tests and instead we should have taken tests from more papers as well as creating our own tests. It could also be an issue of corpus size. The career versus family test had the largest WEAT score by a considerable margin in Caliskan et al. (2017). This bias towards women with regards to career versus family is also well documented in other word embedding literature (Bolukbasi et al., 2016). As such it would appear to be the easiest bias for our model to uncover. So, with models built from small corpuses, this is the test that is most likely to return a significant bias. Seeing as we know at least one WEAT test can successfully uncover bias, it could be that with models built from larger corpuses the other tests would have also uncovered a significant bias. Regardless, the tests that we chose could have impacted the sort of results we found.

Conclusion

We failed to prove our hypothesis that a more casual context would promote more bias in conversation. In fact, we found there to be very little bias in our various corpuses and when bias did appear it was stronger in the formal context than in the casual context. That being said it does show that context impacts bias, just not in the direction we anticipated. Our study used word embeddings, specifically word2vec, and we suspect that our results were negatively impacted by the fact that these embeddings were trained on very small corpuses. We still believe that our hypothesis is plausible, however in the future it would need to be tested on larger corpuses of text. It would also be good to draw our corpuses from more varied sources and to use more test cases. If the hypothesis were to be proven, an interesting next study could be to compare bias in spoken versus written text, again using word embeddings where the word embeddings for spoken texts would be built on their transcription. This would look at whether the fact that written text can be deliberated on longer means that bias would be more likely to be eliminated. However, before moving onto this project it is important to further study the role of context in bias. Although the results found do not align with the hypothesis, we are still hopeful that by using larger corpuses we could eliminate many of the faults of the current study and in fact prove the hypothesis.

References

Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 4349-4357.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186. doi: https://doi.org/10.1126/science.aal4230

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(16), E3635–E3644. doi: https://doi.org/10.1073/pnas.1720347115

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in neural information processing systems*, 1693-1701.

Hod, S. (2018). *Responsibly: Toolkit for auditing and mitigating bias and fairness of machine learning systems.* Retrieved from `https://docs.responsibly.ai/`

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, *84*(5), 905-949. doi: https://doi.org/10.1177/0003122419877135

Lai, S., Liu, K., He, S., & Zhao, J. (2016). How to generate a good word embedding. *IEEE Intelligent Systems*, *31*(6), 5-14. doi: https://doi.org/10.1109/MIS.2016.45

van Dijk, T. A. (2008). *Discourse and context: A sociocognitive approach*. Cambridge University Press.