Scoring Ourselves

Linda Dong
UC Berkeley
lindadong@berkeley.edu

Abstract

Gender biases are unfortunately present in many real-world datasets, as a reflection of our imperfect reality. Researchers have developed methods for recognizing and correcting these biases in machine learning applications; often, biases are silently neutralized. Is this the right approach? What if NLP were used, instead, to highlight and measure the extent of bias, in order to prompt behavioral change? In this project, I generate individual gender bias scores for three individuals (Donald Trump, Hillary Clinton, and myself), leveraging established measurement methodology, fine-tuned BERT language models, and Twitter and Facebook datasets.

1 Introduction

Social biases are unfortunately present in many real-world datasets. Most methods for recognizing and correcting these biases in machine learning applications do so silently - but is this the right approach?

While silent neutralization can resolve the problem within specific applications, it cannot generalize across all circumstances (in settings without machines, for example), and it also does not address the root cause of the issue. Outside of gender biases, there are many cognitive biases that humans are notoriously bad at combating - including confirmation bias, selection bias, recency bias, and loss aversion. What if, instead of silently resolving a subset of the problem, we used machine learning to measure the problem? We can then nudge individuals towards behavioral change, in order to address the root cause and truly resolve the issue at hand.

2 Background

Gender bias in machine learning applications is a relatively well-studied field (Sun et al., 2019). Previous work on measuring gender biases have largely leveraged word embeddings (such as GloVe and word2vec) (Garg et al., 2018) and used vector distances (usually cosine similarities) to measure associations across gendered words (Bolukbasi et al., 2016).

One seminal example of this is the Word Embedding Association Test (WEAT), which applies psychology's established Implicit Association Test (IAT) (Greenwald et al., 1998) to word embeddings. In a paper titled "Semantics derived automatically from language corpora necessarily contain human biases," Caliskan et al. introduced WEAT and showed that many of the implicit associations and attitudes IAT found in humans were also present in GloVe (Caliskan et al., 2017). These researchers used baskets of words to address polysemy (improving upon previous iterations of similar efforts) and tested for both racial and gender biases. They found that female names were more associated with family, while male words were more associated with career; similarly, female terms were more associated with the arts, while male terms were more associated with the sciences.

State-of-the-art NLP methods, however, no longer rely on static embeddings. Language models, such as ELMo and BERT, rely on contextual word embeddings, where each input element can be represented by a different embedding, depending on its surrounding context. This makes it hard to leverage these older methods of measuring bias.

Luckily, measurement methodology have also evolved. Kurita et al., in the 2019 paper "Measuring bias in contextualized word representations," extended BERT's training masking mechanism to formulate an approach to calculate BERT's gen-

der bias (Kurita et al., 2019). Taking inspiration from BERT's mechanism of randomly masking 15 % of words in training, these researchers deliberately masked words post-training, using specially formulated evaluation sentences, to measure whether BERT's predicted words were more likely to be female or male. This methodology required pre-defined evaluation frameworks (detailed in the "Scoring" section below) and generated a gender bias score per evaluation attribute. For this project, I used 6 attributes: career and family attributes (leveraged from WEAT), pleasant and unpleasant attributes (leveraged from Kurita et al), and leader and follower attributes (newly created by me).

3 Summary of Methodology

For this project, I chose to use BERT-base to train my three individual language models (for Donald Trump, Hillary Clinton, and myself). I gravitated towards using BERT not only because it is state-of-the-art, but also because I faced many limitations that using BERT could resolve. First, in order to train individual models, I needed to train on an individual's data. Individual data is limited in both quantity and availability, and often unlabeled. It would not be feasible to train an individual language model from scratch - but I can leverage pre-trained BERT-base, which can be fine-tuned with much less individual data.

For data, I leveraged Twitter archives and Facebook, and decided to build three models - one for Donald Trump (using Tweets), one for Hillary Clinton (using Tweets), and one for myself (using a direct export of my Facebook posts and comments).

After fine-tuning these three models, I then generated gender bias scores for each model across 6 attributes: career and family [1], pleasant and unpleasant [2], and leader and follower [3].

3.1 Novelty

Two things are new about my approach. One is the fact that I generated individual bias scores; the Kurita paper only measured BERT-base's bias. This means I also need to care about model accuracy (i.e., that my model accurately predicts what Trump, or Hillary, or myself would say), in addition to measuring bias.

I also added 2 new attributes for evaluating bias. While I leveraged WEAT's family and career attributes, I modified Kurita's pleasant and unpleasant attributes (by adding more vocabulary to the attribute lists), and I created my own leader and follower attributes (more detail in "Scoring" section below).

3.2 Baseline

Since I'm comparing gender bias scores across models, I defined my baseline to be BERT-base (i.e., I'm comparing these biases to gender bias scores generated by the BERT-base model).

4 Model Training

Using Google's BERT repository [4], I first attempted to fine-tune my Trump model. From a total of 47,055 Tweets downloaded, I filtered out retweets, replaced "\n" with a space in order to format every Tweet on its own line, shuffled each Tweet, and then split the data 90/10 into training and validation sets. This resulted in 26,531 total Tweets, with 23,878 in my training set and 2,653 in my test set.

I tried many different approaches, including filtering out hyperlinks, increasing steps (from 10K to 40K), using the momentum optimizer, changing learning rates, freezing weights (only training the dense layer), and removing the next_sentence_loss term.

Validation accuracy flattened out prior to 10K steps and showed no improvement at 40K:

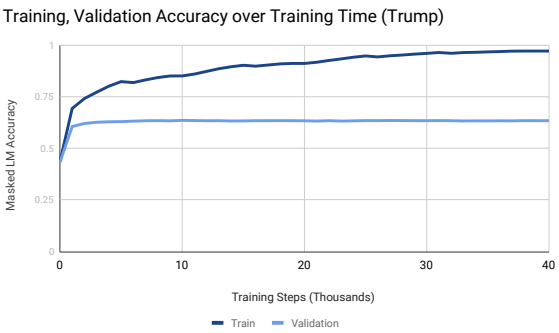

Figure 1: **Accuracy Curve**

[1] Details here: https://github.com/lindadongx/bert/tree/master/data/sets/career_family

[2] Details here: https://github.com/lindadongx/bert/tree/master/data/sets/unpleasant_pleasant

[3] Details here: https://github.com/lindadongx/bert/tree/master/data/sets/follower_leader

[4] Details here: https://github.com/google-research/bert

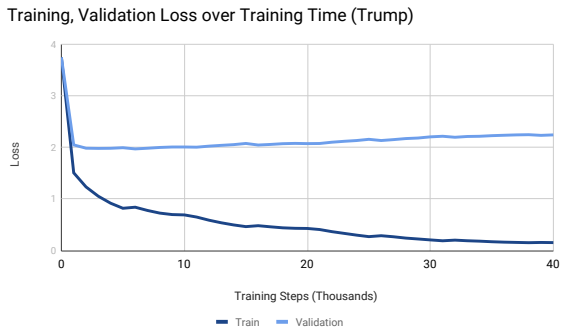

Figure 2: **Loss Curve**

The continuously increasing training accuracy, however, indicates overfitting. For this reason, I froze the weights and trained only the dense layer (and tested adding additional dense layers as well); however, this showed no improvement (Figure 3):

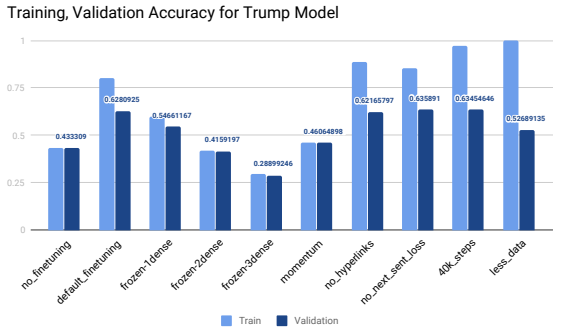

Figure 3: **Trump Model: Fine-Tuning Results**

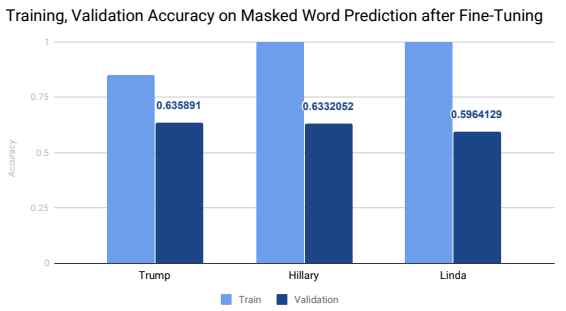

Figure 4: **Accuracy Per Fine-Tuned Model**

In the end, turning off next_sentence_loss performed best, resulting in a validation accuracy of 0.6359. I applied this setting to the two other models (Hillary and Linda models) as well. For the Hillary model, I pulled 6,426 Tweets from two sites [5]; after similar data cleaning, this resulted in 5,151 Tweets (4,635 for training and 516 for validation). For the Linda model, I pulled 6,038 posts and comments from my Facebook profile in a direct export; after filtering out re-posts and non-authored items (such as friends' comments on my posts), this resulted in 3,513 examples (3,162 for training and 351 for validation).

While Hillary and Linda validation sets are much smaller than the Trump set, validation accuracies don't diverge much as seen in Figure 4. One interpretation of this is that Trump may be harder to predict than Hillary and Linda. Also, while these validation accuracies may initially seem low, for training on this little data, performance is not bad, given BERT's vocabulary size is 32K, meaning random predictions would generate accuracies of 1/32K.

5 Scoring

Kurita et al's scoring methodology required the definition of frameworks. Specifically, each framework required three things: a set of "target" gendered words, a set of neutral "attribute" words, and a set of sentence templates to define the relationship between "target" and "attribute" words, as shown below (this example showcases the career/family evaluation framework leveraged from WEAT):

"Male" targets	he, boys, men
"Female" targets	she, girls, women
"Career" attributes	executive, management, professional, corporation, salary, office, business, career
"Family" attributes	home, parents, children, family, cousins, marriage, wedding, relatives
Templates	[TARGET] likes [ATTRIBUTE]. [TARGET] like [ATTRIBUTE]. [TARGET] is interested in [ATTRIBUTE].

Table 1: **Career/Family Framework**

Multiple words are used per list to address polysemy and concept completeness. For each combination of attributes and templates, I constructed evaluation sentences. [MASK] is used in place of the [TARGET] placeholder. For example:

- Attribute "executive": "[MASK] likes executive." "[MASK] likes executive." "[MASK] is interested in executive."

[5] https://www.kaggle.com/benhamner/clinton-trump-tweets, https://www.vicinitas.io/free-tools/download-user-tweets

- Attribute “home”: “[MASK] likes home.” “[MASK] likes home.” “[MASK] is interested in home.”
- And so on, for all attributes. This example generates a total of 96 sentences: 24 for male/career, 24 for male/family, 24 for female/career, and 24 for female/family.

For each of these 96 evaluation sentences, I computed the log probability that the model predicts each of the defined target words in place of [MASK]: log((P ([MASK] = target | sentence))). Here, each sentence generates 6 initial target probabilities - one each for “he”, “boys”, “men”, “she”, “girls”, and “women.” This example generates a total of 576 initial target probabilities.

In order to address a model’s inherent bias towards any of the target words (regardless of attribute), I also calculated the prior probability for each of the target words. First, each evaluation sentence is modified by the replacement of [MASK] in place of [ATTRIBUTE]. For example:

- Attribute “executive”: “[MASK] is [MASK].” “[MASK] likes [MASK].” “[MASK] is interested in [MASK].”
- Attribute “home”: “[MASK] likes [MASK].” “[MASK] likes [MASK].” “[MASK] is interested in [MASK].”
- And so on, for a total of 96 sentences.

Each sentence still generates 6 prior probabilities by predicting against only the first [MASK] in each evaluation sentence (log((P ([MASK] = target | sentence))) - one each for “he”, “boys”, “men”, “she”, “girls”, and “women.” This generates a total of 576 prior probabilities.

I then divided each initial target probability by each corresponding prior probability in order to produce a normalized probability. This generated one normalized probability per combination of target word and evaluation sentence (in this example, 576 combinations). These normalized probabilities are then averaged per gender and per attribute concept, resulting in 4 scores: male/career, male/family, female/career, and male/family. Gender bias is calculated as [female average - male average], which is defined as pro-female bias. This pro-female bias is generated for each attribute concept, resulting in just two scores: one for career, and one for family.

Each framework has different target words, attribute words, and template structures, but the measurement methodology remains the same. I modified the pleasant/unpleasant framework to include more vocabulary in the attribute (it is unfortunately too long to include directly [6]), but here is my own leader/follower evaluation framework:

"Male" Targets	he, boys, man, men, masculine
"Female" Targets	she, girls, woman, women, feminine
"Leader" Attributes	leader, senior, manager, executive, command, powerful, dominant, strong
"Follower" Attributes	follower, junior, subordinate, assistant, obey, powerless, meek, weak
Templates	[TARGET] is [ATTRIBUTE]. [TARGET] are [ATTRIBUTE].

Table 2: **Leader/Follower Framework**

6 Results

For my baseline, BERT-base, my results [7] were as seen in Figure 5 (light blue).

To the left (negative scale) is pro-male, and to the right (positive scale) is pro-female. BERT is generally straddling the middle, but displays clear pro-male bias for career and leader attributes.

To validate scoring performance, I qualitatively evaluated the top target predictions per evaluation sentence [8]. I honed in on the career attribute, given the relatively large bias compared to other attributes (below are normalized log probabilities):

[MASK] likes ...	executive	management	professional
Top 1	he | 0.08	he | 0.35	men | 1.19
Top 2	she | -0.27	she | -0.01	boys | 0.33
Top 3	men | -0.51	boys | -1.98	he | 0.01
Top 4	boys | -0.85	girls | -2.19	women | -0.21
Top 5	girls | -1.00	men | -2.27	girls | -0.22
Top 6	women | -1.06	women | -2.79	she | -0.83

Table 3: **Top 6 Predictions (word | log p)**

These predictions seem consistent, and there are no concerning issues in these results, giving me

[6] Details here: https://github.com/lindadongx/bert/tree/master/data/sets/unpleasant_pleasant

[7] Details here: https://github.com/lindadongx/bert/blob/master/data/results/bias.txt

[8] Details here: https://github.com/lindadongx/bert/tree/master/data/results/topk

more confidence that the scores are valid. Next up is Trump (Figure 5, dark blue).

The Trump model is clearly pro-male for the leader attribute - though oddly also pro-male for the follower attribute, with a surprisingly small pro-male bias on career. The model demonstrates clear pro-female bias for both family and unpleasant attributes.

The Hillary model (Figure 5, yellow) is surprisingly pro-male for the career and family attributes, and, to a lesser extent, unpleasant. She is pro-female for pleasant and leader attributes (no surprises there), though also oddly for follower.

The Linda model results are shown in green in Figure 5. I exhibit strong pro-male bias for career and leader attributes, and the only pro-female bias I have is for the follower attribute. This is a little surprising, but gender bias tends to be a blind spot for most people, and I tend to trust the model more than my conscious mind.

Looking at everything together, I start to see more social biases, where all models tend to exhibit bias in the same direction (Figure 5).

For the career attribute, for example, everyone is pro-male. The follower, pleasant, and unpleasant attributes are pretty evenly split. For the leader attribute, everyone is pro-male except for Hillary. And the magnitude of bias for the family attribute is even, especially between Trump and Hillary.

Taking a simple average across all attributes (Figure 5, last row), it appears everyone is pro-male - while Hillary is least so, I am twice as biased as Trump.

7 Future Work

One thing that can be improved in this project is better metrics. This includes defining a better way to compare results for opposing attributes (like leader and follower) - including potentially developing this into an evaluation metric for scoring. This also includes defining a better way of averaging across attributes for one model, rather than just taking the simple average.

Taking one step back, one thing that was tricky about this project is the difficulty of defining the right attributes and frameworks, as these might reflect bias inherently. It would be ideal if these attributes did not need to be explicitly defined, but the fact is, with gender bias, only some attributes are problematic. Take the sentences "women have xx chromosomes, and men have xy chromosomes" and "women have submissive natures, and men have aggressive natures." The first seems acceptable, while the second does not. This is because gender bias is ultimately dependent on values, beliefs, and sometimes intent, which are hard to infer, even for humans. One improvement here would be to define a machine learning approach to learn the right attributes over time, rather than have to explicitly define them.

8 Conclusion

In conclusion, I would like to propose that we use machines to help us do the things we're worst at. Rather than silently neutralize our shortcomings, they can help us better identify and manage them - by measuring them with precision and accuracy, and nudging us towards behavioral change. Just as Apple Watch applications can measure our heart rates and activity levels, so may future applications bring visibility not just into our health, but into our biases as well - if we can get the metrics right.

References

[Bolukbasi et al.2016] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

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

[Garg et al.2018] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

[Greenwald et al.1998] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

[Kurita et al.2019] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

[Sun et al.2019] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

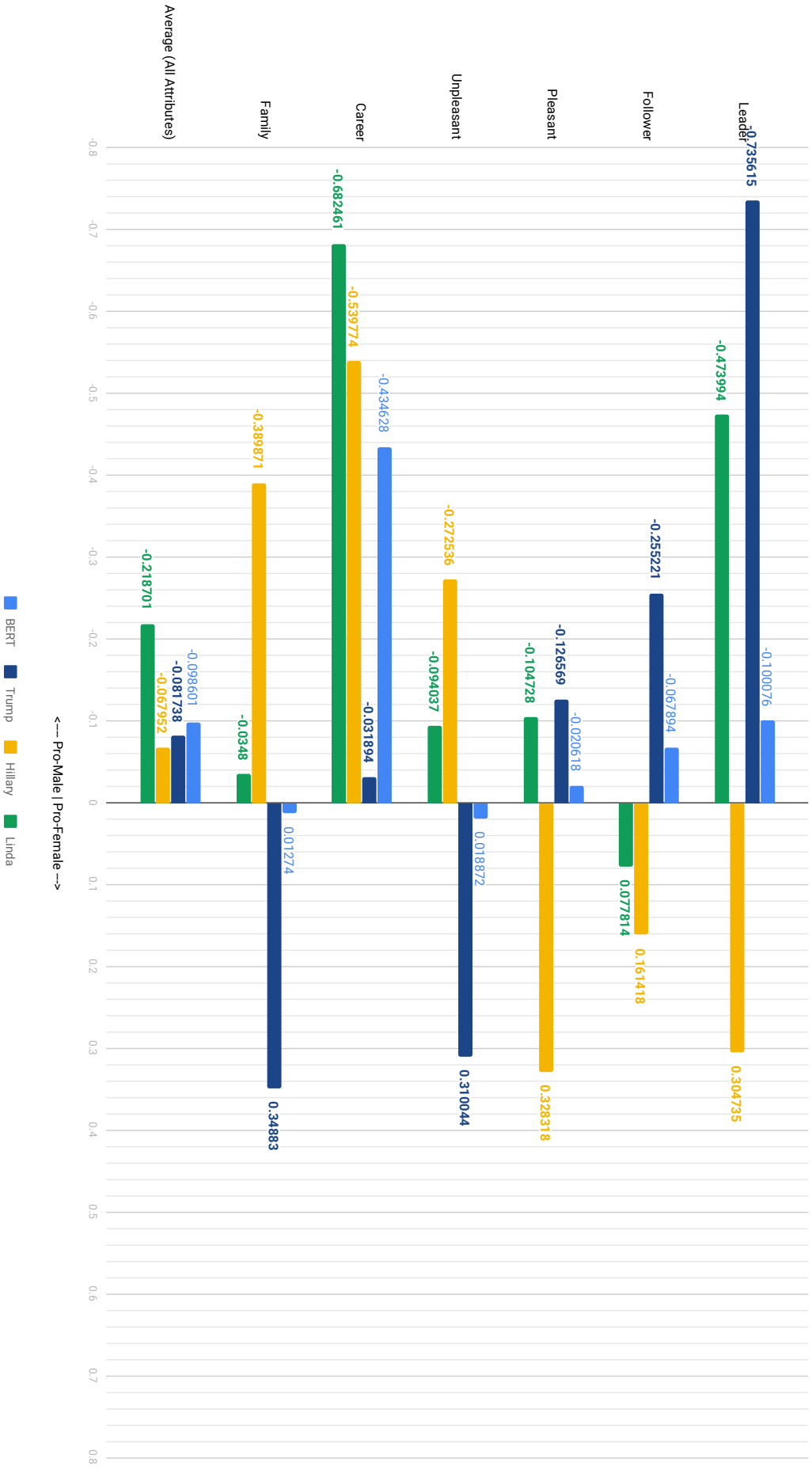

Figure 5: **Bias Scores**