

Quantifying Social Biases in Contextual Word Representations

Keita Kurita Nidhi Vyas Ayush Pareek Alan W Black Yulia Tsvetkov

Carnegie Mellon University

{kkurita, nkvyas, apareek, awb, ytsvetko}@andrew.cmu.edu

Abstract

Contextual word embeddings such as BERT have achieved state of the art performance in numerous NLP tasks. Since they are optimized to capture the statistical properties of training data, they tend to pick up on and amplify social stereotypes present in the data as well. In this study, we (1) propose a template-based method to quantify bias in BERT; (2) show that this method obtains more consistent results in capturing social biases than the traditional cosine based method; and (3) conduct a case study, evaluating gender bias in a downstream task of Gender Pronoun Resolution. Although our case study focuses on gender bias, the proposed technique is generalizable to unveiling other biases, including in multiclass settings, such as racial and religious biases.

1 Introduction

Type-level word embedding models, including word2vec and GloVe (Mikolov et al., 2013; Pennington et al., 2014), have been shown to exhibit social biases present in human-generated training data (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018; Manzini et al., 2019). These embeddings are then used in a plethora of downstream applications, which perpetuate and further amplify stereotypes (Zhao et al., 2017; Leino et al., 2019). To reveal and quantify corpus-level biases in word embeddings, Bolukbasi et al. (2016) used the word analogy task (Mikolov et al., 2013). For example, they showed that gendered male word embeddings like *he*, *man* are associated with higher-status jobs like *computer programmer* and *doctor*, whereas gendered words like *she* or *woman* are associated with *homemaker* and *nurse*.

Contextual word embedding models, such as ELMo and BERT (Peters et al., 2018; Devlin et al., 2019) have become increasingly common, replacing traditional type-level embeddings and attaining new state of the art results in the majority of

NLP tasks. In these models, every word has a different embedding, depending on the context and the language model state; in these settings, the analogy task used to reveal biases in uncontextualized embeddings is not applicable. Recently, May et al. (2019) showed that traditional cosine-based methods for exposing bias in sentence embeddings fail to produce consistent results for embeddings generated using contextual methods. We find similar inconsistent results with cosine-based methods of exposing bias; this is a motivation to the development of a novel bias test that we propose.

In this work, we propose a new method to quantify bias in BERT embeddings (§2). Since BERT embeddings use a *masked* language modelling objective, we directly query the model to measure the bias for a particular token. More specifically, we create simple template sentences containing the attribute word for which we want to measure bias (e.g. *programmer*) and the target for bias (e.g. *she* for gender). We then mask the attribute and target tokens sequentially, to get a relative measure of bias across target classes (e.g. male and female). Contextualized word embeddings for a given token change based on its context, so such an approach allows us measure the bias for similar categories divergent by the target attribute (§2). We compare our approach with the cosine similarity-based approach (§3) and show that our measure of bias is more consistent with human biases and is sensitive to a wide range of biases in the model using various stimuli presented in Caliskan et al. (2017). Next, we investigate the effect of a specific type of bias in a specific downstream task: gender bias in BERT and its effect on the task of Gendered Pronoun Resolution (GPR) (Webster et al., 2018). We show that the bias in GPR is highly correlated with our measure of bias (§4). Finally, we highlight the potential negative impacts of using BERT in downstream real world applications (§5). The code and data used in this work are publicly

available.¹

2 Quantifying Bias in BERT

BERT is trained using a masked language modelling objective i.e. to predict masked tokens, denoted as [MASK], in a sentence given the entire context. We use the predictions for these [MASK] tokens to measure the bias encoded in the actual representations.

We directly query the underlying masked language model in BERT² to compute the association between certain **targets** (e.g., gendered words) and **attributes** (e.g. career-related words). For example, to compute the association between the target *male gender* and the attribute *programmer*, we feed in the masked sentence “[MASK] is a programmer” to BERT, and compute the probability assigned to the sentence ‘*he* is a programmer’ (p_{tgt}). To measure the association, however, we need to measure how much *more* BERT prefers the male gender association with the attribute *programmer*, compared to the female gender. We thus re-weight this likelihood p_{tgt} using the prior bias of the model towards predicting the male gender. To do this, we mask out the attribute *programmer* and query BERT with the sentence “[MASK] is a [MASK]”, then compute the probability BERT assigns to the sentence ‘*he* is a [MASK]’ (p_{prior}). Intuitively, p_{prior} represents how likely the word *he* is in BERT, given the sentence structure and no other evidence. Finally, the difference between the normalized predictions for the words *he* and *she* can be used to measure the gender bias in BERT for the *programmer* attribute.

Generalizing, we use the following procedure to compute the association between a target and an attribute:

1. Prepare a template sentence
e.g. “[TARGET] is a [ATTRIBUTE]”
2. Replace [TARGET] with [MASK] and compute $p_{tgt} = P([MASK] = [TARGET] | \text{sentence})$
3. Replace both [TARGET] and [ATTRIBUTE] with [MASK], and compute prior probability $p_{prior} = P([MASK] = [TARGET] | \text{sentence})$
4. Compute the association as $\log \frac{p_{tgt}}{p_{prior}}$

¹<https://bit.ly/2EkJwh1>

²For all experiments we use the uncased version of BERT_{BASE} https://storage.googleapis.com/bert_models/2018_10_18/uncased_L-12_H-768_A-12.zip.

We refer to this normalized measure of association as the *increased log probability* score and the difference between the increased log probability scores for two targets (e.g. he/she) as *log probability bias score* which we use as measure of bias. Although this approach requires one to construct a template sentence, these templates are merely simple sentences containing attribute words of interest, and can be shared across multiple targets and attributes. Further, the flexibility to use such templates can potentially help measure more fine-grained notions of bias in the model.

In the next section, we show that our proposed *log probability bias score* method is more effective at exposing bias than traditional cosine-based measures.

3 Correlation with Human Biases

We investigate the correlation between our measure of bias and human biases. To do this, we apply the log probability bias score to the same set of attributes that were shown to exhibit human bias in experiments that were performed using the Implicit Association Test (Greenwald et al., 1998). Specifically, we use the stimuli used in the Word Embedding Association Test (WEAT) (Caliskan et al., 2017).

Word Embedding Association Test (WEAT):

The WEAT method compares set of target concepts (e.g. male and female words) denoted as X and Y (each of equal size N), with a set of attributes to measure bias over social attributes and roles (e.g. career/family words) denoted as A and B . The degree of bias for each target concept t is calculated as follows:

$$s(t, A, B) = [\text{mean}_{a \in A} \text{sim}(t, a) - \text{mean}_{b \in B} \text{sim}(t, b)],$$

where *sim* is the cosine similarity between the embeddings. The test statistics is

$$S(X, Y, A, B) = [\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)],$$

where the test is a permutation test over X and Y . The p -value is computed as

$$p = \Pr[S(X_i, Y_i, A, B) > S(X, Y, A, B)]$$

The effect size is measured as

$$d = \frac{S(X, Y, A, B)}{\text{std}_{t \in X \cup Y} s(t, A, B)}$$

Category	Templates
Pleasant/Unpleasant (Insects/Flowers)	T are A, T is A
Pleasant/Unpleasant (EA/AA)	T are A, T is A
Career/Family (Male/Female)	T likes A, T like A, T is interested in A
Math/Arts (Male/Female)	T likes A, T like A, T is interested in A
Science/Arts (Male/Female)	T likes A, T like A, T is interested in A

Table 1: Template sentences used for the WEAT tests (T: target, A: attribute)

Category	Targets	Templates
Pleasant/Unpleasant (Insects/Flowers)	flowers,insects,flower,insect	T are A, the T is A
Pleasant/Unpleasant (EA/AA)	black, white	T people are A, the T person is A
Career/Family (Male/Female)	he,she,boys,girls,men,women	T likes A, T like A, T is interested in A
Math/Arts (Male/Female)	he,she,boys,girls,men,women	T likes A, T like A, T is interested in A
Science/Arts (Male/Female)	he,she,boys,girls,men,women	T likes A, T like A, T is interested in A

Table 2: Template sentences used and target words for the grammatically correct sentences (T: target, A: attribute)

It is important to note that the statistical test is a permutation test, and hence a large effect size does not guarantee a higher degree of statistical significance.

3.1 Baseline: WEAT for BERT

To apply the WEAT method on BERT, we first compute the embeddings for target and attribute words present in the stimuli using multiple templates, such as “TARGET is ATTRIBUTE” (Refer Table 1 for an exhaustive list of templates used for each category). We mask the TARGET to compute the embedding³ for the ATTRIBUTE and vice versa. Words that are absent in the BERT vocabulary are removed from the targets. We ensure that the number of words for both targets are equal, by removing random words from the smaller target set. To confirm whether the reduction in vocabulary results in a change of p -value, we also conduct the WEAT on GloVe with the reduced vocabulary.⁴

3.2 Proposed: Log Probability Bias Score

To compare our method measuring bias, and to test for human-like biases in BERT, we also compute the *log probability bias score* for the same set of attributes and targets in the stimuli. We compute the mean *log probability bias score* for each attribute, and permute the attributes to measure statistical significance with the permutation test. Since many TARGETs in the stimuli cause the template sentence to become grammatically

incorrect, resulting in low predicted probabilities, we fixed the TARGET to common pronouns/indicators of category such as *flower*, *he*, *she* (Table 2 contains a full list of target words and templates). This avoids large variance in predicted probabilities, leading to more reliable results. The effect size is computed in the same way as the WEAT except the standard deviation is computed over the mean *log probability bias scores*.

We experiment over the following categories of stimuli in the WEAT experiments: Category 1 (flower/insect targets and pleasant/unpleasant attributes), Category 3 (European American/African American names and pleasant/unpleasant attributes), Category 6 (male/female names and career/family attributes), Category 7 (male/female targets and math/arts attributes) and Category 8 (male/female targets and science/arts attributes).

3.3 Comparison Results

The WEAT on GloVe returns similar findings to those of Caliskan et al. (2017) except for the European/African American names and pleasant/unpleasant association not exhibiting significant bias. This is due to only 5 of the African American names being present in the BERT vocabulary. The WEAT for BERT fails to find any statistically significant biases at $p < 0.01$. This implies that WEAT is not an effective measure for bias in BERT embeddings, or that methods for constructing embeddings require additional investigation. In contrast, our method of querying the underlying language model exposes statistically significant association across all categories, showing that BERT does indeed encode biases and that our method is more sensitive to them.

³We use the outputs from the final layer of BERT as embeddings

⁴WEAT was originally used to study the GloVe embeddings

Category	WEAT on GloVe	WEAT on BERT	Ours on BERT <i>Log Probability Bias Score</i>
Pleasant/Unpleasant (Insects/Flowers)	1.543*	0.6688	0.8744*
Pleasant/Unpleasant (EA/AA)	1.012	1.003	0.8864*
Career/Family (Male/Female)	1.814*	0.5047	1.126*
Math/Arts (Male/Female)	1.061	0.6755	0.8495*
Science/Arts (Male/Female)	1.246*	0.8815	0.9572*

Table 3: Effect sizes of bias measurements on WEAT Stimuli. (* indicates significant at $p < 0.01$)

Gender	Prior Prob.	Avg. Predicted Prob.
Male	10.3%	11.5%
Female	9.8%	13.9%

Table 4: Probability of pronoun referring to neither entity in a sentence of GPR

4 Case Study: Effects of Gender Bias on Gendered Pronoun Resolution

Dataset We examined the downstream effects of bias in BERT using the Gendered Pronoun Resolution (GPR) task (Webster et al., 2018). GPR is a sub-task in co-reference resolution, where a pronoun-containing expression is to be paired with the referring expression. Since pronoun resolving systems generally favor the male entities (Webster et al., 2018), this task is a valid test-bed for our study. We use the GAP dataset⁵ by Webster et al. (2018), containing 8,908 human-labeled ambiguous pronoun-name pairs, created from Wikipedia. The task is to classify whether an ambiguous pronoun P in a text refers to entity A , entity B or neither. There are 1,000 male and female pronouns in the training set each, with 103 and 98 of them not referring to any entity in the sentence, respectively.

Model We use the model suggested on Kaggle,⁶ inspired by Tenney et al. (2019). The model uses BERT embeddings for P , A and B , given the context of the input sentence. Next, it uses a multi-layer perceptron (MLP) layer to perform a naive classification to decide if the pronoun belongs to A , B or neither. The MLP layer uses a single hidden layer with 31 dimensions, a dropout of 0.6 and L2 regularization with weight 0.1.

Results Although the number of male pronouns associated with no entities in the training data is

⁵<https://github.com/google-research-datasets/gap-coreference>

⁶<https://www.kaggle.com/mateiionita/taming-the-bert-a-baseline>

slightly larger, the model predicted the female pronoun referring to no entities with a significantly higher probability ($p = 0.007$ on a permutation test); see Table 4. As the training set is balanced, we attribute this bias to the underlying BERT representations.

We also investigate the relation between the topic of the sentence and model’s ability to associate the female pronoun with no entity. We first extracted 20 major topics from the dataset using non-negative matrix factorization (Lee and Seung, 2001) (refer to Appendix for the list of topics). We then compute the bias score for each topic as the sum of the *log probability bias* score for the top 15 most prevalent words of each topic weighted by their weights within the topic. For this, we use a generic template “[TARGET] are interested in [ATTRIBUTE]” where TARGET is either men or women. Next we compute a bias score for each sample in the training data as the sum of individual bias scores of topics present in the sample, weighted by the topic weights. Finally, we measured the Spearman correlation coefficient to be 0.207 (which is statistically significant with $p = 4e - 11$) between the bias scores for male gender across all samples and the model’s probability to associate a female pronoun with no entity. We conclude that models using BERT find it challenging to perform coreference resolution when the gender pronoun is female and if the topic is biased towards the male gender.

5 Real World Implications

In previous sections, we discussed that BERT has human-like biases, which are propagated to downstream tasks. In this section, we discuss another potential negative impact of using BERT in a downstream model. Given that three quarters of US employers now use social media for recruiting job candidates (Segal, 2014), many applications are filtered using job recommendation systems and other AI-powered services. Zhao et al. (2018)

discussed that resume filtering systems are biased when the model has strong association between gender and certain professions. Similarly, certain gender-stereotyped attributes have been strongly associated with occupational salary and prestige (Glick, 1991). Using our proposed method, we investigate the gender bias in BERT embeddings for certain occupation and skill attributes.

Datasets: We use three datasets for our study of gender bias in employment attributes:

- *Employee Salary Dataset*⁷ for Montgomery County of Maryland- Contains 6882 instances of “Job Title” and “Salary” records along with other attributes. We sort this dataset in decreasing order of salary and take the first 1000 instances as a proxy for high-paying and prestigious jobs.
- *Positive and Negative Traits Dataset*⁸- Contains a collection of 234 and 292 adjectives considered “positive” and “negative” traits, respectively.
- *O*NET 23.2 technology skills*⁹ Contains 17649 unique skills for 27660 jobs, which are posted online

Discussion We used the following two templates to measure gender bias:

- “TARGET is ATTRIBUTE”, where TARGET are male and female pronouns viz. *he* and *she*. The ATTRIBUTE are job titles from the Employee Salary dataset, or the adjectives from the Positive and Negative traits dataset.
- “TARGET can do ATTRIBUTE”, where the TARGETs are the same, but the ATTRIBUTE are skills from the O*NET dataset.

Table 5 shows the percentage of attributes that were more strongly associated with the male than the female gender. The results prove that BERT expresses strong preferences for male pronouns, raising concerns with using BERT in downstream tasks like resume filtering.

⁷<https://catalog.data.gov/dataset/employee-salaries-2017>

⁸<http://ideonomy.mit.edu/essays/traits.html>

⁹<https://www.onetcenter.org/database.html#individual-files>

Dataset	Percentage
Salary	88.5%
Pos-Traits	80.0%
Neg-Traits	78.9%
Skills	84.0%

Table 5: Percentage of attributes associated more strongly with the male gender

6 Related Work

NLP applications ranging from core tasks such as coreference resolution (Rudinger et al., 2018) and language identification (Jurgens et al., 2017), to downstream systems such as automated essay scoring (Amorim et al., 2018), exhibit inherent social biases which are attributed to the datasets used to train the embeddings (Barocas and Selbst, 2016; Zhao et al., 2017; Yao and Huang, 2017). There have been several efforts to investigate the amount of intrinsic bias within uncontextualized word embeddings in binary (Bolukbasi et al., 2016; Garg et al., 2018; Swinger et al., 2019) and multiclass (Manzini et al., 2019) settings.

Contextualized embeddings such as BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018) have been replacing the traditional type-level embeddings. It is thus important to understand the effects of biases learned by these embedding models on downstream tasks. However, it is not straightforward to use the existing bias-exposure methods for contextualized embeddings. For instance, May et al. (2019) used WEAT on sentence embeddings of ELMo and BERT, but there was no clear indication of bias. Rather, they observed counterintuitive behavior like vastly different p -values for results concerning gender.

Along similar lines, Basta et al. (2019) noted that contextual word-embeddings are less biased than traditional word-embeddings. Yet, biases like gender are propagated heavily in downstream tasks. For instance, Zhao et al. (2019) showed that ELMo exhibits gender bias for certain professions. As a result, female entities are predicted less accurately than male entities for certain occupation words, in the coreference resolution task. Field and Tsvetkov (2019) revealed biases in ELMo embeddings that limit their applicability across data domains. Motivated by these recent findings, our work proposes a new method to expose and measure bias in contextualized word embeddings, specifically BERT. As opposed to previ-

ous work, our measure of bias is more consistent with human biases. We also study the effect of this intrinsic bias on downstream tasks, and highlight the negative impacts of gender-bias in real world applications.

7 Conclusion

In this paper, we showed that querying the underlying language model can effectively measure bias in BERT and expose multiple stereotypes embedded in the model. We also showed that our measure of bias is more consistent with human-biases, and outperforms the traditional WEAT method on BERT. Finally we showed that these biases can have negative downstream effects. In the future, we would like to explore the effects on other downstream tasks such as text classification, and devise an effective method of debiasing contextualized word embeddings.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS1812327.

References

- Evelin Amorim, Marcia Cançado, and Adriano Veloso. 2018. Automated essay scoring in the presence of biased ratings. In *Proc. of NAACL*, pages 229–237.
- Solon Barocas and Andrew D Selbst. 2016. Big data’s disparate impact. *Calif. L. Rev.*, 104:671.
- Christine Basta, Marta R Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783*.
- 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 *Proc. of NIPS*, pages 4349–4357.
- Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.
- Anjalie Field and Yulia Tsvetkov. 2019. Entity-centric contextual affective analysis. In *Proc. of ACL*.
- 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.
- Peter Glick. 1991. Trait-based and sex-based discrimination in occupational prestige, occupational salary, and hiring. *Sex Roles*, 25(5-6):351–378.
- Anthony Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of personality and social psychology*, 74:1464–80.
- David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proc. of ACL*, pages 51–57.
- Daniel Lee and Hyunjeong Seung. 2001. Algorithms for non-negative matrix factorization. In *Proc. of NIPS*.
- Klas Leino, Matt Fredrikson, Emily Black, Shayak Sen, and Anupam Datta. 2019. Feature-wise bias amplification. In *Proc. of ICLR*.
- Thomas Manzini, Yao Chong, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multi-class bias in word embeddings. In *Proc. of NAACL*.
- Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proc. of NAACL*.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositional-ity. In *Proc. of NIPS*, pages 3111–3119.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proc. of EMNLP*, pages 1532–1543.
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proc. of NAACL*.
- J Segal. 2014. Social media use in hiring: Assessing the risks. *HR Magazine*, 59(9).
- Nathaniel Swinger, Maria De-Arteaga, Neil Heffernan IV, Mark Leiserson, and Adam Kalai. 2019. What are the biases in my word embedding? In *Proc. of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*.

- Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations.
- Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*.
- Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems*, pages 2921–2930.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *NAACL (short)*.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proc. of EMNLP*.
- Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings.

Appendix

Topic Id	Top 5 Words
1	match,round,second,team,season
2	times,city,jersey,york,new
3	married,son,died,wife,daughter
4	best,award,actress,films,film
5	friend,like,work,mother,life
6	university,music,attended,high,school
7	president,general,governor,party,state
8	songs,solo,song,band,album
9	medal,gold,final,won,world
10	best,role,character,television,series
11	kruse,moved,amy,esme,time
12	usa,trunchbull,pageant,2011,miss
13	american,august,brother,actress,born
14	sir,died,church,song,john
15	natasha,days,hospital,helene,later
16	played,debut,sang,role,opera
17	january,december,october,july,married
18	academy,member,american,university,family
19	award,best,played,mary,year
20	jersey,death,james,king,paul

Table 6: Extracted topics for the GPR dataset