Exploring Debiasing Sentence Representations

Grace Cuenca
gcuenca@

Leslie Kim
lesliek@

Nicole Meister
nmeister@

Abstract

SENT-DEBIAS is a method proposed by Liang et al. (2020) to reduce the social bias present in sentence-level embeddings from Bidirectional Encoder Representations from Transformers (BERT). In this study, we first reproduce their results to demonstrate the effectiveness of the proposed SENT-DEBIAS algorithm in debiasing BERT fine-tuned on three different datasets, SST, COLA, and QNLI. We then present three new ablations on the model as well as a discussion of the results. We experiment with the input representation, the evaluation data, and the pre-processing of the training data.

1 Introduction

Recently, word and sentence embeddings have become one of the most powerful tools used in natural language processing technology. Word and sentence embeddings are essentially dense vectors used to represent words and sentences in natural text. Models such as FastText (Bojanowski et al., 2016), ELMo (Peters et al., 2018), and BERT (Devlin et al., 2019) offer pre-trained embeddings generated from large textual corpora. These embeddings can simply be fine-tuned for any downstream tasks.

One issue that is becoming of increasing concern to natural language processing researchers is the presence of social bias within word level embeddings. For example, a study on word embeddings showed that male terms were more closely associated with career terms and female terms more closely associated with family terms (Islam et al., 2016). These sorts of bias are learned by the models and encoded in the embeddings. There have been recent successful efforts to reduce these biases present in word embeddings (Bolukbasi et al., 2016a). But it has been more difficult to do so for sentence level embeddings because of the immense computational power needed to retrain sentence-based embedding models, and the wider variety of possible sentences than words (Liang et al., 2020).

In the paper "Towards Debiasing Sentence Representations," Liang et al. (2020) present a method called SENT-DEBIAS which successfully reduces bias in the sentence level representations from BERT. They use SENT-DEBIAS on the pre-trained BERT encoder, as well as BERT fine-tuned on datasets such as the Stanford Sentiment Treebank (SST-2) sentiment classification (Socher et al., 2013), the Corpus of Linguistic Acceptability (CoLA) grammatical acceptability judgment (Warstadt et al., 2018), and the Question Natural Language Inference (QNLI) (Wang et al., 2018). Throughout the paper, we refer to these, respectively, as BERT, BERT post SST, BERT post CoLA, and BERT post QNLI. In addition, they evaluate the performance of these debiased sentence embeddings on classification tasks in order to make sure SENT-DEBIAS does not hurt performance on downstream tasks.

In section 2 we present an in-depth summary of the original paper as well as a description of work related to our ablations. Next, we present the results from reproducing some of the experiments in the original paper. In section 6, we share motivations, results, and analysis of each of the three ablations performed.

We developed these ablations by reading Liang et al. (2020)'s paper and code in depth and researching under-explored areas or ideas might help the performance of SENT-DEBIAS in reducing bias in embeddings. The first ablation uses representations for n sentence pairs instead of 1 sentence to estimate the bias subspace. For the second ablation, we replace the very short and repetitive evaluation sentences with ones which were hand-written to be longer more complex, and more diverse. The third ablation pre-processes the training data to in-

clude more people's names as bias-attribute words for gender. We present our explanations for our obtained results, and offer potential directions for future work.

2 Related Work

2.1 Measuring Presence of Bias in Sentence Representations

Much work has been done in recent years to measure and mitigate the amount of bias present in many machine learning fields, with natural language processing being no exception. We found Liang et al. (2020)'s related work to be extremely relevant in shedding light into our debiasing ablations. In particular, bias in sentence representations has been a primary focus, as seen in May et al. (2019) and Basta et al. (2019). An issue with this approach is that pretrained sentence representations are difficult to remove bias from. Zhao et al. (2019), Park et al. (2018) and Garg et al. (2019), who have also looked into debiasing sentence representations, have not been able to find a method to debias sentences after the sentence embeddings have already been produced. In order to circumvent this, they modify the individual word embeddings within the sentence and need to perform costly retraining. Bordia and Bowman (2019) start to look into language models that deal with individual word embeddings, but they too run into the issue of costly retraining. Kurita et al. (2019) performs the word-level Word Embedding Association Test (WEAT) (Caliskan et al., 2017) to measure the bias present in BERT, but runs into the problem of difficult post-hoc debiasing as well as costly retraining.

2.2 Overview of "Towards Debiasing Sentence Representations"

Liang et al. (2020) propose a method SENT-DEBIAS to reduce biases in sentence-level representations. We illustrate their method in Figure 1, summarize the four main steps of SENT-DEBIAS below, and describe the evaluation process.

1. **Define Bias Attribute:** Identify a d-class *bias attributes* and define a set of *bias attribute words* that are indicative of these attributes. For example, when mitigating gender bias in sentence representations, gender is the *bias attribute* with $d = 2$. The *bias attribute words* are defined as word pairs where each pair consists of words that has an equivalent meaning except for the bias attribute. Liang et al. (2020) uses 10 pairs of bias attribute words and 4 examples of those bias attribute words are depicted in Figure 1.1.

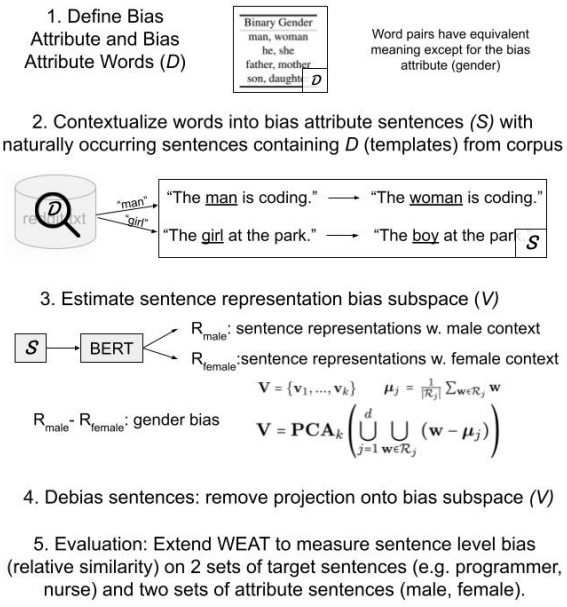

Figure 1: Our Illustration of SENT-DEBIAS Algorithm and Evaluation Method.

2. **Contextualize bias attribute words into bias attribute sentences:**
   (a) Problem: Converting words to sentences to obtain a representation from pre-trained encoders is difficult because the potential number of sentences these bias attribute words can occur in is unbounded.
   (b) Solution: Find naturally occurring sentences *(templates)* containing bias attribute words and replace bias attribute word for complementary word. As illustrated in 1.2, when the sentence "he went to the store" appears in the text corpus, they generate a new sentence "she went to the store" to populate S, the set of bias attribute sentences.
   (c) The text corpora originate from:
      i. WikiText-2 (Merity et al., 2016) (a dataset of formally written Wikipedia articles)
      ii. Stanford Sentiment Treebank (Zhang et al., 2018) (a collection of 10000 polarized written movie reviews)

iii. Reddit (data collected from discussion forums related to politics, electronics, and relationships)

iv. MELD (Poria et al., 2019) (a large-scale multimodal multi-party emotional dialog dataset collected from the TV-series Friends)

v. POM (Park et al., 2014) (a dataset of spoken review videos collected across 1,000 individuals spanning multiple topics)

3. **Estimate sentence representation bias subspace:**

   (a) Define R_i, R_j as sets that collect all BERT sentence representations from sentences containing man (i) and woman (j) bias attribute words respectively. Each set R defines a vector space which has a specific bias attribute present across all its contexts.

   (b) $R_i - R_j$ represents the presence of gender bias.

   (c) Define bias subspace $V = \{v_1, ..., v_k\}$ given by first k components of principal component analysis (PCA).

$$V = \text{PCA}_\text{k}(\bigcup_{j=1}^{d} \bigcup_{w \in R_j} w - \mu_j) \quad (1)$$

$$\mu_j = \frac{1}{R_j} \sum_{w \in R_j} \quad (2)$$

4. **Debias All Sentences by Removing Projection Onto Bias Subspace**: Liang et al. (2020) apply a partial version of the HARD-DEBIAS algorithm (Bolukbasi et al., 2016b) to remove bias from new sentence representations. Bias components are removed from sentences that are not gendered and should not contain gender bias (e.g., I am a doctor, That nurse is taking care of the patient.) by removing the projection onto the bias subspace. Given a representation h of a sentence and the estimated gender subspace $V = \{v_1, ..., v_k\}$, the debiased representation $\hat{h}$ is given by first obtaining h_v, the projection of h onto the bias subspace V before subtracting h_V from h. This results in a vector that is orthogonal to the bias subspace V, thus containing no bias.

To evaluate SENT-DEBIAS, Liang et al. (2020) extend the Word Embedding Association Test (WEAT) (Caliskan et al., 2017) to a Sentence Embedding Association Test (SEAT). WEAT measures bias by comparing two sets of target words *(programmer, nurse)* to two sets of attribute words *(male, female)* and measuring relative similarity. Unbiased word representations should display no difference between the two target words in terms of their relative similarity to the two sets of attribute words. The relative similarity as measured by WEAT is measured through the effect size where an absolute effect size closer to 0 represents lower bias.

Liang et al. (2020) also provided the sentences that they evaluated their models on. In particular, their gender tests contained several categories, with concrete label translations shown in Table 1. For example, test C6 contains male and female name sentences (i.e "This is John." and "This is Lisa.") as well as career and family term sentences that do not explicitly refer to gender (i.e "These are executives." and "Here is a relative."). The idea of these evaluation tests is to examine the specific word embeddings that are explicitly male or female names within the sentences and how close they are to career and family terms in the bias subspace.

2.3 Related Works Regarding Our Ablations

We initially sought to expand upon Liang et al. (2020) by measuring the effect of debiasing using additional metrics. There has been previous work to measure the presence of bias in addition to WEAT metrics. Liang et al. (2020) acknowledge that WEAT is not a perfect metric as it can only be used to detect the presence of biases, not absence. Wang and Russakovsky (2021) propose a Bias Amplification metric to quantify how much models amplify the biases present in the data they are trained on. However, given the nature of the SENT-DEBIAS technique where we only have access to the sentence embeddings from biased and debiased models to compare, applying this metric did not work and we did not pursue this route.

Instead we pursued the following three ablations: N-sentence representation, evaluation on "natural" sentences, and extending the bias attribute words to names. We originally decided to pursue the N-sentence representation ablation because we encountered many different nuances within the code that we explored to help understand the way nat-

ural language processing models learn bias. Regarding the N-sentence representation ablation, we believed it followed a natural hierarchy to go from word, to sentence, to potential paragraph embeddings.

However, we also found some interesting subtleties within the code as we dove deeper into this subject. For example, the evaluation corpus contains simple sentences for the model to validate on. Thus, we thought it would be worthwhile to assess how debiasing increases or decreases when more complex and descriptive sentence structures were introduced. A second interesting point we found was that the code did not give all gendered names equal weight when templating embeddings. We found in the code that the training data only recognizes "John" and "Mary" as male and female names. Thus, if other names were presented, they would not be appropriately classified as male or female names.

3 Statement of Purpose

In their 2007 paper, Liang et al. (2020) explored methods to mitigate the social bias present in sentence-level word embeddings like BERT and ELMo. They propose an algorithm called SENT-DEBIAS as a way to reduce the bias from these sentence representations. In this paper, we reproduce the baseline results of the SENT-DEBIAS algorithm as well as present three new model ablations.

4 Setup

Accessing the existing codebase was easy as the original authors had an updated GitHub and a helpful README. We met virtually with Paul Liang, one of the authors of the original paper, who informed us that the compute available on Colab for free should be enough to run the experiments we were planning to conduct. For the process of evaluating the bias of the word embeddings, the Tesla T4 GPU available on Colab was sufficient and usually took less than an hour. On the other hand, the process of fine-tuning BERT on the different datasets was one of the biggest difficulties in our project. We originally tried running the fine-tuning process on Google Colab with the Tesla T4 GPU. Even though the process was relatively fast, it was unreliable as after a few hours Colab would without warning disconnect us from the runtime and we would not be able to view the results. Therefore, we decided to run the fine-tuning on our local computers, which used Radeon Pro 555X. Fine-tuning locally took two to four days.

5 Reproduction of Baseline Results

To begin investigating effective methods of debiasing, we first reproduced the results of Liang et al. (2020) using a pre-trained BERT encoder and 3 varieties of fine-tuned BERT encoders. Each cell contains an entry $a \rightarrow b$ that represents a SEAT score (a measure of bias) from biased a to debiased b. In general, a SEAT score closer to 0 means there is less bias present. Tables 2 and 3 show the results of our baseline reproduction experiment.

Test Name	Categories
C6	M/F Names, Career/Family
C6b	M/F Terms, Career/Family
C7	M/F Terms, Math/Arts
C7b	M/F Names, Math/Arts
C8	M/F Terms, Science/Arts
C8b	M/F Names, Science/Arts

Table 1: Category Legend
[Note. Adapted from (Liang et al., 2020)]

Test	BERT	BERT post SST-2
C6	0.477 → **-0.089**	**-0.036** → -0.131
C6b	**0.108** → -0.434	**-0.010** → -0.088
C7	0.252 → **0.193**	**-0.219** → -0.564
C7b	0.254 → **0.193**	1.153 → **-0.671**
C8	0.399 → **-0.071**	**0.103** → 0.126
C8b	0.636 → **0.542**	0.222 → **0.054**

Table 2: Debiasing Results on BERT Variations Pt.1

Test	BERT post CoLA	BERT post QNLI
C6	**-0.004** → 0.0974	**-0.261** → 0.793
C6b	**-0.062** → 0.209	**-0.155** → 0.467
C7	**0.299** → 0.325	-0.584 → **0.328**
C7b	**0.152** → 0.331	-0.581 → **0.337**
C8	0.198 → **-0.188**	**-0.087** → 0.222
C8b	**-0.092** → 0.201	-0.521 → **-0.043**

Table 3: Debiasing Results on BERT Variations Pt.2

Furthermore, we also reproduced the visualizations that the paper provides. Figure 2 and Figure 3 show the t-SNE plots of average sentence representations of a word across its sentence templates in

the bias subspace before and after debiasing respectively. Notice that in Figure 2, the term "woman" is closer to the term "family" than "man" is. However, in Figure 3, "man" and "woman" are virtually equidistant from all other terms in the bias subspace.

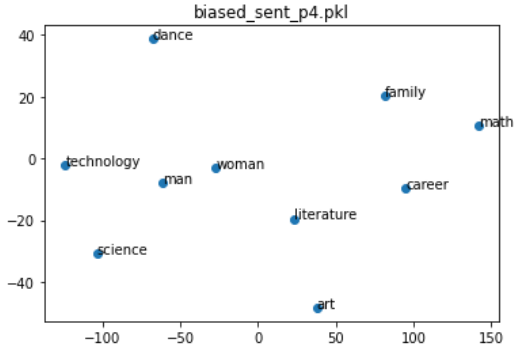

Figure 2: Sentence Embeddings Before Debiasing

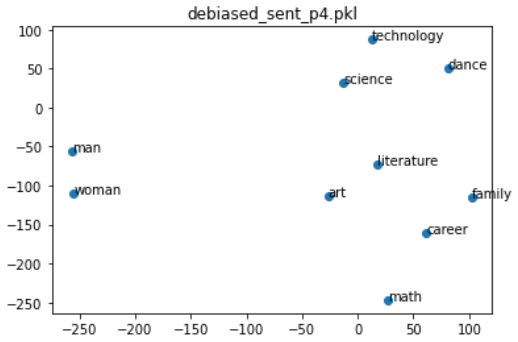

Figure 3: Sentence Embeddings After Debiasing

6 Ablations

6.1 N-Sentence Representation

Liang et al. (2020) aim to debias the representation of singular sentence representations and use individual sentences to estimate the bias subspace. We extend the SENT-DEBIAS algorithm to debias N-sentence representations. This is accomplished by editing the step depicted in Figure 1.2. Instead of contextualizing bias attribute words into single bias attribute sentences, the bias attribute words are contextualized into bias attribute sentences of size N. We then estimate the corresponding bias subspace, debias, and measure the average absolute values of all 6 SEAT effect sizes, which is now an N-Sentence Embedding Association Test as this algorithm is evaluated on sentences of size N. The core difference in this ablation is that sentences of size $N > 1$ are used to estimate the bias subspace.

We hypothesized that by increasing N (up to a certain threshold), the debiasing will be more effective because the bias subspace will be estimated with the context of more than just a singular sentence embedding. However, there will reach a point where the additional sentences does not provide additional benefit.

We evaluated our N-SENT-DEBIAS Algorithm in two experiments. First, we evaluated the N-SENT-DEBIAS Algorithm on $N = 2$ for all the Caliskan tests that Liang et al. (2020) used. Given the high variances in average absolute effect size, we ran this algorithm twice and took the average of the two runs. Secondly, we stress tested this algorithm by evaluating the debiasing effect of size $N = 1$ to $N = 8$ and plotted the average absolute effect size and the average standard deviation:

$$\sigma_N = \sqrt{\frac{\sum_{\text{test=C6}}^{\text{C8b}} \sigma^2_{\text{test}}}{6}}$$

The results of our first experiment regarding the 2-Sentence representation ablation is shown in Tables 4 and 5. From these results, we observe that when fine-tuning the BERT encoder for specific tasks, the bias is occasionally reduced.

The results of our second experiment are displayed in Figure 4. From these results, it seems as if N=2 and N=7 result in the smallest average absolute effect size when finetuning the BERT encoder for a specific task. SST-2 displayed the highest variance in average absolute effect size. Additionally, we observe a trend where bias is reduced with N=2, but the bias increases as N increases.

Test	BERT post SST-2
C6	**0.036** → 0.137
C6b	0.010 → **0.077**
C7	**0.219**→ 0.956
C7b	1.153→ **0.852**
C8	0.103 → **0.0532**
C8b	0.222 → **0.014**

Table 4: Debiasing Results when using $N = 2$ Sentences to Estimate Bias Subspace Pt.1 (Absolute average effect size of fine-tuning the model twice)

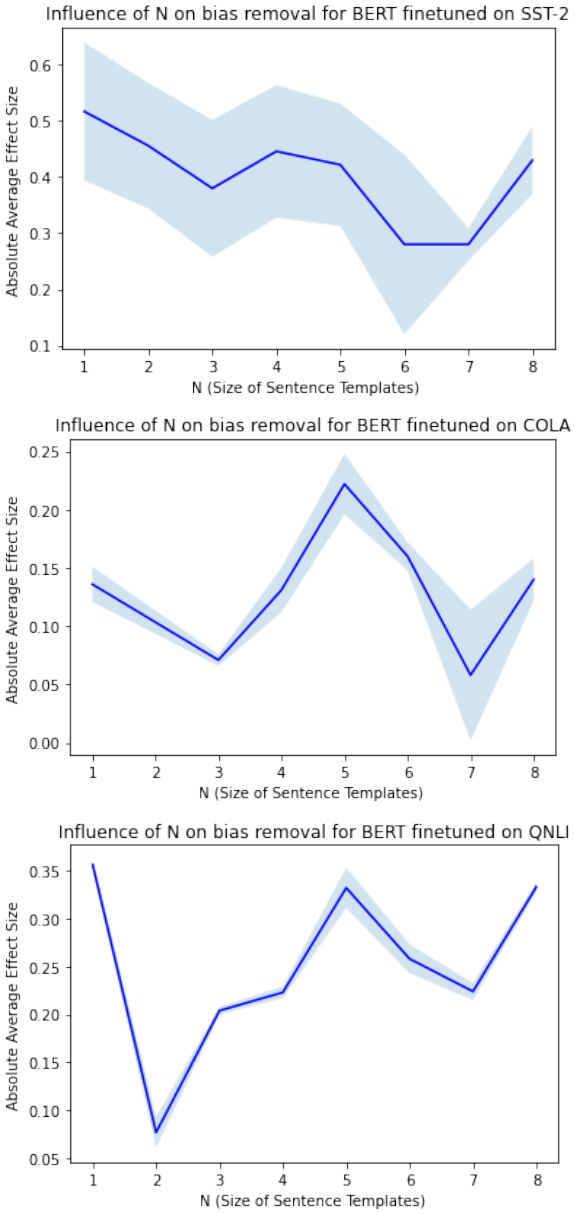

Figure 4: Influence of N on bias removal for BERT finetuned on SST-2, COLA, and QNLI. The solid line represents the mean over different combinations of domains and the shaded area represents the average standard deviation.

Test	BERT post CoLA	BERT post QNLI
C6	**0.004** → 0.188	0.261→ **0.168**
C6b	**0.062** → 0.121	**0.155** → 0.306
C7	0.299 → **0.118**	0.584 → **0.311**
C7b	0.152 → **0.066**	0.581 → **0.333**
C8	0.198 → **0.108**	**0.087** → 0.532
C8b	**0.092** → 0.143	0.521 → **0.218**

Table 5: Debiasing Results when using $N = 2$ Sentences to Estimate Bias Subspace Pt.2 (Absolute average effect size of fine-tuning the model twice)

Analysis and Discussion This ablation did decrease the bias occasionally in some downstream tasks, possibly due to the fact that in the N=2 case depicted in Table 4 and Table 5 there are twice as many templates which allows the model to use more examples to accurately estimate the bias subspace. However, fine-tuning does not lead to reliable and consistent decreases in bias and cannot be used as a standalone debiasing method, as Liang et al. (2020) noted with their original SENT-DEBIAS method.

6.2 Evaluating on Natural Sentences

When observing the sentences the BERT models were evaluated on, we noticed that their structures were all simple and very similar. For example, the C6 category (containing M/F names and career/family terms) contains sentences such as "This is John" and "The person's name is Lisa." Other interesting sentences from this category are "Corporations are things" and "That is a career." One can observe that these sentences are rather awkward and verbose. Thus, one ablation we focused on was adding more natural sounding sentences into the evaluation corpus to see how well the SENT-DEBIAS model would debias. Results are shown in Tables 6 and 7.

Test	BERT	BERT post SST-2
C6e	**-0.007** → -0.010	**-0.028** → -0.069
C6be	**-0.101** → -0.242	0.016 → **-0.001**
C7e	0.267 → **0.191**	**0.330** → 0.818
C7be	0.180 → **0.157**	-0.904 → **-0.695**
C8e	**0.133** → 0.261	-0.199 → **0.169**
C8be	**-0.075** → 0.285	0.170 → **0.007**

Table 6: Results on Adding Natural Sentences Ablation Pt.1

Test	BERT post CoLA	BERT post QNLI
C6e	**-0.007** → 0.055	**-0.041** → 0.394
C6be	**-0.101** → 0.191	**-0.037** → 0.234
C7e	**0.267** → 0.318	-0.670 → **0.330**
C7be	**0.180** → 0.347	-0.631 → **0.310**
C8e	**0.133** → 0.206	-0.631 → **-0.059**
C8be	**-0.075** → 0.287	-0.579 → **0.128**

Table 7: Results on Adding Natural Sentences Ablation Pt.2. Note that the 'e' added to the end of the tests refers to .jsonl files (for example C6, C6b, C7... etc.) that contain the authors' original sentences plus additional sentences that we created.

Analysis and Discussion Overall, this ablation did decrease the bias on the BERT post QNLI and BERT post SST-2 models, but not the BERT and BERT post CoLA models. From Table 7, we can observe that the BERT post CoLA model produced more biased sentence representations across all categories. As opposed to the BERT post SST-2 models and BERT post QNLI models, BERT post CoLA produces a model that is better at measuring the grammatical acceptability of a sentence, and adding more complex sentences that did not fit the typical simple sentence mold of the provided sentences most likely affected its debiasing score.

Although the other models did relatively well, one important factor to note is the lack of adjectives used to describe M/F terms/names and career/family/science/math/art terms in the original provided evaluation sentence files (sentence weat files). In order to add more complexity to our natural sentences; however, we experimented with adding adjectives and different sentence structures to the weat files to reflect what humans would say or type more naturally. For example, a sentence that we added to the C7 category under the Art heading was "This dance recital is very beautiful." The word "beautiful" may have associations to female names or terms that are reflected in the precomputed weights of the models we experimented on. However, because of the addition of the adjective, it was possibly difficult for the model to debias the sentence since the bias does not come from the arts category, but rather the biased nature of the word "beautiful."

6.3 Extending Bias Attribute Words to Names

The key idea of this ablation was pre-processing the sentences used to fine-tune BERT during training in order to better estimate the bias subspace. The existing code creates templates to estimate the bias subspace by looking for gendered pronouns such as "he"/"she" or "man"/"woman" as well as the single named pair "John"/"Mary." However, when looking through the training data, we found that many of the sentences included names other than John and Mary.

With the original method of identifying names, these sentences would not be recognized by the templates at all and would not contribute to estimating the bias subspace. We used Named Entity Recognition (NER)[1] capabilities from spaCy to identify entities with the label "PERSON" and the python tool gender_guesser[2] to change presumptive male names to "John" and presumptive female names to "Mary." Then the original code would identify these words as bias attribute words and use templates to generate a second corresponding sentence with the appropriate name pair.

We then used this new pre-processed version of the training data during the bias evaluation phase for BERT post SST-2, BERT post CoLA, and BERT post QNLI. The results are displayed in Tables 8 and 9.

Test	BERT post SST-2
C6	**-0.036** → -0.130
C6b	**-0.010** → 0.077
C7	**-0.219** → -0.991
C7b	1.153 → **1.096**
C8	**0.103** → 0.164
C8b	**0.222** →-0.261

Table 8: Results from Extending Bias Attribute Words to Names Pt.1

Test	BERT post CoLA	BERT post QNLI
C6	**-0.004** →-0.382	-0.261→ **0.190**
C6b	-0.062 → **0.041**	**-0.155** →-0.241
C7	0.299 → **-0.008**	-0.584 → **-0.296**
C7b	0.152 → **0.114**	-0.581 → **-0.294**
C8	0.198 → **0.148**	-0.087 →**-0.037**
C8b	**-0.092** → 0.174	-0.521 →**-0.235**

Table 9: Results from Extending Bias Attribute Words to Names Pt.2

Analysis and Discussion We hypothesized that this ablation would help SENT-DEBIAS further reduce bias in sentence representations because with more training examples, there could be better estimation of the bias subspace. We see that this was indeed true for BERT post CoLA and BERT post QNLI. With this ablation, SENT-DEBIAS reduced bias for these two variations in almost all tests conducted. On the other hand, the ablation produced worse results for BERT post SST-2; all tests except for one resulted in increased bias. In the original paper, Liang et al. (2020) cite results from Kiritchenko and Mohammad (2018) to state that it "has

[1]https://spacy.io/usage/linguistic-featuresnamed-entities
[2]https://pypi.org/project/gender-guesser/

been shown that sentiment analysis datasets have labels that correlate with gender information and therefore contain gender bias" (Liang et al., 2020). We hypothesize that debiasing in this ablation may not have worked as well for BERT post SST-2 because the test set was more skewed towards gendered biases connected to people's names. Also, in the original paper, it was found that the variance in effect size was higher for BERT post SST-2 than for BERT post CoLA or BERT post QNLI. High variance could also have contributed to the poor performance of SENT-DEBIAS on BERT post SST-2 in this ablation; more iterations of the ablation experiment may have countered this effect.

7 Conclusion

In this paper, we reproduced the binary gender test results from "Towards Debiasing Sentence Representations." In the original paper, SENT-DEBIAS reduced bias in BERT for the majority of binary tests. However, in our results which can be viewed in Tables 2 and 3, SENT-DEBIAS reduced bias in BERT for just under half of the total number of binary tests. We believe that some of the discrepancy is attributable to high variance in effect size in the experiment and recommend multiple iterations of each evaluation process.

But in a number of key respects, our reproduced results are consistent with those of the original paper. Like the original findings, reduction in bias was most consistent for the tests on the pre-trained BERT encoder. For the pre-trained BERT encoder, SENT-DEBIAS successfully decreased bias in 6 out of the 7 tests, only displaying increased bias on C6B (M/F Terms, Career/Family). Also consistent with the original findings, SENT-DEBIAS performed most poorly in debiasing for BERT post CoLA.

We focused on ablations which we hypothesized would help the performance of SENT-DEBIAS in reducing bias in sentence representations. One of our most interesting findings was that changing the embedding representation from 1 sentence to 2 sentences resulted in reduction in bias across all fine-tuning tasks. This first ablation reduced bias in BERT for the great majority of our binary tests. On the other hand, our second ablation of adding natural sentences to the evaluation data made SENT-DEBIAS reduction perform more poorly. Thirdly, expanding the bias attribute terms in sentences to include more people's names gave great reduction in bias for BERT post CoLA and BERT post QNLI, but not for BERT post SST.

Future work should increase the number of tests run in order to decrease the variance in effect size. We also believe an interesting area of further research would be to use a rigorous and quantitative approach to understand why the SENT-DEBIAS algorithm performed well for BERT fine-tuned on certain tasks but not on others.

Acknowledgments

We would like to thank Professor Chen and Professor Narasimhan for their support in this class, as well as our advisor Chris Sciavolino for his advice throughout the project. Lastly, we thank Paul Liang for giving his time to meet with us.

References

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2016. Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016a. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016b. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 219–226, New York, NY, USA. Association for Computing Machinery.

Aylin Caliskan Islam, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, abs/1608.07187.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. *CoRR*, abs/2007.08100.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *CoRR*, abs/1609.07843.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, page 50–57, New York, NY, USA. Association for Computing Machinery.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461.

Angelina Wang and Olga Russakovsky. 2021. Directional bias amplification. *CoRR*, abs/2102.12594.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural network acceptability judgments. *CoRR*, abs/1805.12471.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA. Association for Computing Machinery.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.