Reducing Gender Bias During Fine-Tuning of a Pre-Trained Language Model

Natasha Flowers and Sam Temlock

12/04/2021

1 Abstract

The issue of machine learning models inheriting biases from the data they are trained on, and subsequently passing that bias along to downstream tasks, is a well-known problem in data science that manifests itself across many applications. Pre-trained language models, such as BERT and DistilBERT, contain biases from the huge amounts of real world data they are trained on, but it is impractical for most users to try to remove this bias from the model itself. We therefore sought a method to reduce bias that is contained to the final fine-tuning stage of usage of a pre-trained model. We chose to focus on using simple counterfactual data augmentation in the fine-tuning task to reduce the gender bias in a pre-trained DistilBERT model for the task of sentiment classification. We observed a small but consistent reduction in gender bias in the model as measured by the difference in average model sentiment on text manipulated to be heavily gendered towards the traditional binary genders of male or female, as well as using the StereoSet bias metrics. In addition, the de-biased model is still able to closely match the performance of the original model on the IMDb movie reviews dataset.

2 Introduction

As the usage of so-called "black-box" machine learning algorithms becomes more and more prevalent across many fields, so too does the problem of bias propagated by these algorithms. The ability to learn how to approximate human decision-making is both a strength and weakness of this approach, as the models that are developed tend to pick up on the underlying human biases reflected in corpora and enforce them in downstream applications in ways that can be difficult for both the developers and users to understand. Many studies have shown that static word embeddings can exhibit social biases that are picked up implicitly from the training data, and more recent research has focused on conducting similar studies on contextualized word embedding models. However, bias in such models can be much harder to measure given that traditional methods used to quantify bias in static embeddings do not translate effectively. It is also difficult for most users to attempt to de-bias these contextualized word embeddings without undertaking extensive retraining or modification of the model.

In this work, we sought a more approachable method to measure and address gender bias in DistilBERT. Rather than directly looking at the contextualized embeddings that can vary given the sentence and trying to derive a measure of bias, we instead examine gender bias in the form of a downstream sentiment analysis task. We take advantage of the fact that gender is directly encoded in a specific subset of English language words, such as pronouns and titles; this gives us the opportunity to mask those specific words and measure the impact on overall sentiment as scored by our model on a corpus of movie reviews. To assess bias, we compare sentiment scores on text reviews that we have manipulated to contain wholly female-associated words and the same set of text reviews that we have manipulated to contain wholly male-associated words. The difference in average sentiment score between these two versions of the reviews gives us a simple metric by which to judge what impact gender has on sentiment scores produced by the model. We show that simple fine-tuning of DistilBERT on text without gender associated words can help debias the sentiment scores while maintaining model accuracy.

3 Background

Inherent social biases have been shown to exist within static language models themselves (Bolukbasi et al., 2016, Caliskan et al., 2017), such as embedding models like word2vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014), and also in their associated downstream systems, such as in automated essay grading (Amorim et al., 2018) and Google Translate (Prates et al., 2020). Furthermore, work around identifying bias in contextual word embeddings (Kurita et al., 2019) and debiasing these models (Zhao et al., 2019) has also been conducted to some extent. However, in most studies, the debiasing techniques fail to produce consistent results across models (May et al., 2019), require retraining of the word embeddings (Zhao et al., 2019), or exhibit a trade-off between debiasing and accuracy (Kaneko and Bollegala, 2021). In our approach, we attempt to use a simple counterfactual data augmentation method to fine tune a pre-trained language model, with the goal of reducing the difference in sentiment scores across the set of reviews with female-associated words vs the same set of reviews but with male-associated words without requiring significant modification of pre-trained embeddings or sacrificing model accuracy.

In keeping with our goal of simplicity and ease of use, we decided to use the DistilBERT model (Sanh et al., 2019) with a linear classification layer on top of the pooled output for fine-tuning. DistilBERT is trained using knowledge distillation (Hinton et al., 2015) to reduce the size of a BERT model (Devlin et al., 2018) by 40%, while still maintaining roughly 97% of NLP performance, as measured on the GLUE benchmark (Wang et al., 2018). By utilizing this smaller model, we were able to iterate quickly on models to examine the impact on bias and accuracy.

For our fine-tuning task, we chose sentiment classification on Stanford's IMDb movie review dataset (Maas et al., 2011), as we felt that movies reflect real-world dynamics and situations, and sentiment scores can act as a representation of bias or preferences. We therefore expect that movie reviews are influenced by societal gender stereotypes, among other factors (we do not expect that this is the main factor driving any one individual review score; we do, however, expect there to be a systemic reflection of gender bias in the reviews overall). The dataset contains 50,000 reviews, originally split into 25,000 test and 25,000 train. The original review scores are out of 10, and reviews that received neutral scores (defined as a score of 5 or 6) are not included. We followed the typical approach of using a binary classification scheme, with all reviews scoring 4 or less classified as negative and reviews scoring 7 or higher classified as positive.

4 Methods

4.1 Gender Vocabulary

In order to reduce the gender associations of the reviews, we needed to identify gender-related words throughout the text. We began with a list of gendered word pairs (ex. 'king' ↔ 'queen') used in the gender neutral GN-GloVe word embedding (Zhao et al., 2018), which we augmented with plurals of existing words in the list as well as several entirely new words. Using this gendered vocabulary, we created a new gender-neutral version of the IMDb movie review dataset by replacing all gendered words with DistilBERT's unknown symbol, '[UNK]'. In order to more easily assess the differences in gender associations, we created datasets that were unambiguously 'male' or 'female' associated: for the 'female' dataset, all male gendered words were replaced with their corresponding female equivalent; and for the 'male' dataset, all female gendered words were replaced with their corresponding male equivalent. In cases where there were multiple corresponding words in our gendered word list (ex. 'his' ↔ 'her', 'hers) , we randomly selected a replacement from among the options. An example of the four variations on a single review is provided in Appendix A.1. Approximately 16% of reviews did not contain any gendered words, and therefore would not change at all between our different datasets; we removed these reviews from subsequent steps after confirming that the distribution of review scores was similar among this subset of reviews as in the total, as they would provide no additional information about gender associations.

Separately, we created versions of all the afore-mentioned datasets using an expanded gender vocabulary that included first names, taken from the names corpus created by Mark Kantrowitz and accessed through the NLTK package (Bird et al., 2009). For the neutral dataset, we simply replaced all names in the lists with the '[UNK]' token in addition to replacing the words from the original gendered vocabulary. For the all-male and all-female datasets, since we did not have direct gender equivalents for most names as we did for terms in the original vocabulary, we chose to replace all female names with the most common male name over the last century ('James', SSA, 2021) to create the all-male dataset and all male names with the most common female name ('Mary', SSA, 2021) to create the all-female dataset.

4.2 Model Implementation

After performing the preprocessing steps described above using the original gender vocabulary on the combined train and test IMDb movie review datasets (originally containing 25,000 reviews each for a total of 50,000 reviews), we were left with a total of 41,830 reviews that contained at least one gendered word, of which 50.2% were classified as positive. When expanding the original vocabulary to include first names, we retained 49,241 reviews that had at least one gendered word, of which 50.1% were classified as positive. We split each of these datasets into train, development, and testing subsets using an 80-10-10 split.

We used the DistilBertforSequenceClassification class with the pre-trained distilbert-base-uncased default configuration from the HuggingFace model library, which consists of a linear layer for classification built on top of the pooled output from the DistilBERT transformer model. For our baseline model comparison, we fine-tuned our classifier using the cross-entropy loss function.

We first fine-tuned the classifier on the original review wording with no gender masking or replacement, and then used it to classify the female and male versions of the test data, giving us an average female sentiment score and an average male sentiment score. We then fine-tuned a separate copy of the same classifier on the gender-neutral review wording (with gendered words replaced by '[UNK]'), and again classified the female and male test data to get average sentiment scores. We compared the average positive sentiment score of the review classification between the male and female datasets in order to check for overall sentiment associations with male vs. female terms.

We experimented with the number of training epochs, learning rate, and whether or not to include first names as part of the gendered vocabulary[1]. We report the results for the first model that was run and evaluated for each variation; we also ran multiple iterations of all models to ensure that reported results were representative of the overall capabilities of the model.

4.3 Metrics

We report two distinct sets of metrics: metrics for model performance, and metrics for model bias.

For model performance, we selected accuracy and cross-entropy loss to compare our results. It should be noted that while we attempt to optimize for model performance, the focus of our work is not to surpass any state-of-the-art results in literature, but rather to get close enough to be a realistic use-case for evaluating reduction in bias.

For the bias metrics, we use the difference in average predicted sentiment of the model on the male vs female datasets, as well as the StereoSet metrics (Nadeem et al., 2020) of Language Modeling Score (LMS), Stereotype Score (SS), and ICAT Score (ICAT), although we are primarily concerned with the SS and ICAT metrics. The LMS score (on a scale of 0-100) is concerned with the model performance on the language modeling task, while the SS score (also on a scale of 0-100), directly attempts to measure the model bias, with a score of 50 representing a completely unbiased model. An ideal model has an LMS score of 100 and SS score of 50. The ICAT score is simply a weighted average of the two scores, computed using the following equation:

$$ICAT = LMS * \frac{min(SS, 100 - SS)}{50}$$

[1] The expanded datasets with first names replaced are denoted as 'original+' and 'neutral+'; the original+ dataset differs from the original dataset in that it includes more of the reviews, as we restrict the dataset to only reviews that have at least one gendered word.

5 Results and Discussion

5.1 Results

We found a small but consistent reduction in the average sentiment difference when training on our neutral dataset (where we replaced gendered words with '[UNK]') as compared to the original dataset, shown in Table 1. The accuracy was similar between models that were otherwise identical but trained on the original vs. neutral dataset. Reducing the learning rate to $1e^{-5}$ from $5e^{-5}$ resulted in a notable improvement in both accuracy and loss for models trained on either dataset. The models trained on the expanded review set (incorporating names into the gendered vocabulary list) showed a slight additional improvement in accuracy, likely due to the increased size of the training set. Our best-performing models were very close to the 92.82% accuracy given by the original DistilBERT paper (Sanh et al., 2019) on this same IMDb sentiment analysis task; note that we modified the originally provided train-test split to use more labeled data for training (80% for training vs. the original 50%), and also restricted the set of reviews to those with gendered words. We chose the best-performing model using the gendered vocabulary without names (shown in bold below) as our primary model for further evaluation and error analysis, as the neutral model had the smallest difference between average male and female sentiment of the variations we tried.

Training Epochs	Learning Rate	Training Dataset	Loss	Accuracy	Avg. Male Sentiment	Avg. Female Sentiment	Sentiment Difference
5	$5e^{-5}$	Original Neutral	0.3141 0.3795	0.8793 0.8716	0.5559 0.5924	0.5430 0.5984	0.0129 -0.0060
10	$5e^{-5}$	Original Neutral	0.6274 0.4842	0.8743 0.8783	0.5475 0.5317	0.5348 0.5354	0.0127 -0.0037
5	$1e^{-5}$	**Original** **Neutral**	**0.2439** **0.2386**	**0.9283** **0.9285**	**0.5332** **0.5398**	**0.5308** **0.5396**	**0.0024** **0.0002**
5	$1e^{-5}$	Original+ Neutral+	0.2142 0.2197	0.9324 0.9330	0.5254 0.5066	0.5183 0.5040	0.0071 0.0025

Table 1: Results of fine-tuning DistilBERT on the original and gender-neutral IMDb review datasets. 'Avg. Male Sentiment' is the average sentiment score for the test dataset with entirely male gendered words, while 'Avg. Female Sentiment' is the average sentiment score for the test dataset with entirely female gendered words. 'Original+' and 'Neutral+' training datasets include names in the list of gendered tokens for replacement. The bolded model was selected as our primary model and used for subsequent analyses.

In addition to the results on the IMDb dataset, we also evaluated both original and neutral models on the StereoSet development dataset, as shown in Table 2. Although StereoSet measures model preferences across gender, race, religion, and profession, given that our focus is on gender bias, we only show the results on the gender data samples. The full results on all metrics can be found in Appendix A.2.

Task	Dataset	Task Count	Model	LMS	SS	ICAT
Intrasentence	Gender	765	Original Neutral	26.11 **40.47**	44.43 **51.67**	23.20 **39.11**
Intersentence	Gender	726	Original Neutral	35.32 34.05	54.20 54.62	32.36 30.91

Table 2: Results of evaluating both original and neutral models on the StereoSet development dataset. The results are from gender subset of the dataset, and excludes scores for the race, religion and profession subsets.

From Table 2, we see that the debiased, neutral model greatly outperforms the original model on the intrasentence task on both the LMS and SS metrics. However, the results for the intersentence task are roughly similar. This may indicate that our debiasing approach is more effective for individual sentences than for compositions of multiple sentences; if so, this could help explain why the reduction in sentiment difference that we observe is quite small overall, as most reviews are longer than a single sentence.

5.2 Error Analysis

We confirm that the neutral model has been at least slightly debiased by examining the confusion matrix in Figure 1. From the matrix, we can see that the original model has a greater proportion of false-positives (and a lesser proportion of male false-negatives) in the male dataset than the female dataset, showing a slight tendency to incorrectly rate male reviews higher than female. The neutral model has a much closer proportion of false-positives and false-negatives for the female dataset vs the male dataset, illustrating the decrease in bias.

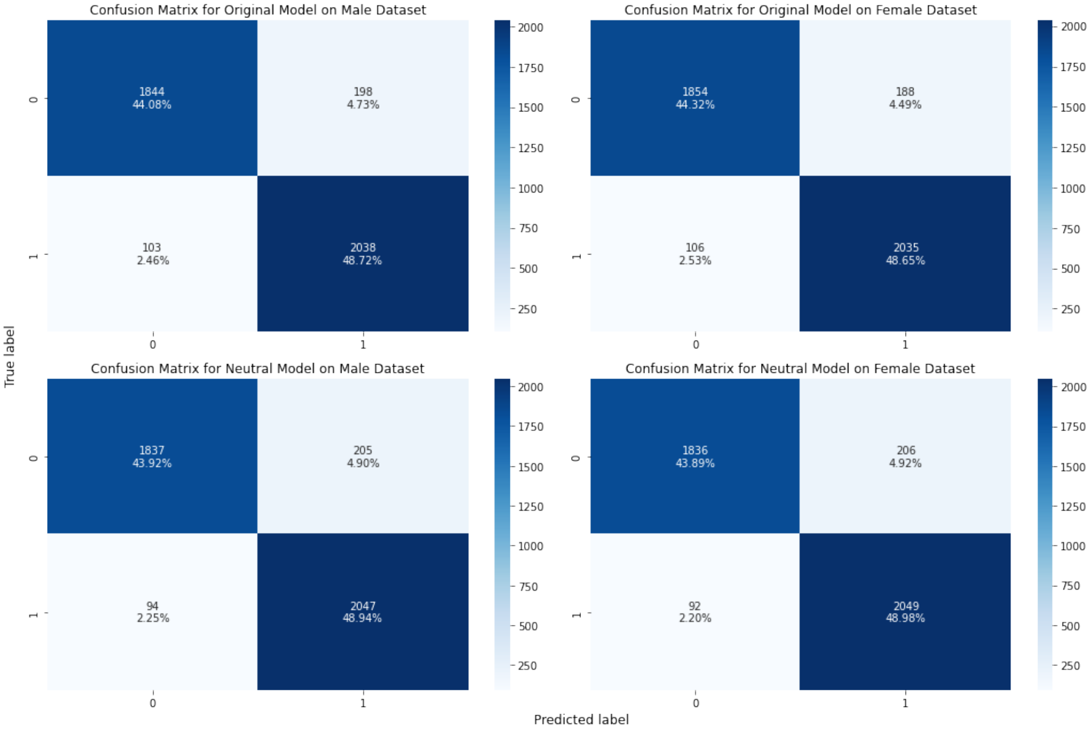
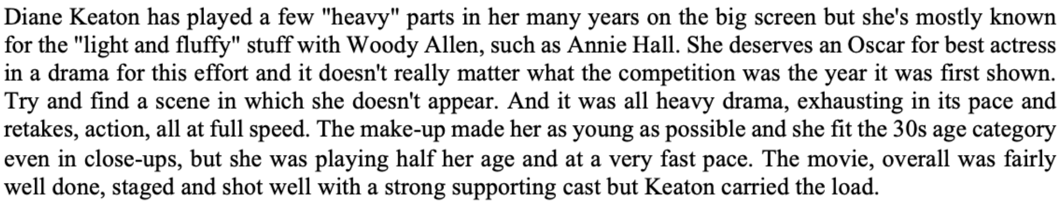


Figure 1: Confusion matrix for the original and neutral model predictions on the male and female datasets (from top left to bottom right): original model on male dataset, original model on female dataset, neutral model on male dataset, neutral model on female dataset.

One example of a review that illustrates this slight negative bias towards female gendered words is shown in Figure 2.



Diane Keaton has played a few "heavy" parts in her many years on the big screen but she's mostly known for the "light and fluffy" stuff with Woody Allen, such as Annie Hall. She deserves an Oscar for best actress in a drama for this effort and it doesn't really matter what the competition was the year it was first shown. Try and find a scene in which she doesn't appear. And it was all heavy drama, exhausting in its pace and retakes, action, all at full speed. The make-up made her as young as possible and she fit the 30s age category even in close-ups, but she was playing half her age and at a very fast pace. The movie, overall was fairly well done, staged and shot well with a strong supporting cast but Keaton carried the load.

Figure 2: Example of the original text of a review with review score of 8 (positive sentiment, label 1). There are 9 female gendered words in the review (and no male gendered words).

This review about a Diane Keaton movie should be classified as positive (receiving a review score of 8), and doesn't contain many words with strong negative connotations that could potentially confuse the model. The original review text is heavily female gendered, with 9 female gendered words and no male gendered words. The original model misclassifies this example's sentiment as negative, with a low probability of positive sentiment at 0.3553. However, when all female gendered words are replaced with their male word counterparts, the model correctly classifies this review as positive, with sentiment of 0.5517. In comparison, the neutral model correctly classifies both male and female versions of this review, with sentiment scores of 0.9838 and 0.9848 respectively. In this case, it appears that the inclusion of either male or female gendered words decrease the model's positive sentiment score on this review, but female gendered words decrease positive sentiment even more than male gendered words. The different versions of this review (with gendered tokens masked and swapped) can be found in Appendix A.1.

Additional analysis of model performance on varying review word lengths is provided in Appendix A.3.

5.3 Limitations

In seeking to characterize gender bias in a simple and understandable way, we have sacrificed some nuance in the identification of gender-related tokens. Our find-and-replace approach to removing or switching the gender of reviews is unable to distinguish context for words that have multiple meanings (some of which may be gendered, some of which may not). This means that there are situations where we may end up masking a word even though it is not used in a gendered context (for example, 'tailor' as a profession has a gender association, but 'tailor' as a verb does not). There is also the reverse situation where we chose not to include a word that sometime has a gendered connotation because it is more commonly used in a neutral sense (for example, 'count' as a title has a gender association but 'count' as a verb does not).

In addition, we did not compare the effect of other tokens that may be correlated with gender identifiers but do not directly identify gender by itself. As a result, there could be confounding tokens that may weaken our ability to effectively create a gender-neutral dataset using this masking approach.

We also note that, given that we are using a pre-trained version of DistilBERT, our maximum embedding length is set at 512 tokens, but this dataset contains movie reviews that exceed that length. We are therefore losing available information about longer movie reviews, and may not be able to classify these reviews as effectively as we can for shorter reviews.

Finally, we recognize that the baseline difference in sentiment between reviews with a majority of female words vs those with a majority of male words is small; this is not surprising to us given that we expect that many other factors have a much stronger influence on review sentiment than underlying gender bias. However, it does make it more challenging to evaluate the utility of the small amount of debiasing we were able to accomplish. The StereoSet evaluation metrics help to confirm that the small shift we observed in our classifications does in fact correspond to reduction of gender bias in other tasks as well.

6 Conclusion

We found that masking gendered words for the fine-tuning of a DistilBERT model did result in a consistent reduction in size of the difference in average sentiment between reviews that had been coerced to use only female-gendered words and reviews that had been coerced to use only male-gendered words. We were able to achieve this without a decrease in the overall model accuracy (in fact, the neutral model had slightly higher accuracy for 3 out of our 4 model variations). Evaluating our model using the StereoSet bias metric also confirmed that the model trained on a neutralized dataset had less gender-associated bias than the model trained on the original dataset. However, the difference was quite small across the board, and it is unclear if it would translate to a meaningful reduction in bias for downstream tasks where the real-world impact is more crucial. Nevertheless, this kind of simple token replacement is a quick and easy way to reduce the gender bias of a model without loss of accuracy, and does not require adjustment of the pre-trained transformer model. Future work could expand beyond the traditional gender binary to look at bias associated with other genders, apply this method to different datasets to see if our findings hold in other domains, and compare the magnitude of bias reduction achieved using this counterfactual data augmentation method to other methods of debiasing.

References

Amorim, E., Cançado, M., & Veloso, A. (2018). Automated essay scoring in the presence of biased ratings. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 229–237.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python* [https://www.nltk.org/book]. O'Reilly Media Inc.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*, 4349–4357.

Caliskan, A., Bryson, J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Kaneko, M., & Bollegala, D. (2021). Debiasing pre-trained contextualised embeddings. *arXiv preprint arXiv:2101.09523*.

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150.

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nadeem, M., Bethke, A., & Reddy, S. (2020). Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Prates, M. O., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: A case study with google translate. *Neural Computing and Applications*, *32*(10), 6363–6381.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

SSA. (2021). Top names over the last 100 years [https://www.ssa.gov/OACT/babynames/decades/century.html, accessed: 2021-12-01].

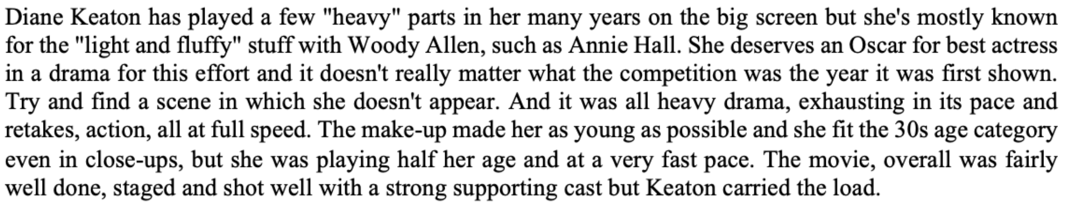
Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K.-W. (2019). Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*.

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K.-W. (2018). Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.

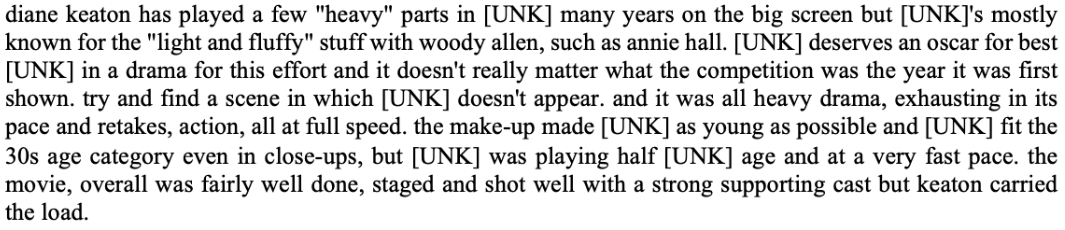
A Appendix

A.1 Example Review with Substitutions



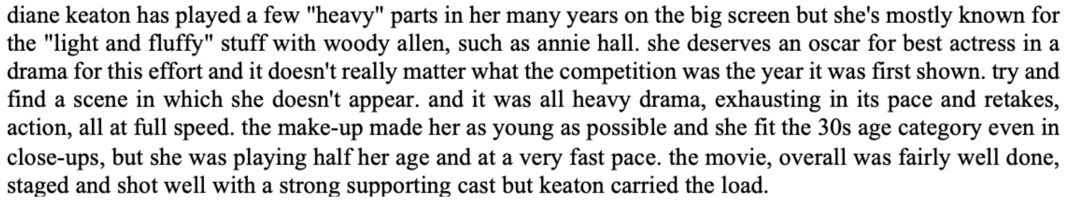
Diane Keaton has played a few "heavy" parts in her many years on the big screen but she's mostly known for the "light and fluffy" stuff with Woody Allen, such as Annie Hall. She deserves an Oscar for best actress in a drama for this effort and it doesn't really matter what the competition was the year it was first shown. Try and find a scene in which she doesn't appear. And it was all heavy drama, exhausting in its pace and retakes, action, all at full speed. The make-up made her as young as possible and she fit the 30s age category even in close-ups, but she was playing half her age and at a very fast pace. The movie, overall was fairly well done, staged and shot well with a strong supporting cast but Keaton carried the load.

Figure 3: Original version of the Diane Keaton movie review. There are 9 female gendered words in the review (and no male gendered words).



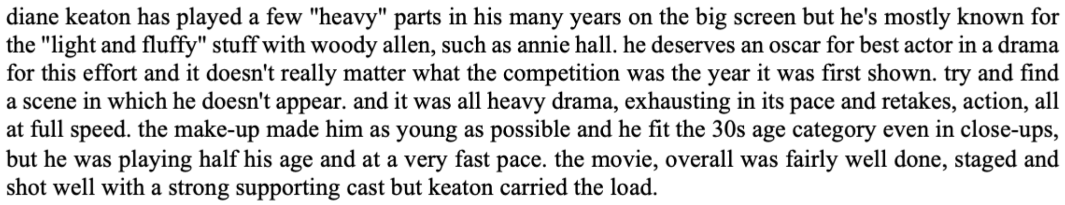
diane keaton has played a few "heavy" parts in [UNK] many years on the big screen but [UNK]'s mostly known for the "light and fluffy" stuff with woody allen, such as annie hall. [UNK] deserves an oscar for best [UNK] in a drama for this effort and it doesn't really matter what the competition was the year it was first shown. try and find a scene in which [UNK] doesn't appear. and it was all heavy drama, exhausting in its pace and retakes, action, all at full speed. the make-up made [UNK] as young as possible and [UNK] fit the 30s age category even in close-ups, but [UNK] was playing half [UNK] age and at a very fast pace. the movie, overall was fairly well done, staged and shot well with a strong supporting cast but keaton carried the load.

Figure 4: Neutral version of the Diane Keaton movie review, where gendered words have been replaced with '[UNK]'.



diane keaton has played a few "heavy" parts in her many years on the big screen but she's mostly known for the "light and fluffy" stuff with woody allen, such as annie hall. she deserves an oscar for best actress in a drama for this effort and it doesn't really matter what the competition was the year it was first shown. try and find a scene in which she doesn't appear. and it was all heavy drama, exhausting in its pace and retakes, action, all at full speed. the make-up made her as young as possible and she fit the 30s age category even in close-ups, but she was playing half her age and at a very fast pace. the movie, overall was fairly well done, staged and shot well with a strong supporting cast but keaton carried the load.

Figure 5: Female gendered version of the Diane Keaton movie review (identical to the original review in this case, since all gendered words were already female).



diane keaton has played a few "heavy" parts in his many years on the big screen but he's mostly known for the "light and fluffy" stuff with woody allen, such as annie hall. he deserves an oscar for best actor in a drama for this effort and it doesn't really matter what the competition was the year it was first shown. try and find a scene in which he doesn't appear. and it was all heavy drama, exhausting in its pace and retakes, action, all at full speed. the make-up made him as young as possible and he fit the 30s age category even in close-ups, but he was playing half his age and at a very fast pace. the movie, overall was fairly well done, staged and shot well with a strong supporting cast but keaton carried the load.

Figure 6: Male gendered version of the Diane Keaton movie review, where female gendered words have been replace with their male word counterparts.

A.2 Full StereoSet Results

Below we provide the full StereoSet results for our original and neutral models. Note that we did not attempt to debias reviews on any characteristic other than gender; interestingly, most other characteristics were still impacted by our debiasing approach.

Task	Dataset	Count	LMS	SS	ICAT
Intrasentence	Gender	765	26.11	44.43	23.20
	Profession	2430	29.64	51.54	28.73
	Race	2886	30.03	50.30	29.85
	Religion	237	24.32	57.47	20.69
Intersentence	Gender	726	35.32	54.20	32.36
	Profession	2481	39.00	53.55	36.23
	Race	2928	39.03	52.65	36.97
	Religion	234	39.71	56.28	34.72
Overall	-	**4229**	**33.87**	**51.75**	**32.68**

Table 3: Original Model Results

Task	Dataset	Count	LMS	SS	ICAT
Intrasentence	Gender	765	40.47	51.67	39.11
	Profession	2430	39.49	48.59	38.37
	Race	2886	46.95	45.32	42.56
	Religion	237	54.00	47.68	51.49
Intersentence	Gender	726	34.05	54.62	30.91
	Profession	2481	39.18	57.15	33.58
	Race	2928	37.19	56.16	32.61
	Religion	234	37.25	61.12	28.97
Overall	-	**4229**	**40.49**	**52.06**	**38.82**

Table 4: Neutral Model Results

A.3 Model Performance against Length of Reviews

Figure 7 shows the performance of each model across the different review lengths. Review lengths are binned in groups (0-100, 101-200, 201-300, 301-400, 401-500, 501-512) to show how the accuracy is affected by the length of the review. It can be seen that the model performances do not differ much aside from the lower performance of the neutral model on the shorter reviews. This may be a result of less information being provided by the review itself, and therefore forcing each model to rely more heavily on its own on understanding of word embeddings/language to make predictions, such as biases/associations with gendered words. As these biases have been reduced in the neutral model, there is less information for the model to use to make accurate predictions.

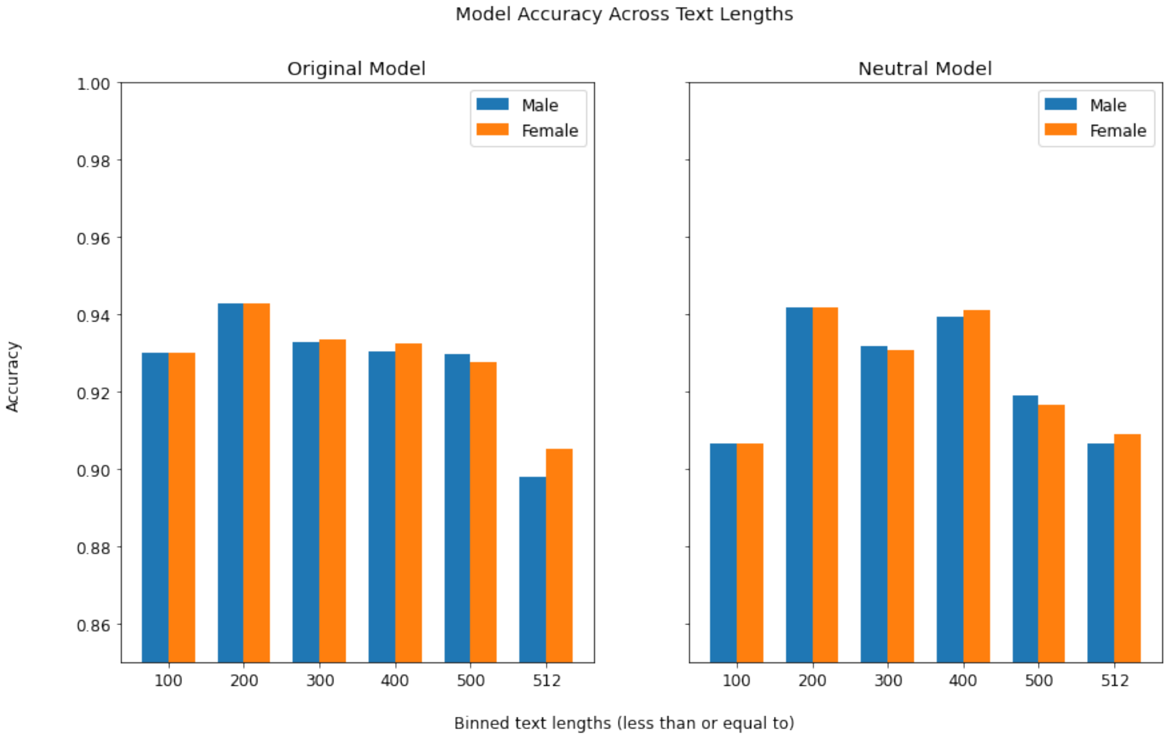


Figure 7: Model accuracy across different lengths of texts. Bins displayed on the x-axis include all reviews that are equal to or smaller than the tick mark itself, and greater than the bin to the left of it. Note that the 512 bin also includes reviews that were longer in length but truncated to 512 when tokenized.