# FairnessNLP: Final Report
# CSE 481N

Erica Eaton, Brandon Ko, Wilson Tang

June 2021

**Abstract**

As sentiment analysis models are more frequently used for sensitive tasks, it is important to ensure the model does not favor some groups of people over others. Previous work introduced different bias mitigation techniques, but without analyzing these techniques under a common evaluation metric, it is difficult to compare different techniques to each other. Our research evaluates models trained with individual as well as multiple bias mitigation techniques. We show how different techniques perform relative to each other and present the most effective combination of bias mitigation techniques.

## 1 Introduction

Sentiment analysis is increasingly used in a wide variety of applications, particularly for understanding users' sentiment towards events, people, or topics on social media. However, sentiment analysis models have been shown to encode or express bias towards certain groups of people based on attributes such as gender, race, or nationality [Kiritchenko and Mohammad, 2018]. Focusing on gender biases (Section 5), we define bias in sentiment analysis models as the predicted sentiment changing when words used to identify a particular group of people, such as gendered words, are changed. For example, if changing the word in a sentence from "he" to "she" also changes the predicted sentiment of the sentence, then there is gender bias.

For many applications, bias in sentiment analysis models could harm specific groups. For example, when using sentiment analysis of tweets for political campaigns [Joyce and Deng, 2017], bias could result in inaccurate sentiment for tweets written about the candidates, based on their gender for example, and affect their poll scores. When gauging employee satisfaction via sentiment analysis [Phoong, 2017], tweets could contain gendered words referencing the author's boss or coworkers, which should not influence the tweet's sentiment. However, not all sentiment analysis applications fall within the scope of our bias definition. When sentiment analysis of social media posts is used to predict depression [Hassan et al., 2017], bias would include bias towards the author of the posts, but our bias definition only captures bias within the raw text of sentences. Although we recognize that incorporating information about the tweet author when training a model can introduce bias, our work focuses on mitigating bias in terms of the content of tweets.

Prior work on bias mitigation only focuses on a specific part of the model training pipeline, such as the dataset, word embeddings, or the model's predictions (Section 3). Different bias mitigation techniques also use different evaluation metrics, making it difficult to compare multiple mitigation techniques. Our research uses a common set of bias evaluation metrics to evaluate and better compare the effectiveness of different bias mitigation techniques (Sections 2 and 3). We also combine multiple bias mitigation techniques to determine if debiasing multiple parts of the training pipeline can result in a more debiased model overall (Sections 4 and 6).[1]

[1] Our code is on GitHub and our project video is on YouTube

## 2 Bias Evaluation Techniques

To evaluate the effectiveness of the bias mitigation methods, we implemented multiple bias evaluation techniques to measure bias in different components of the training pipeline.

### 2.1 Bias in the Dataset

To measure bias in a dataset, we implement co-occurrence and conditional co-occurrence bias, defined by Qian et al. [2019]. Co-occurrence bias compares the occurrence of gender-neutral words with gendered words, where a gender-neutral word is biased, whether positively or negatively, towards a particular gender, if it occurs more often with words from that gender. Ideally, each gender-neutral word should occur about the same number of times with male words as female words, yielding a co-occurrence bias close to zero. However, co-occurrence bias will be large if male words, for example, occur much more often than female words, as gender-neutral words will occur more often with male words [Qian et al., 2019]. Conditional co-occurrence bias addresses this concern by normalizing the co-occurrence counts of male and female words with a gender-neutral word by the total number of male and female words [Qian et al., 2019].

### 2.2 Bias in Model Predictions

In accordance with our definition of bias in sentiment analysis models, we implement four metrics to analyze the model's predictions and compare the predicted sentiment of text as well as the probability the model assigns to that sentiment when changing gendered words in the text.

Metamorphic testing (MT) for model fairness violations, defined by Ma et al. [2020], measures bias in the model's predicted output classes. Given an input sentence, we find perturbations of the sentence using analogy and active mutations [Ma et al., 2020]. Analogy mutations replace a person noun with its gendered counterpart (i.e., "actress" → "actor"). Active mutations add a gendered word in front of a non-gendered person noun (i.e., "person" → "female person"). An MT violation occurs when the predicted sentiment for a perturbation does not match the predicted sentiment for the original sentence [Ma et al., 2020]. The final measurement is the total number of violations in the training set.

Average individual fairness, from Huang et al. [2020], is the average Wasserstein-1 distance between the sentiment probabilities of pairs of sentences. These sentences are formed from the same sentence template, but with different sensitive attributes (i.e., "this man feels happy" and "this woman feels happy"). Similarly, average group fairness is the Wasserstein-1 distance between the sentiment probabilities for all sentence templates and only sentence templates in that subgroup [Huang et al., 2020]. We compute average individual and group fairness on male and female names and noun phrases, where each gender is a subgroup. Also, the original paper uses generated sentences, but we use sentence templates.

The Equity Evaluation Corpus (EEC), created by Kiritchenko and Mohammad [2018], contains sentence templates such as "The conversation with <person object>was <emotional situation word>", where "person object" is a male or female name or noun phrase and "emotional situation word" is an emotion. This corpus was designed to help measure gender and racial bias in models that predict emotions and their intensity in tweets. Kiritchenko and Mohammad [2018] compute the p-value from the paired t-test between a list of sentiment scores when using male vs. female person objects. The null hypothesis that there is on average no difference between the lists of sentiment scores for male vs. female person objects is not accepted if the p-value is below $0.05$. In other words, the higher the p-value, the greater the similarity between sentiment scores for sentences with male vs. female person objects.

## 3 Bias Mitigation Techniques

We classify the bias mitigation techniques into three categories based on the component of the training pipeline they affect: dataset, training, and evaluation level mitigations.

### 3.1 Dataset

A biased dataset can introduce its biases to the model during training, as the model will learn the patterns reflected in the dataset and sometimes amplify those biases [Zhao et al., 2017]. For example, if all instances of the word "he" occur in text with positive sentiment, then the model may learn to assign positive sentiment to text containing "he", resulting in a biased model.

Counterfactual data augmentation (CDA) [Dinan et al., 2020] creates a more balanced dataset. For each sample in the dataset, CDA generates an identical sample but with all gendered words replaced by the corresponding word in the opposite gender [Dinan et al., 2020]. For example, applying CDA to "He is an actor" would yield "She is an actress". We implemented CDA and used the same gendered word lists as Dinan et al. [2020].

AFLite [Bras et al., 2020] removes easy-to-predict samples from the training dataset, generating a new dataset that is much more difficult for the model to learn. The intuition behind this method is to prevent the model from relying on biased patterns. For example, if there are many samples with female words that all have the same sentiment, the model may associate female words with a specific sentiment. AFLite removes these samples, forcing the model to build a better understanding of the underlying patterns in the dataset without being able to exploit features in the easy-to-predict and potentially biased samples, but at a cost to model performance. We implemented AFLite according to [Bras et al., 2020].

### 3.2 Training

Previous bias mitigation techniques at the model level involve debiasing word embeddings by removing the calculated gender subspace [Bolukbasi et al., 2016], which makes the embeddings for gender-neutral words approximately equidistant to embeddings for gendered word pairs, such as the distance from "doctor" to "he" vs. "she". However, Gonen and Goldberg [2019] showed that most gender-neutral words still retain their bias even after applying these bias mitigation techniques, as gender-neutral words biased towards one gender are still grouped together. Thus, Liang et al. [2020] introduce Sent-Debias, which removes the calculated gender subspace in the model's hidden state outputs (sentence representations). The gender subspace is built using sentence representations outputted by the model for sentences that contain gendered words. To remove the gender subspace, each sentence representation output from the model is projected onto the gender subspace and the projection is subtracted from the original sentence representation to effectively remove the bias components. This debiasing technique focuses on biases that occur after the model builds its contextualized representations as opposed to the word embeddings fed into the model. We use the Sent-Debias code provided by Liang et al. [2020] in their GitHub repository.

### 3.3 Evaluation

Certified mitigation, defined by Ma et al. [2020], is a black-box, bias mitigation technique that does not require any dataset modification or model re-training. Given an input sentence, we find perturbations using the same technique as for metamorphic testing. We then feed the original input sentence and all perturbations through the model, and output a smoothed average of all scores. For the weight to put on the original sentence's output, Ma et al. [2020] and our implementation use $0.1$.

## 4 Experiments and Results

For all experiments, we use the BERT for Sequence Classification model from HuggingFace [Devlin et al., 2018] and a Twitter sentiment dataset [Rosenthal et al., 2017]. The Twitter dataset contains about $27,500$ tweets from 2013 to 2017, each labeled as positive or negative [Rosenthal et al., 2017]. Each tweet was labeled by at least five crowdsourced workers who underwent quality control tests [Rosenthal et al., 2017]. We split the Twitter dataset into $70\%$ training, $20\%$ validation, and $10\%$ test. The model for each experiment is trained with the same hyperparameters (batch size = 32, learning rate = $2\mathrm{e}{-5}$, epochs = $3$). These hyperparameters showed promising results in the experiment with no debiasing techniques, and no further hyperparameter tuning was done for any experiment.

The following sections detail our results when applying each bias mitigation technique individually and when combining multiple techniques, fine-tuning BERT on the Twitter dataset with these techniques applied. However, we will not discuss our results on the EEC for each model, as we found all models, including the baseline, perform quite well on the EEC, achieving an average p-value across EEC sentences between $0.95$ and $0.99$, where the maximum average p-value is $1$. These results indicate the model's predicted sentiment for each sentence in the EEC when using male vs. female names and noun phrases are quite similar.

### 4.1 Individual Experiments

Table 1 lists the different datasets used in our experiments. For AFLite, we generated two training datasets. The first is weakly filtered, where only $10\%$ of samples are filtered out, resulting in a dataset with $17,000$ samples. The second dataset is strongly filtered, where $50\%$ of the samples are filtered out, resulting in a dataset with $10,000$ samples. The parameters for AFLite are $k$=300, $m$=4, $n$=17000 for the first dataset and $k$=3500, $m$=8, $n$=10000 for the second dataset. Comparing CDA and AFLite, the dataset debiasing techniques, CDA strongly outperforms AFLite in reducing co-occurrence bias and conditional co-occurrence bias. Since CDA is balancing the number of male and female words in the dataset, each gender-neutral word in the dataset will occur about the same number of times with male words as female words, yielding very low (conditional) co-occurrence bias. However, applying AFLite with the weak or strong filter, did not change the (conditional) co-occurrence bias much, indicating the easy-to-predict samples AFLite filters out are generally not biased towards a particular gender since the resulting dataset does not seem less biased.

| Dataset | Tweets | Co-Occurrence Bias (↓) | Cond. Co-Occurrence Bias (↓) |
|---|---|---|---|
| Original | 19166 | 0.982 | 0.556 |
| With CDA | 24326 | 0.017 | 0.020 |
| With Weak AFLite | 17000 | 0.989 | 0.568 |
| With Strong AFLite | 10000 | 0.935 | 0.507 |

Table 1: Dataset size and bias when applying CDA, AFLite, or no bias mitigation techniques

Results when using each bias mitigation technique individually are in Figure 1. Applying CDA improved 3 of the fairness metrics (average individual and group fairness on names and noun phrases), reducing average individual and group fairness on names to nearly zero. For MT violations, CDA greatly reduced all violations caused by "male", "female", "masculine", and "feminine". CDA also slightly increased model performance, likely because of the $27\%$ increase in dataset size from CDA, since the model has more data to train on. Sent-Debias also improved on most fairness metrics, reducing MT violations the most and only making violations on "masculine", "feminine", and "gods". All violations involving "male" and "female" were resolved, likely because "male" and "female" were in the gender subspace but "masculine" and "feminine" were not. However, Sent-Debias decreased model performance, as did weak and strong AFLite. Also, weak AFLite did not improve on the fairness metrics. Strong AFLite increased MT violations, but improved on all fairness metrics. Nevertheless, based on the primary metric (MT violations), neither AFLite model significantly decreased bias. Certified mitigation (CM) increased model performance the most out of all models, while slightly reducing MT violations and improving on all fairness metrics. Intuitively, this increase in model performance could be because CM outputs a smoothed average of slight mutations of the sentence, which is less likely to be inaccurate, and CM corrects biases towards individual words.

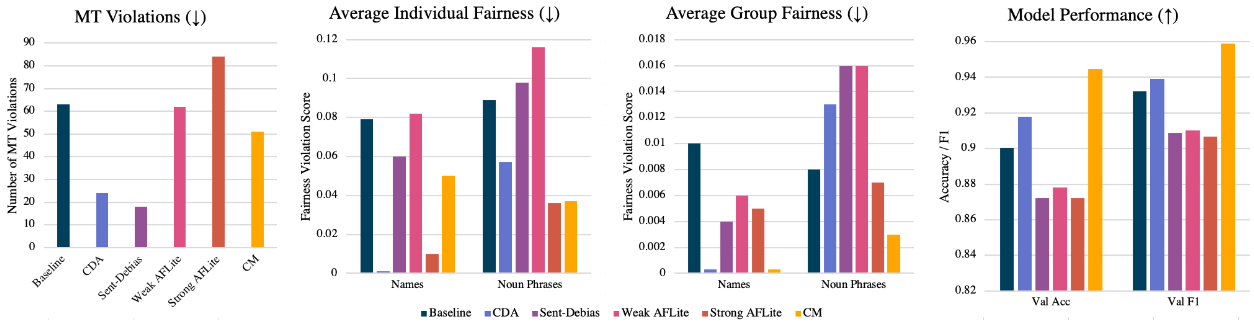

Figure 1: MT Violations, fairness metrics, and model performance for each individual bias mitigation technique

### 4.2 Combinations

Combining Sent-Debias and CDA resulted in a model that performed the best on the MT violations metric. Sent-Debias eliminated all violations caused by "male" and "female", but still had violations caused by "masculine" and "feminine". CDA reduced all violations caused by "male", "female", "masculine", and "feminine". Thus, combining these two techniques drastically decreased violations caused by "male", "female", "masculine", and "feminine", resulting in the fewest total MT violations.

When combining AFLite with other bias mitigation techniques, we observe the other techniques are simply correcting for AFLite's mistakes. AFLite individually did not meaningfully improve the number of MT violations, as described above. When other techniques were added on top of AFLite, the number of MT violations improved, but their counterparts without AFLite still performed better. Thus, we conclude AFLite is not a successful bias mitigation technique for our use case.

For CM, we observed that MT violations increased when combining CM with Sent-Debias or Sent-Debias and CDA, but decreased when combining CM with CDA or strong AFLite. Additionally, in all experiments where CM was part of the debiasing pipeline, all four average individual and group fairness metrics as well as the validation accuracy and F1 scores improved from the baseline.

### 4.3 Best Results

Table 2 lists the debiasing pipeline that yielded the lowest value for each bias evaluation technique and the highest values for validation accuracy and F1. From Table 2, we observe that combining CDA and Sent-Debias results in the fewest MT violations, while CM and Sent-Debias gives the best model performance. Although different models performed best for average individual/group fairness, each model includes CM, which individually improves on all fairness metrics. The following debiasing pipeline had the best performance overall for debiasing a sentiment analysis model: CDA→Sent-Debias→CM. CDA debiases the dataset, Sent-Debias the model, and CM the model predictions. CDA and Sent-Debias together greatly reduce MT violations, while CM improves average individual and group fairness on names and noun phrases and increases model accuracy and F1 score.

## 5 Limitations and Future Work

A key limitation in debiasing with respect to gender is the presence of gender-neutral words, such as using "they" and "them" in place of the male gendered words "he" and "him" or the female gendered words "she" and "her". None of the techniques explored in this paper is quite able to handle situations where the model can be biased against gender-neutral words. For example, adding the "neuter" (neutral) gender into the metamorphic testing process caused the number of MT violations to increase substantially. Future work for gender debiasing should include the representation of gender-neutral words.

| Model | MT V(↓) | AIF N (↓) | AGF N (↓) | AIF NPs (↓) | AGF NPs (↓) | V Acc (↑) | V F1 (↑) |
|---|---|---|---|---|---|---|---|
| Baseline | 63 | 0.079 | 0.010 | 0.089 | 0.008 | 0.900 | 0.932 |
| CDA + Sent-Debias | **16** | 0.010 | 0.005 | 0.102 | 0.011 | 0.910 | 0.935 |
| Strong AFLite + Sent-Debias + CM | 98 | **0** | **0** | 0.025 | 0.004 | 0.911 | 0.934 |
| Weak AFLite + Sent-Debias + CM | 73 | 0.021 | 7.35e-4 | **0.009** | 0.002 | 0.915 | 0.937 |
| Strong AFLite + CDA + CM | 57 | 0.032 | 0.004 | 0.049 | **0.0002** | 0.903 | 0.927 |
| Sent-Debias + CM | 20 | 0.019 | 0.001 | 0.042 | 0.002 | **0.946** | **0.960** |
| **CDA + Sent-Debias + CM** | 26 | 0.014 | 0.005 | 0.030 | 0.002 | 0.915 | 0.937 |

Table 2: Best performing models in terms of validation accuracy (V Acc), validation F1 (V F1), and each bias evaluation metric. MT V is metamorphic testing violations. AIF N or NPs is average individual fairness for names or noun phrases, respectively. AGF N or NPs is average group fairness for names or noun phrases, respectively.

Another limitation is the reliance of CDA and Sent-Debias on a pre-defined list of gendered words. These pre-defined lists can contain bias in what we consider gendered words, such as missing certain gendered words. Future work can explore changing these gendered lists as a form of hyperparameter tuning. Certified mitigation attempts to address this limitation by using a knowledge graph. Future work could also attempt to apply knowledge graph ideas to CDA and Sent-Debias.

Another area for future work is extending the bias mitigation and evaluation techniques we used to other sensitive groups besides gender, such as race or nationality. All techniques in our bias evaluation and mitigation pipeline can be extended to other sensitive groups, whether by changing sentence templates or hyperparameters for the sensitive attribute.

## 6 Conclusion

Prior work on bias mitigation focused on mitigating bias in one part of the training pipeline, with each technique evaluated under a different bias evaluation metric. By evaluating bias mitigation techniques for multiple parts of the training pipeline under a common set of evaluation metrics, we were able to compare the effectiveness of these techniques and create a debiasing pipeline for sentiment analysis models that shows improvement across multiple bias evaluation metrics and increases overall accuracy. Our pipeline mitigates gender bias in multiple parts of the training process: CDA debiases the dataset, Sent-Debias debiases the model's hidden states, and certified mitigation debiases the model's predictions. We leave it to future work to extend this pipeline for gender-neutral words and other bias domains, such as race and religion.

## Acknowledgments

We would like to thank Xuhui Zhou, Noah Smith, Jungo Kasai, Ofir Press, and the students in CSE 481 NLP Capstone for their feedback and advice throughout this project.

## References

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

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. Queens are powerful too: Mitigating gender bias in dialogue generation, 2020.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them, 2019.

Anees Ul Hassan, Jamil Hussain, Musarrat Hussain, Muhammad Sadiq, and Sungyoung Lee. Sentiment analysis of social networking sites (sns) data using machine learning approach for the measurement of depression. In *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 138–140, 2017. doi: 10.1109/ICTC.2017.8190959.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.7. URL https://www.aclweb.org/anthology/2020.findings-emnlp.7.

Brandon Joyce and Jing Deng. Sentiment analysis of tweets for the 2016 us presidential election. In *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*, pages 1–4, 2017. doi: 10.1109/URTC.2017.8284176.

Svetlana Kiritchenko and Saif M. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems, 2018.

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

Pingchuan Ma, Shuai Wang, and Jin Liu. Metamorphic testing and certified mitigation of fairness violations in nlp models. In *IJCAI*, 2020.

Seuk Wai Phoong. Social media sentiment analysis on employment in malaysia. 12 2017.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. Reducing gender bias in word-level language models with a gender-equalizing loss function, 2019.

Sara Rosenthal, Noura Farra, and Preslav Nakov. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2088. URL https://www.aclweb.org/anthology/S17-2088.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1323. URL https://www.aclweb.org/anthology/D17-1323.

## A Results for Experiments with Individual Bias Mitigation Techniques

| Model | MT V(↓) | AIF N (↓) | AGF N (↓) | AIF NPs (↓) | AGF NPs (↓) | V Acc (↑) | V F1 (↑) |
|---|---|---|---|---|---|---|---|
| Baseline | 63 | 0.079 | 0.010 | 0.089 | 0.008 | 0.900 | 0.932 |
| CDA | 24 | 3.86e-5 | 4.90e-5 | 0.057 | 0.013 | 0.918 | 0.939 |
| Sent-Debias | 18 | 0.060 | 0.004 | 0.098 | 0.007 | 0.872 | 0.909 |
| Weak AFLite | 62 | 0.082 | 0.006 | 0.116 | 0.016 | 0.878 | 0.910 |
| Strong AFLite | 84 | 0.010 | 0.005 | 0.036 | 0.007 | 0.872 | 0.907 |
| CM | 51 | 0.050 | 1.47e-4 | 0.037 | 0.003 | 0.945 | 0.960 |

Table 3: Model performance and prediction bias when applying each bias mitigation technique individually. MT V is metamorphic testing violations. AIF N or NPs is average individual fairness for names or noun phrases, respectively. AGF N or NPs is average group fairness for names or noun phrases, respectively. V Acc and V F1 are validation accuracy and F1 score, respectively.