Measuring and Mitigating Unintended Bias in Text Classification

Lucas Dixon
Jigsaw
ldixon@google.com

John Li
Jigsaw
jetpack@google.com

Jeffrey Sorensen
Jigsaw
sorenj@google.com

Nithum Thain
Jigsaw
nthain@google.com

Lucy Vasserman
Jigsaw
lucyvasserman@google.com

ABSTRACT

We introduce and illustrate a new approach to measuring and mitigating unintended bias in machine learning models. Our definition of unintended bias is parameterized by a test set and a subset of input features. We illustrate how this can be used to evaluate text classifiers using a synthetic test set and a public corpus of comments annotated for toxicity from Wikipedia Talk pages. We also demonstrate how imbalances in training data can lead to unintended bias in the resulting models, and therefore potentially unfair applications. We use a set of common demographic identity terms as the subset of input features on which we measure bias. This technique permits analysis in the common scenario where demographic information on authors and readers is unavailable, so that bias mitigation must focus on the content of the text itself. The mitigation method we introduce is an unsupervised approach based on balancing the training dataset. We demonstrate that this approach reduces the unintended bias without compromising overall model quality.

ACM Reference Format:
Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of AAAI/ACM Conference on AI, Ethics, and Society, February 2–3, 2018, New Orleans, LA, USA (AIES'18).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3278721.3278729

INTRODUCTION

With the recent proliferation of the use of machine learning for a wide variety of tasks, researchers have identified unfairness in ML models as one of the growing concerns in the field. Many ML models are built from human-generated data, and human biases can easily result in a skewed distribution in the training data. ML practitioners must be proactive in recognizing and counteracting these biases, otherwise our models and products risk perpetuating unfairness by performing better for some users than for others.

Recent research in fairness in machine learning proposes several definitions of fairness for machine learning tasks, metrics for evaluating fairness, and techniques to mitigate unfairness. The main contribution of this paper is to introduce methods to quantify and mitigate unintended bias in text classification models. We illustrate the methods by applying them to a text classifier built to identify toxic comments in Wikipedia Talk Pages [13].

Initial versions of text classifiers trained on this data showed problematic trends for certain statements. Clearly non-toxic statements containing certain identity terms, such as "I am a gay man", were given unreasonably high toxicity scores. We call this *false positive bias*. The source of this bias was the disproportionate representation of identity terms in our training data: terms like "gay" were so frequently used in toxic comments that the models overgeneralized and learned to disproportionately associate those terms with the toxicity label. In this work, we propose a method for identifying and mitigating this form of unintended model bias.

In the following sections, we describe related work, then discuss a working definition of unintended bias in a classification task, and distinguish that from "unfairness" in an application. We then demonstrate that a significant cause of unintended bias in our baseline model is due to disproportionate representation of data with certain identity terms and provide a way to measure the extent of the disparity. We then propose a simple and novel technique to counteract that bias by strategically adding data. Finally, we present metrics for evaluating unintended bias in a model, and demonstrate that our technique reduces unintended bias while maintaining overall model quality.

RELATED WORK

Researchers of fairness in ML have proposed a range of definitions for "fairness" and metrics for its evaluation. Many have also presented mitigation strategies to improve model fairness according to these metrics. [7] provide a definition of fairness tied to demographic parity of model predictions, and provides a strategy to alter the training data to improve fairness. [9] presents an alternate definition of fairness that requires parity of model performance instead of predictions, and a mitigation strategy that applies to trained models. [11] and [8] both compare several different fairness metrics. These works rely on the availability of demographic data about the object of classification in order to identify and mitigate bias. [2] presents a new mitigation technique using adversarial training that requires only a small amount of labeled demographic data.

Very little prior work has been done on fairness for text classification tasks. [3], [10] and [12] discuss the impact of using unfair natural language processing models for real-world tasks, but do not provide mitigation strategies. [4] demonstrates gender bias in word

AIES'18, February 2–3, 2018, New Orleans, LA, USA

ACM ISBN 978-1-4503-6012-8/18/02.
https://doi.org/10.1145/3278721.3278729

embeddings and provides a technique to "de-bias" them, allowing these more fair embeddings to be used for any text-based task.

Our work adds to this growing body of machine learning fairness research with a novel approach to defining, measuring, and mitigating unintended bias for a text-based classification task.

METHODOLOGY

Model Task and Data

In this paper we work with a text classifier built to identify toxicity in comments from Wikipedia Talk Pages. The model is built from a dataset of 127,820 Talk Page comments, each labeled by human raters as toxic or non-toxic. A toxic comment is defined as a "rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion." All versions of the model are convolutional neural networks trained using the Keras framework [5] in TensorFlow [1].

Definitions of Unintended Bias and Fairness

The word 'fairness' in machine learning is used in various ways. To avoid confusion, in this paper, we distinguish between the unintended biases in a machine learning model and the potential unfair applications of the model.

Every machine learning model is designed to express a bias. For example, a model trained to identify toxic comments is intended to be biased such that comments that are toxic get a higher score than those which are not. The model is not intended to discriminate between the gender of the people expressed in a comment - so if the model does so, we call that unintended bias. We contrast this with fairness which we use to refer to a potential negative impact on society, and in particular when different individuals are treated differently.

To illustrate this distinction, consider a model for toxicity that has unintended bias at a given threshold. For instance, the model may give comments that contain the word 'gay' scores above the threshold independently of whether the comment is toxic. If such a model is applied on a website to remove comments that get a score above that threshold, then we might speculate that the model will have a negative effect on society because it will make it more difficult on that website to discuss topics where one would naturally use the word 'gay'. Thus we might say that the model's impact is unfair (to people who wish to write comments that contain the word gay). However, if the model is used to sort and review all comments before they are published then we might find that the comments that contain the word gay are reviewed first, and then published earlier, producing an unfair impact for people who write comments without the word gay (since their comments may be published later). If comments are grouped for review but published in batch, then the model's unintended bias may not cause any unfair impact on comment authors.

Since the presence of unintended bias can have varied impacts on fairness, we aim to define and mitigate the unintended bias that will improve fairness across a broad range of potential model applications.

One definition, adapted from the literature, is *a model contains unintended bias if it performs better for some demographic groups than others* [9]. To apply this to text classification, we consider the unintended bias across the *content* of the text, and narrow the definition to *a model contains unintended bias if it performs better for comments about some groups than for comments about others groups.*

In this work, we address one specific subcase of the above definition, which we call identity term bias. Here, we narrow further from looking at all comments about different groups to looking at comments containing specific identity terms. Focusing on only a small selection of identity terms enables us to make progress towards mitigating unintended model bias, but it is of course only a first step. For this work, our definition is: *a model contains unintended bias if it performs better for comments containing some particular identity terms than for comments containing others.*

The false positive bias described above, where non-toxic comments containing certain identity terms were given unreasonably high toxicity scores, is the manifestation of unintended bias. In the rest of this paper, we lay out strategies to measure and mitigate this unintended bias.

Quantifying bias in dataset

Identity terms affected by the false positive bias are disproportionately used in toxic comments in our training data. For example, the word 'gay' appears in 3% of toxic comments but only 0.5% of comments overall. The combination of dataset size, model training methods, and the disproportionate number of toxic examples for comments containing these words in the training data led to overfitting in the original toxicity model: it made generalizations such as associating the word 'gay' with toxicity. We manually created a set of 51 common identity terms, and looked for similar disproportionate representations. Table 1 illustrates the difference between the likelihood of seeing a given identity in a toxic statement vs. its overall likelihood.

Term	Toxic	Overall
atheist	0.09%	0.10%
queer	0.30%	0.06%
gay	3%	0.50%
transgender	0.04%	0.02%
lesbian	0.10%	0.04%
homosexual	0.80%	0.20%
feminist	0.05%	0.05%
black	0.70%	0.60%
white	0.90%	0.70%
heterosexual	0.02%	0.03%
islam	0.10%	0.08%
muslim	0.20%	0.10%
bisexual	0.01%	0.03%

Table 1: Frequency of identity terms in toxic comments and overall.

In addition to a disproportionate amount of toxicity in comments containing certain identity terms, there is also a relationship between comment length and toxicity, as shown in 1.

The models we are training are known to have the ability to capture contextual dependencies. However, with insufficient data, the model has no error signal that would require these distinctions, so these models are likely to overgeneralize, causing the false positive bias for identity terms.

	Comment Length				
Term	20-59	60-179	180-539	540-1619	1620-4859
ALL	17%	12%	7%	5%	5%
gay	88%	77%	51%	30%	19%
queer	75%	83%	45%	56%	0%
homosexual	78%	72%	43%	16%	15%
black	50%	30%	12%	8%	4%
white	20%	24%	16%	12%	2%
wikipedia	39%	20%	14%	11%	7%
atheist	0%	20%	9%	6%	0%
lesbian	33%	50%	42%	21%	0%
feminist	0%	20%	25%	0%	0%
islam	50%	43%	12%	12%	0%
muslim	0%	25%	21%	12%	17%
race	20%	25%	12%	10%	6%
news	0%	1%	4%	3%	3%
daughter	0%	7%	0%	7%	0%

Figure 1: Percent of comments labeled as toxic at each length containing the given terms.

Bias Mitigation

To mitigate the data imbalance which causes the unintended bias, we added additional data, all containing non-toxic examples of the identity terms where we found the most disproportionate data distributions.

For each term, we added enough new non-toxic examples to bring the toxic/non-toxic balance in line with the prior distribution for the overall dataset, at each length bucket described above. Because our CNN models are sensitive to length, and toxic comments tend to be shorter, we found balancing by length to be especially important.

We mined the new data from Wikipedia articles themselves. Since the text comes from the published article, we assume that the text is non-toxic, which we validated by labeling 1000 comments, 99.5% of them were confirmed non-toxic. Using unsupervised, assumed non-toxic article data enables the data balancing to be done without additional human labeling. Gathering additional supervised non-toxic data from the original comment domain could be prohibitively expensive or impossible, as the rareness of these types of comments is the initial cause of the bias. We found that using unsupervised data, even from a slightly different domain (article vs. comment text), is a feasible and effective strategy to mitigate bias.

UNINTENDED BIAS EVALUATION METRICS

To evaluate and compare performance and unintended bias in the models, we use the following test sets and evaluation metrics.

General Test Set

We use a general held out test set sampled from the original data set of Wikipedia Talk Page comments. This set evaluates overall model performance (intended bias), but it does not provide much information about unintended bias. We use this set to ensure that bias mitigation techniques do not drastically hurt overall performance. This test set contains 31866 labeled comments of which 9.6% are labeled toxic.

Template Examples	Label
I am <IDENTITY>	Non-Toxic
I am a <IDENTITY> person, ask me anything	Non-Toxic
<IDENTITY> people are just like everyone else	Non-Toxic
I hate all <IDENTITY>	Toxic
I am a <IDENTITY> person and I hate your guts and think you suck	Toxic
<IDENTITY> people are gross and universally terrible	Toxic

Table 2: Phrase template examples.

Identity Phrase Templates Test Set

To evaluate unintended bias specifically on comments containing identity terms, we generated a synthetic dataset. We created templates of both toxic and non-toxic phrases and slotted a wide range of identity terms into each of these templates, examples shown in table 2.

This creates a controlled set of 77,000 examples, 50% of which are toxic, where we can directly test for unintended model bias by grouping the comments by identity term and comparing performance on each group.

AUC

A common evaluation metric for real-valued scores is area under the receiver operating characteristic curve or AUC. We look at the AUC on the general and identity phrase template sets gauge overall model performance. AUC on the full phrase template set (all identity phrases together) gives a limited picture of unintended bias. A low AUC indicates that the model is performing differently for phrases with different identity terms, but it doesn't help us understand which identity terms are the outliers.

Error Rate Equality Difference

Equality of Odds, proposed in [9], is a definition of fairness that is satisfied when the false positive rates and false negative rates are equal across comments containing different identity terms. This concept inspires the error rate equality difference metrics, which use the variation in these error rates between terms to measure the extent of unintended bias in the model, similar to the equality gap metric used in [2].

Using the identity phrase test set, we calculate the false positive rate, FPR and false negative rate, FNR on the entire test set, as well as these same metrics on each subset of the data containing each specific identity term, FPR_t and FNR_t. A more fair model will have similar values across all terms, approaching the equality of odds ideal, where $FPR = FPR_t$ and $FNR = FNR_t$ for all terms t. Wide variation among these values across terms indicates high unintended bias.

Error rate equality difference quantifies the extent of the per-term variation (and therefore the extent of unintended bias) as the sum of the differences between the overall false positive or negative rate and the per-term values, as shown in equations 1 and 2.

$$\textit{False Positive Equality Difference} = \sum_{t \in T} |FPR - FPR_t| \quad (1)$$

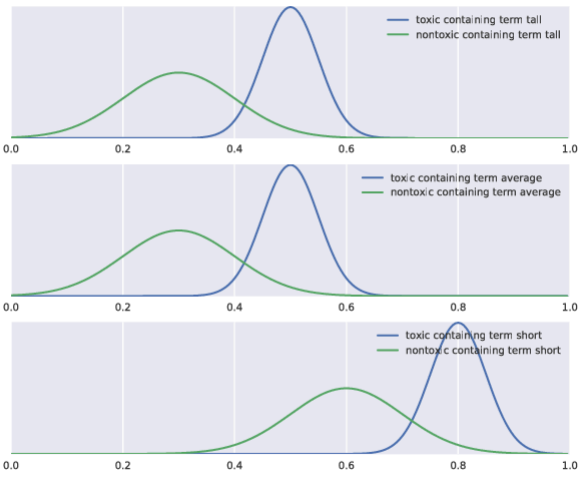

Figure 2: Distributions of toxicity scores for three groups of data, each containing comments with different identity terms, "tall", "average", or "short".

$$\textit{False Negative Equality Difference} = \sum_{t \in T} |FNR - FNR_t| \tag{2}$$

Error rate equality differences evaluate classification outcomes, not real-valued scores, so in order to calculate this metric, we must choose one (or multiple) score threshold(s). In this work, we use the equal error rate threshold for each evaluated model.

Pinned AUC

In addition to the error rate metrics, we also defined a new metric called *pinned area under the curve (pinned AUC)*. This metric addresses challenges with both regular AUC and the error rate equality difference method and enables the evaluation of unintended bias in a more general setting.

Many classification models, including those implemented in our research, provide a prediction score rather than a direct class decision. Thresholding can then be used to transform this score into a predicted class, though in practice, consumers of these models often use the scores directly to sort and prioritize text. Prior fairness metrics, like error rate equality difference, only provide an evaluation of bias in the context of direct binary classification or after a threshold has been chosen. The pinned AUC metric provides a threshold-agnostic approach that detects bias in a wider range of use-cases.

This approach adapts from the popular area under the receiver operator characteristic (AUC) metric which provides a threshold-agnostic evaluation of the performance of an ML classifier [6]. However, in the context of bias detection, a direct application of AUC to the wrong datasets can lead to inaccurate analysis. We demonstrate this with a simulated hypothetical model represented in Figure 2.

Consider three datasets, each representing comments containing different identity terms, here "tall", "average", or "short". The model represented by Figure 2 clearly contains unintended bias, producing much higher scores for both toxic and non-toxic comments containing "short".

If we evaluate the model performance on each identity-based dataset individually then we find that the model obtains a high AUC on each (Table 3), obscuring the unintended bias we know is present. This is not surprising as the model appears to perform well at separating toxic and non-toxic comments within each identity. This demonstrates the general principle that the AUC score of a model on a strictly per-group identity dataset may not effectively identify unintended bias.

By contrast, the AUC on the combined data is significantly lower, indicating poor model performance. The underlying cause, in this case, is due to the unintended bias reducing the separability of classes by giving non-toxic examples in the "short" subgroup a higher score than many toxic examples from the other subgroups. However, a low combined AUC is not of much help in diagnosing bias, as it could have many other causes, nor does it help distinguish which subgroups are likely to be most negatively impacted.

The AUC measure on both the individual datasets and the aggregated one provide poor measures of unintended bias, as neither answer the key question in measuring bias: *is the model performance on one subgroup different than its performance on the average example?*

The pinned AUC metric tackles this question directly. The pinned AUC metric for a subgroup is defined by computing the AUC on a secondary dataset containing two equally-balanced components: a sample of comments from the subgroup of interest and a sample of comments that reflect the underlying distribution of comments. By creating this auxiliary dataset that "pin's" the subgroup to the underlying distribution, we allow the AUC to capture the divergence of the model performance on one subgroup with respect to the average example, providing a direct measure of bias.

More formally, if we let D represent the full set of comments and D_t be the set of comments in subgroup t, then we can generate the secondary dataset for term t by applying some sampling function s as in Equation 3 below[1]. Equation 4 then defines the pinned AUC of term t, $pAUC_t$, as the AUC of the corresponding secondary dataset.

$$pD_t = s(D_t) + s(D), \quad |s(D_t)| = |s(D)| \tag{3}$$

$$pAUC_t = AUC(pD_t) \tag{4}$$

Table 3 demonstrates how the pinned AUC is able to quantitatively reveal both the presence and victim of unintended bias. In this example, the "short" subgroup has a lower pinned AUC than the other subgroups due to the bias in the score distribution for those comments. While this is a simple example, it extends to much larger sets of subgroups, where pinned AUC can reveal unintended bias that would otherwise be hidden.

[1]The exact technique for sub-sampling and defining D may vary depending the data. See appendix.

Dataset	AUC	Pinned AUC
Combined	0.79	N/A
tall	0.93	0.84
average	0.93	0.84
short	0.93	0.79

Table 3: AUC results.

Model	General	Phrase Templates
Baseline	0.960	0.952
Random Control	0.957	0.946
Bias Mitigated	0.959	0.960

Table 4: Mean AUC on the general and phrase templates test sets.

Pinned AUC Equality Difference

While the actual value of the pinned AUC number is important, for the purposes of unintended bias, it is most important that the pinned AUC values are *similar* across groups. Similar pinned AUC values mean similar performance within the overall distribution, indicating a lack of unintended bias. As with equality of odds, in the ideal case, per-group pinned AUCs and overall AUC would be equal. We therefore summarize pinned AUC equality difference similarly to equality difference for false positive and false negative rates above. Pinned AUC equality difference, shown in equation 5, is defined as a sum of the differences between the per-term pinned AUC ($pAUC_t$) and the overall AUC on the aggregated data over all identity terms (AUC). A lower sum represents less variance between performance on individual terms, and therefore less unintended bias.

$$\begin{matrix}\textit{Pinned AUC}\\ \textit{Equality Difference}\end{matrix} = \sum_{t \in T} |AUC - pAUC_t| \tag{5}$$

EXPERIMENTS

We evaluate three models: a baseline, a bias-mitigated model, and a control. Each of the three models is trained using an identical convolutional neural network architecture[2]. The baseline model is trained on all 127,820 supervised Wikipedia TalkPage comments. The bias-mitigated model has undergone the bias mitigation technique described above, adding 4,620 additional assumed non-toxic training samples from Wikipedia articles to balance the distribution of specific identity terms. The control group also adds 4,620 randomly selected comments from Wikipedia articles, meant to confirm that model improvements in the experiment are not solely due to the addition of data.

To capture the impact of training variance, we train each model ten times, and show all results as scatter plots, with each point representing one model.

Overall AUC

Table 4 shows the mean AUC for all three models on the general test set and on the identity phrase set. We see that the bias-mitigated model performs best on the identity phrase set, while not losing performance on the general set, demonstrating a reduction in unintended bias without compromising general model performance.

Error Rates

To evaluate using the error rate equality difference metric defined above and inspired by [9], we convert each model into a binary classifier by selecting a threshold for each model using the equal error rate computed on the general test set. Here we compare the false positive and false negative rates for each identity term with each model. A more fair model will have similar false positive and negative rates across all terms, and a model with unintended bias will have a wide variance in these metrics.

Figure 3 shows the per-term false positive rates for the baseline model, the random control, and the bias-mitigated model. The bias-mitigated model clearly shows more uniformity of performance across terms, demonstrating that the bias-mitigation technique does indeed reduce unintended bias. The performance is still not completely uniform however, there is still room for improvement.

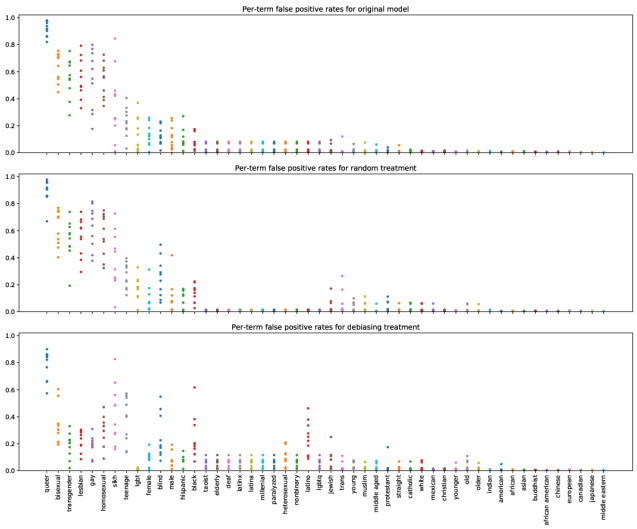

Figure 3: Per-term false positive rates for the baseline, random control, and bias-mitigated models.

Figure 4 shows the per-term false negative rates for the three experiments. The effect is less pronounced here since we added non-toxic (negative) data only, aiming specifically to combat false positives. Most importantly, we do not see an increase in variance of false negative rates, demonstrating that the bias mitigation technique reduces unintended bias on false positives, while not introducing false negative bias on the measured terms.

Pinned AUC

We also demonstrate a reduction in unintended bias using the new pinned AUC metric introduced in this work. As with error rates, a more fair model will have similar performance across all terms. Figure 5 show the per-term pinned AUC for each model, and we again see more uniformity in from the bias-mitigated model. This

[2]The details of the model and code are available at https://github.com/conversationai/unintended-ml-bias-analysis

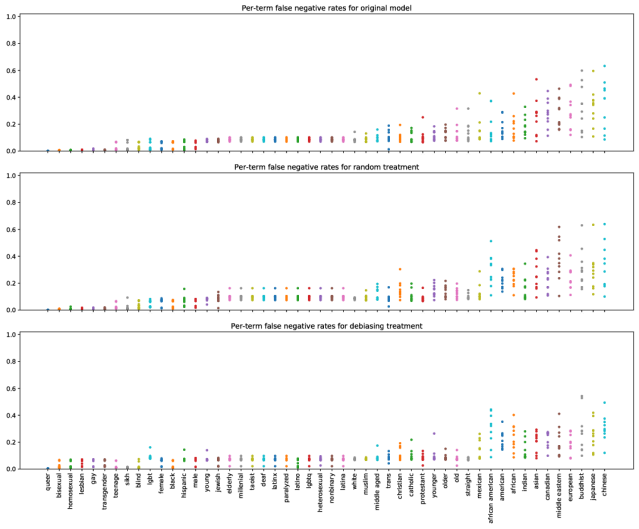

Figure 4: Per-term false negative rates for the baseline, random control, and bias-mitigated models.

demonstrates that the bias mitigation technique reduces unintended bias of the model's real-valued scores, not just of the thresholded binary classifier used to measure equality difference.

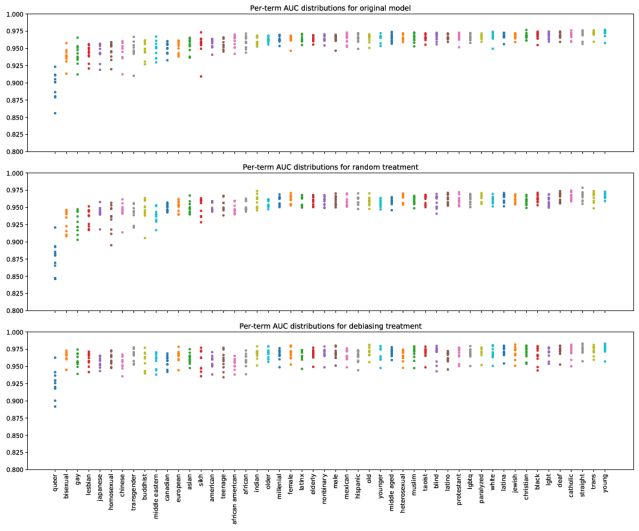

Figure 5: Per-term pinned AUC for the baseline, random control, and bias-mitigated models.

Equality Difference Summary

Finally, we look at the equality difference for false positives, false negatives, and pinned AUC to summarize each chart into one metric, shown in 5. The bias-mitigated model shows smaller sums of differences for all three metrics, indicating more similarity in performance across identity terms, and therefore less unintended bias.

	Sums of differences		
Metric	Baseline	Control	Bias-Mitigated
False Positive Equality Difference	74.13	77.72	52.94
False Negative Equality Difference	36.73	36.91	30.73
Pinned AUC Equality Difference	6.37	6.84	4.07

Table 5: Sums of differences between the per-term value and the overall value for each model.

FUTURE WORK

This work relies on machine learning researchers selecting narrow a definition of unintended bias tied to a specific set of identity terms to measure and correct for. For future work, we hope to remove the human step of identifying the relevant identity terms, either by automating the mining of identity terms affected by unintended bias or by devising bias mitigation strategies that do not rely directly on a set of identity terms. We also hope to generalize the methods to be less dependent on individual words, so that we can more effectively deal with biases tied to words used in many different contexts, e.g. white vs black.

CONCLUSION

In this paper, we have proposed a definition of unintended bias for text classification and distinguished it from fairness in the application of ML. We have presented strategies for quantifying and mitigating unintended bias in datasets and the resulting models. We demonstrated that applying these strategies mitigate the unintended biases in a model without harming the overall model quality, and with very little impact even on the original test set.

What we present here is a first step towards fairness in text classification, the path to fair models will of course require many more steps.

APPENDIX

Pinned AUC

We defined pinned AUC as copied below.

$$pD_t = s(D_t) + s(D), \quad |s(D_t)| = |s(D)| \tag{6}$$

$$pAUC_t = AUC(pD_t) \tag{7}$$

Depending on the exact data in the full set D, there are many options for sub-sampling down to $s(D)$, each with impacts on the pinned AUC metric and it's ability to reveal unintended bias. A full evaluation of these are left for future work, but here is a quick summary of the possible variants:

(1) Replacement: While $D_t \subset D$, it may make sense to sample such that $D_t \not\subset s(D)$. The results shown in this work sample this way.
(2) Other subgroups: If D contains many subgroups in different amounts, $s(D)$ could be sampled to ensure equal representation from each group. In this work, D is synthetically constructed such that each subgroup is equally represented.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. *CoRR* abs/1707.00075 (2017). http://arxiv.org/abs/1707.00075

[3] Su Lin Blodgett and Brendan O'Connor. 2017. Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. *CoRR* abs/1707.00061 (2017). http://arxiv.org/abs/1707.00061

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *CoRR* abs/1607.06520 (2016). http://arxiv.org/abs/1607.06520

[5] François Chollet et al. 2015. Keras. https://github.com/fchollet/keras.

[6] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27, 8 (2006), 861–874.

[7] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311

[8] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. *CoRR* abs/1609.07236 (2016). http://arxiv.org/abs/1609.07236

[9] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *CoRR* abs/1610.02413 (2016). http://arxiv.org/abs/1610.02413

[10] Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *ACL*.

[11] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. *CoRR* abs/1609.05807 (2016). http://arxiv.org/abs/1609.05807

[12] R. Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. *Ethics in Natural Language Processing*.

[13] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1391–1399. https://doi.org/10.1145/3038912.3052591