Individual Calibration in NLP
Reliability in ML Project
236601 048100

Zachary Bamberger, Edan Kinderman

August 14, 2023

Abstract

Bias detection and mitigation in Natural Language Processing (NLP) pose significant challenges, particularly in addressing individual instances and nuanced forms of bias that transcend predefined protected attributes. Building upon the work of Zhao et al., (2020) [19], we introduce Individualized Random Calibration for Language Forecasters (IRCLF), a new method that integrates individually calibrated random forecasters with empirical transformer encoder models [5]. Unlike conventional methods, IRCLF seamlessly balances accuracy and *individualized* fairness during the training phase, without dependence on predefined demographic information.

We leverage IRCLF for the toxicity prediction task [7, 1], measuring the model's learned bias towards nine protected groups [1]. Our findings reveal a successful replication of the trends in [19], where a single hyper-parameter, α, adeptly negotiates between an accuracy-oriented loss (NLL) and a fairness-oriented loss (PAIC). Yet, a critical examination employing more expressive and practical classification metrics uncovered shortcomings in achieving fairness by contemporary standards. This work not only advances the understanding of individualized fairness in NLP but also identifies vital areas for future research and improvement [2].

1 Introduction

The pervasive influence of biases in NLP models [2, 12] has prompted significant research into detection and mitigation strategies. This work focuses on *social bias*, which encompasses well known forms of discrimination such as gender [20] and racial biases [4, 3, 17]. However, it is challenging to identify and model *all* the experiences of *all* protected groups. It is likely that some demographics will be left out of a particular study, and therefore not get the benefits of the resulting bias-mitigation method. In essence, it is challenging to identify all confounding variables when measuring the causal effect of protected groups on a classifier's behaviour [14, 13] as depicted in Figure 1.

This paper leverages a more subtle and context-dependent analysis of bias, known as *individualized social bias*[3]. Traditional methods often rely on predefined demographic information or fail to address individual instances of bias. This paper presents a novel approach that builds

[1] These groups are: Black People, White People, Christians, Women, Men, LGBTQ, Jewish People, Muslims, and those suffering from psychiatric or mental-illness

[2] The code utilized for our experiments can be accessed at the following GitHub repository.

[3] I.e., social bias on the granularity of individuals, and without relying on pre-defined protected groups.

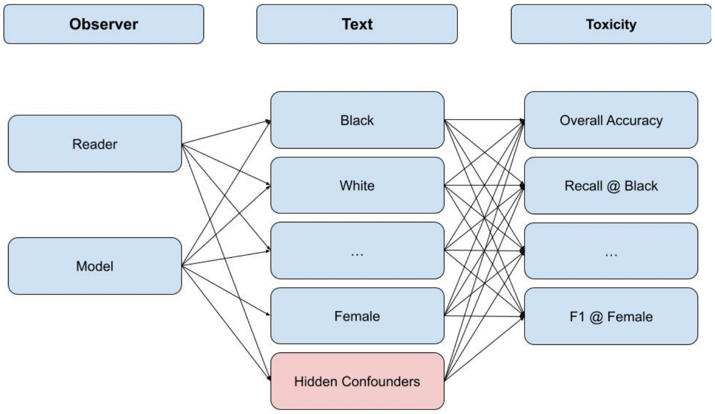

Figure 1: Causal Effects of textual properties on toxicity

on recent advancements in individualized calibration and randomized forecasting [19]. By exploring the explicit trade-off between fairness and accuracy through a single hyper-parameter, α, we offer a comprehensive and nuanced approach to bias detection and mitigation.

2 Background - Fairness in NLP

Bias detection and mitigation in NLP rely on metrics such as TPR, FPR, BPSN, and BNSP [1, 10] to evaluate biases across demographic groups. However, these metrics may overlook individual instances and more nuanced forms of bias [19]. Concretely, a model's form of bias toward an example may not be related to the example's protected attributes.

Recent advancements have introduced a variety of debiasing techniques:

- **BLIND** ([11]): A method that emphasizes context-aware evaluation and bias removal without specifying demographics, offering a flexible approach to bias mitigation.
- **Counterfactual Data Augmentation (CDA)** ([20]): This technique involves augmenting the training data with counterfactual examples, addressing bias by diversifying the dataset. Its effectiveness varies depending on the nature of the bias, the model that generates the counterfactual, and the data on which the model is applied.
- **Iterative Nullspace Projection (INLP)** ([15]): INLP has demonstrated effectiveness in mitigating both gender and non-gender biases. It attempts to remove information from token representations, but its assumption that users can identify all sources of bias can be a limitation.

3 Problem Setup

Our focus is on a regression task, approached from a probabilistic perspective. This implies that our model, denoted as h, processes the feature set $x \in \mathbb{R}^d$, and rather than predicting a specific target value $y \in \mathbb{R}$, it generates a Gaussian Cumulative Distribution Function (CDF). This CDF outlines the probability of the target achieving a particular output. Therefore, $h[x]$

represents the model output, a CDF function, and $h[x](y) \in \mathbb{R}$ signifies the model's estimated probability that the target value is y. We derive this CDF by designing a model that yields two scalar outputs: the mean $\mu \in \mathbb{R}$ and the standard deviation $\sigma \in \mathbb{R}^+$, which together define the Gaussian CDF. We will call such a model a forecaster.

We note that $\boldsymbol{X}$, $\boldsymbol{Y}$ denote random variables, and x, y denote fixed values.

4 Base Method

This project takes its inspiration from the article "Individual Calibration with Randomized Forecasting" [19]. The authors reference the Inverse CDF Theorem, which posits that $F_z(\boldsymbol{Z})$ is a random variable with a uniform distribution in [0,1]. Here, F_z is the CDF of a random variable z, and the function's inputs are sampled from the z distribution. Leveraging this theorem, we can articulate the average calibration of a regression model.

A forecaster H is ϵ approximately average calibrated (with respect to distance metric d) if:

$$d(F_{\boldsymbol{H}[\boldsymbol{X}](\boldsymbol{Y})}\, ,\, F_{\boldsymbol{U}}) \leq \epsilon$$

Here, d is a distance function between CDFs (the paper employs the Wasserstein-1 distance [8]). However, the challenge with average calibration is that it only calibrates on an average basis, potentially leading to discrimination against certain data subgroups (for instance, minor ethnic groups). To ensure fairness, the authors argue for an individually calibrated model, which guarantees calibration for each sample, and therefore guarantees fairness across all subgroups.

A forecaster H is (ϵ , δ)-probably approximately individually calibrated (PAIC) (with respect to distance metric d) if:

$$Pr[\text{err}_{\boldsymbol{H}}(\boldsymbol{X}) \leq \epsilon] \geq 1 - \delta$$

$$\text{err}_{\boldsymbol{H}}(x) = d(F_{\boldsymbol{H}[x](\boldsymbol{Y})}, F_{\boldsymbol{U}})$$

The authors then highlight a known finding that a deterministic forecaster cannot achieve individual calibration using training data [18, 6].

To address this, they suggest the use of a Randomized model, $\bar{h}[x, \boldsymbol{R}]$, where h is deterministic, but $r \in \mathbb{R}$ is a random uniform variable ranging from 0 to 1. They demonstrate that in this scenario, this randomized model, when trained on training data, can achieve individual calibration, as per the following theorem:

$$\bar{h}[x, r] = r\, ,\, \forall r \in [0, 1] \;\Rightarrow$$

$$\bar{h}[x, \boldsymbol{R}] \text{ is individually calibrated}$$

While this indicates that the model is fair, it doesn't necessarily mean it's useful. The model also needs to be "sharp", implying that the predicted probability should be concentrated around the true value.

5 The Article Experiments

The authors applied their model to the tabular UCI Crime and Communities dataset [16], aiming to predict crime rates based on neighborhood features. They used a small Multilayer Perceptron (MLP) model and trained it with two loss functions, $\mathcal{L}_{\text{PAIC}}$ for individual calibration (or fairness) and $\mathcal{L}_{\text{NLL}}$ for sharpness. They defined as:

$$\mathcal{L}_{\text{PAIC}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} |\bar{h}_\theta[x_i, r_i](y_i) - r_i|$$

$$\mathcal{L}_{\text{NLL}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{d}{dy}\bar{h}_\theta[x_i, r_i](y_i)$$

The overall loss was a combination of these two.

$$\mathcal{L}_\alpha(\theta) = (1-\alpha)\cdot\mathcal{L}_{\text{PAIC}}(\theta) + \alpha\cdot\mathcal{L}_{\text{NLL}}(\theta)$$

Their results showed a trade-off in the loss functions: a larger α led to a sharper model but with a larger fairness loss, while a smaller α resulted in a better fairness loss but worsen the sharpness loss. They also implemented an average calibration method [9], which improved average calibration but still showed discrimination against certain subgroups.

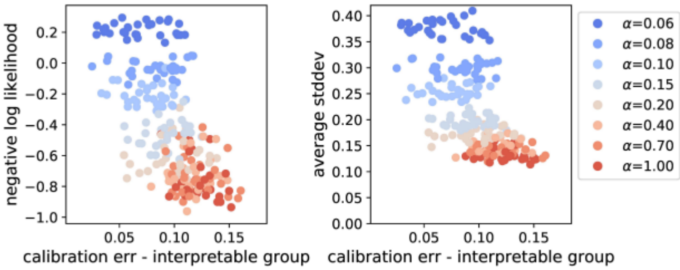

Figure 2: The original article results

In our view, a key constraint of this study is that it was only checked in the case of a very simple scenario: a minor tabular dataset processed by a simple model. This raises questions about the method's efficacy in more challenging tasks, complex models, and diverse domains. Another notable issue is the adverse effect of their fairness strategy on the model's performances ("sharpness"), indicating a potential trade-off between fairness and performance.

Another issue, is that the selected metrics for evaluating the model. While the authors highlight the ability of α to regulate the "sharpness-fairness" tradeoff, evidenced in their loss functions, these metrics offer limited insight into the model's practical utility. It remains unclear how the model's "fairness" translates to tangible outcomes in real-world tasks. Concretely, the authors only measured model performance and fairness with respect to the losses on which the model was trained, and not on more interpretable and robust metrics such as precision, recall, or f1.

6 Our Extension

We aim to extend the article's methodology to a new domain: Natural Language Processing (NLP). Our work involves the CivilComments dataset [7, 1], which contains 1.78 million comments, which we will define as $x \in \mathbb{R}^d$. Each comment is assigned a toxicity level $y \in \mathbb{R}$ ranging from 0 to 1, with higher values indicating greater toxicity. Our objective is to train a model that, given a comment, outputs a mean and standard deviation, defining a Gaussian CDF as per the article. Here, $h[x](y) \in \mathbb{R}$ represents the probability that the model h assigns to the comment $x \in \mathbb{R}^d$ having a toxicity level $y \in [0, 1]$, when $y \in \mathbb{R}$. When viewed as a binary classification problem, comments are considered toxic when $y \geq 0.5$

Some comments contain additional metadata, and are also labeled with group identifiers. These group identifiers contain values in the range of $[0, 1]$, where larger values specify that the comment pertains to a specific group to a greater extent. We focused on nine sufficiently represented groups $g \in G$, in the dataset, where $G = \{$'black', 'white', 'christian', 'female', 'male', 'homosexual gay or lesbian', 'jewish', 'muslim', 'psychiatric or mental illness'$\}$. We denote the value of the group identifier g for some comment x_i as x_{i_g}. A comment, x_i with a high $g = "women"$ label, for instance, indicates that the comment is very much about women, but not necessarily that it is toxic. We define a comment x_i to be associated with a subgroup $g \in G$, if $x_{i_g} \geq 0.5$.

Our goal is to assess the model's fairness across different subgroups and determine if the article's method yields a model that is fairer for all groups by enforcing individual calibration. This task is motivated by the importance of identifying highly toxic comments for content filtering. It's crucial that such a model is "fair", i.e., it doesn't permit toxic comments against specific subgroups or minorities, or falsely flags comments about a particular group as discriminative.

From a research perspective, we find it intriguing to examine whether the article's method can be effectively applied to more complex data and models, as encountered in this NLP task.

We utilize a pre-trained BERT (Bidirectional Encoder Representations from Transformers) [5] as our base model ("bert-base-uncased") to derive the comment embeddings ($h_{x_i} = BERT(x_i)$). Following the final layer of this model, we apply a small Multilayer Perceptron (MLP) to the CLS token embedding to obtain the mean and standard deviation as a function of the hidden representation, h_{x_i}. We train this model using the loss function from [19], experimenting with different α values to observe their impact on performance and fairness. The resulting architecture is depicted in Figure 3

In this new task and model, we introduced two hyperparameters for enhanced performance.

- *input r dim*: replicates the random variable r 'input r dim' times, before integrating it into the MLP's input, amplifying its signal. While the original study used an 'input r dim' of 1, we explored higher values.
- *input r upper bound*: samples a uniform random variable from $[0, \text{input } r \text{ upper bound}]$. This aligns 'input r' more closely with other MLP inputs, a known advantage. We should note that the foundational article set this upper bound at 1, and deviating from it hurts the theoretical guarantees of the article (the Inverse CDF Theorem).

7 Evaluation Methods

We employ the following metrics:

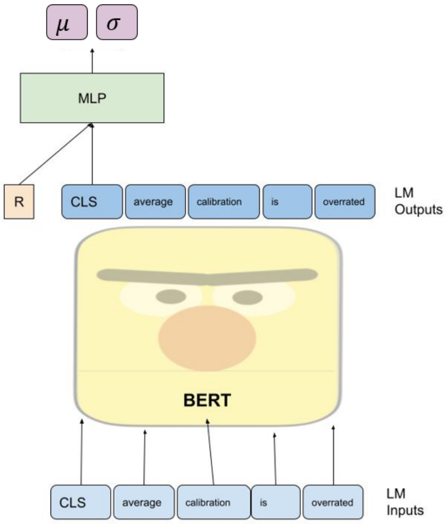

Figure 3: Our forecaster: pre-trained BERT and a regression head

- Loss functions from the article: $\mathcal{L}_{\text{PAIC}}$ to measure the model's individual calibration (and also fairness), $\mathcal{L}_{\text{NLL}}$ and $\mathcal{L}_{\text{stdevv}}$ (the CDF std) to assess the model's sharpness. A smaller standard deviation is desirable as it indicates a sharper model. Those are the metrics used in the article to evaluate the model.

- We extend the article's metrics from strictly the regression setting to also accommodate the binary classification setting. As you might recall, we classify comments with $y_i \geq 0.5$ as toxic, and the rest as non-toxic. This approach is practical as it enables us to effectively balance our data as is common in binary classification tasks. For this classification task, we evaluate the model's True Positive Rate (TPR), False Positive Rate (FPR), Precision, F1 score, and Accuracy. The newly introduced metrics, not present in the original article, provide a more comprehensive evaluation of the proposed method. They serve as indicators of whether the derived model is both practical and fair in real-world applications.

We calculate all these metrics and losses on all the subgroups in our test set, and also on the full test set (without considering sub-group) in Figures 4 and 5. When assessing the metrics in each group, we ensure that samples from this group are balanced. This means that we have an equal number of toxic and non-toxic comments associated with that group. This is done by sampling without replacement. This approach enhances the interpretability of our evaluation. By maintaining an equal ratio, any observed discrepancies in the model's performance across groups can be attributed solely to potential biases towards specific groups, rather than imbalances in the toxic-to-non-toxic sample distribution within the group. We should note that in the "general metric" we calculate on all the dataset (can be seen in the black plot), we don't perform this toxicity balancing.

8 Our Results

Our conclusion is based on 38 trained BERT forecasters. First, our analysis confirmed that the model consistently exhibits biases against particular groups. We can see an example for that in Figure 4. This discriminatory behavior persists across various seeds, α values, and hyperparameters. For example, the model often produces less accurate results for the groups 'homosexual', 'black', 'white' and 'Muslim'.

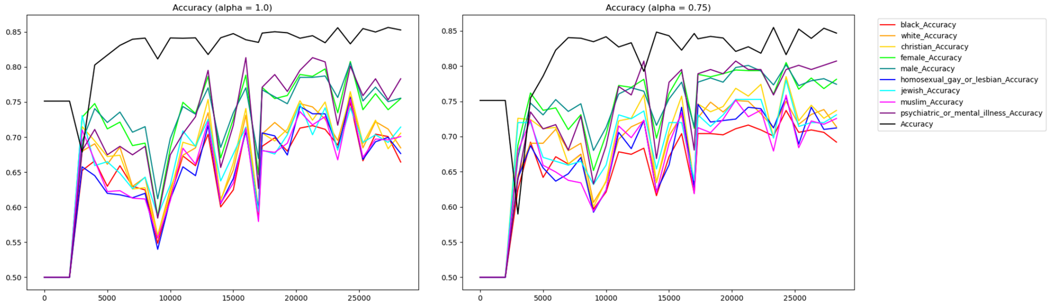

Figure 4: Accuracy on the validation set during training

Furthermore, as shown in Figure 5, the models consistently demonstrate high precision yet suffer from low recall. This indicates that while the system's positive predictions are typically accurate, it frequently overlooks numerous genuine positive instances.

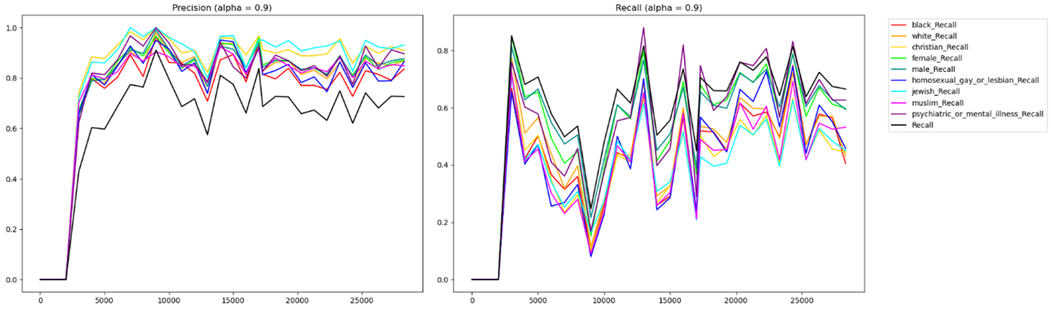

Figure 5: Precision and recall on the validation set during training

In our next experiment, depicted in 6, we found that even with very small α (including $\alpha = 0$, which represents solely the fairness loss), the models struggled to reduce the fairness loss $\mathcal{L}_{\text{PAIC}}$ (sometimes called $\mathcal{L}_{\text{CDF}}$) during training. However, a smaller α did prevent the fairness loss from increasing too much. It was only by reducing the 'input r upper bound' to 0.1 that we achieved a model capable of effectively decreasing the fairness loss. While this method works empirically, it does not maintain the theoretical guarantees established in [19].
The scatter plots depicted in Figure 7 show the performance of various models trained with different α values on the test set. Here, "worst group" denotes the poorest performance among all groups, while "best group" signifies the top-performing group. We can see that we succeed to replicate the findings of the original article. Specifically, models trained with a larger α exhibit lower $\mathcal{L}_{\text{NLL}}$ (indicative of sharpness) but higher $\mathcal{L}_{\text{CDF}}$ / $\mathcal{L}_{\text{PAIC}}$ (reflecting fairness). Conversely, smaller α values display the opposite trend. Although these patterns were evident with 'input r dim'$= 1$, the distinctions became more pronounced when 'input r dim' was increased to 8.

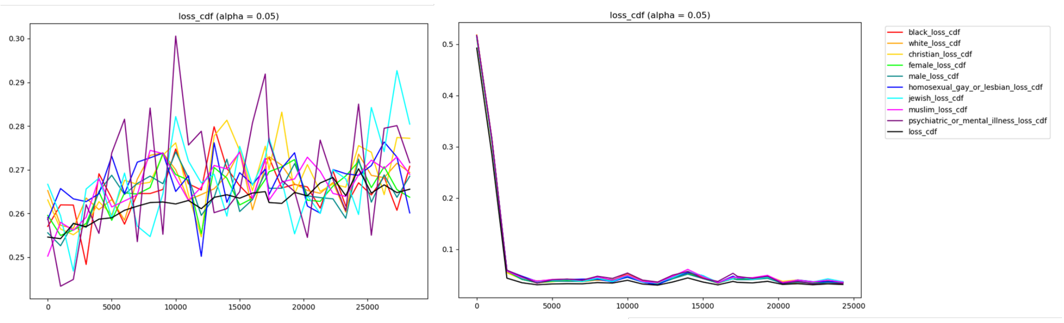

Figure 6: $\mathcal{L}_{\text{PAIC}}$ on the validation set. Left: 'input r dim' = 8, 'input r upper bound' = 1. Right: 'input r dim' = 8, 'input r upper bound' = 0.1.

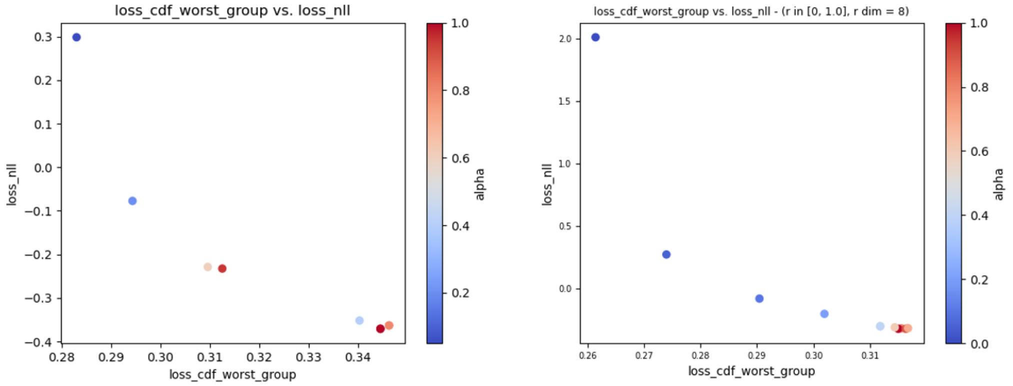

Figure 7: Visualization of performance/fairness trade-off as originally demonstrated in [19]

The weakness of this method becomes clearer when we employ the more application-oriented metrics from the binary classification problem we formulated. In Figure 8, the anticipated sharpness-fairness tradeoff, governed by α, is not evident. In most instances, a higher α (indicative of a stronger NLL loss) simply yields a model that is both sharper and "fairer" (in the sense that even the most disadvantaged group benefits from improved performance). We should note, that by setting 'input r upper bound' to 0.1, we can see this tradeoff in the F1 scores. However, this property does not exist clearly in the other metrics we considered.

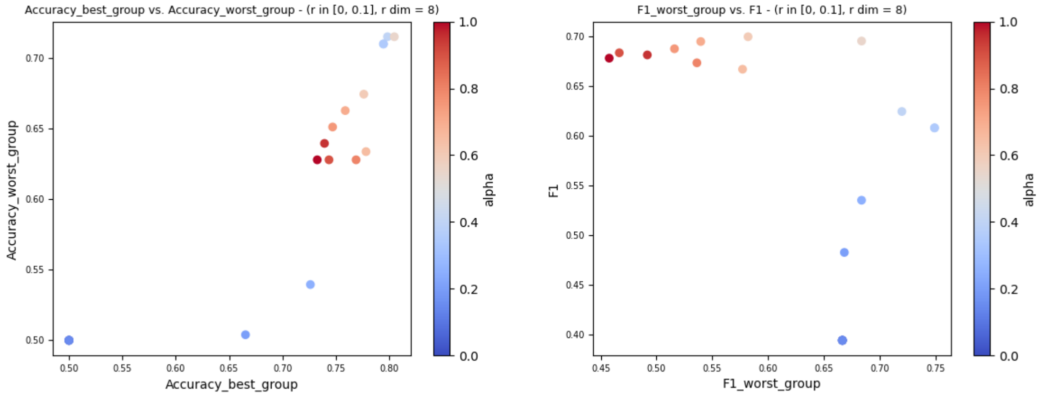

Figure 8: We find that the individualized fairness proposed by [19] does not generalize effectively to binary classification fairness metrics.

9 Future Work

1. Introduce Sufficiency as a metric, in order to check our model fairness.
2. Perform average calibration after randomized individual calibration.
3. Experiment with freezing BERT, and only training the MLP.
4. Compare our method to BLIND, INLP, and Counterfactual Augmentation.
5. Using stronger models than BERT (e.g., SentenceBERT or large T5 model).
6. Exploring the behavior of low α. How effectively can the model optimize the CDF loss?
7. Consider doing individual calibration after fine-tuning BERT with coefficient $\alpha = 1$.
8. Compare average and individual calibration on top of a pre-trained BERT that has been fine-tuned on toxicity prediction in the traditional fashion (i.e., a binary classification head). Specifically, apply these forms of calibration after the model's been fine tuned, and the classification head has been removed. Experiment with/without freezing the underlying BERT encoder model.

References

[1] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA, 2019. Association for Computing Machinery.

[2] A. Chalfin, O. Danieli, A. Hillis, Z. Jelveh, M. Luca, J. Ludwig, and S. Mullainathan. Productivity and selection of human capital with machine learning. *American Economic Review*, 106(5):124–27, May 2016.

[3] T. Davidson, D. Bhattacharya, and I. Weber. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy, Aug. 2019. Association for Computational Linguistics.

[4] M. De-Arteaga, A. Romanov, H. M. Wallach, J. T. Chayes, C. Borgs, A. Chouldechova, S. C. Geyik, K. Kenthapadi, and A. T. Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] R. Foygel Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

[7] Kaggle. Jigsaw unintended bias in toxicity classification, 2023. [Online; accessed 25-July-2023].

[8] L. V. Kantorovich. Mathematical methods of organizing and planning production. *Manage. Sci.*, 6(4):366–422, jul 1960.

[9] V. Kuleshov, N. Fenner, and S. Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804. PMLR, 2018.

[10] H. Orgad and Y. Belinkov. Choose your lenses: Flaws in gender bias evaluation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 151–167, Seattle, Washington, July 2022. Association for Computational Linguistics.

[11] H. Orgad and Y. Belinkov. BLIND: Bias removal with no demographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8801–8821, Toronto, Canada, July 2023. Association for Computational Linguistics.

[12] J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[13] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 12 1995.

[14] J. Pearl. *Causality.* Cambridge university press, 2009.

[15] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online, July 2020. Association for Computational Linguistics.

[16] U. M. L. Repository. Welcome to the uc irvine machine learning repository. [Online; accessed 25-July-2023].

[17] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics.

[18] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[19] S. Zhao, T. Ma, and S. Ermon. Individual calibration with randomized forecasting. In *International Conference on Machine Learning*, pages 11387–11397. PMLR, 2020.

[20] R. Zmigrod, S. J. Mielke, H. Wallach, and R. Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy, July 2019. Association for Computational Linguistics.