# Perturbation Augmentation for Fairer NLP

**Rebecca Qian**
Facebook AI Research
rebeccaqian@meta.com

**Candace Ross**
Facebook AI Research
ccross@meta.com

**Jude Fernandes**
Facebook AI Research
judef@meta.com

**Eric Smith**
Facebook AI Research
ems@meta.com

**Douwe Kiela***
HuggingFace
douwe@huggingface.co

**Adina Williams***
Facebook AI Research
adinawilliams@meta.com

## Abstract

Unwanted and often harmful social biases are becoming ever more salient in NLP research, affecting both models and datasets. In this work, we conceptualize fairness as robustness to demographic differences in text input. We explore the robustness of language models (LMs) to demographic changes using a perturber model and ask whether training on demographically perturbed data leads to fairer language models. We find that (i) language models (LMs) pre-trained on demographically perturbed corpora are typically more fair, and (ii) LMs finetuned on perturbed GLUE datasets exhibit more robustness (meaning less demographic bias) on downstream tasks, and (iii) increased robustness and fairness improvements do not come at the expense of performance on downstream tasks. We hope that this exploration of neural demographic perturbation will help drive more improvement towards fairer NLP.

## 1 Introduction

There is increasing evidence that models can instantiate social biases (Buolamwini and Gebru, 2018; Stock and Cissé, 2018; Fan et al., 2019; Merullo et al., 2019; Prates et al., 2020), often replicating or amplifying harmful statistical associations in their training data (Caliskan et al., 2017; Chang et al., 2019). Training models on data with representational issues can lead to unfair or poor treatment of particular demographic groups Barocas et al. (2017); Mehrabi et al. (2021), a problem that is particularly egregious for historically marginalized groups, including people of color (Field et al., 2021), and women (Hendricks et al., 2018).

In this work, we explore the efficacy of a dataset alteration technique that rewrites demographic references in text, such as changing "women like shopping" to "men like shopping". Similar demographic perturbation approaches have been fruitfully used to measure and often lessen the severity of social bias in text data (Prabhakaran et al., 2019; Zmigrod et al., 2019; Dinan et al., 2020; Webster et al., 2020; Ma et al., 2021; Smith and Williams, 2021; Renduchintala and Williams, 2022). Most approaches for perturbing demographic references, however, rely on rule-based systems, which unfortunately tend to be rigid and error prone, resulting in noisy and unnatural perturbations (see section 1.1). While some have suggested that a neural demographic perturbation model may generate

Workshop on Robustness in Sequence Modeling, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

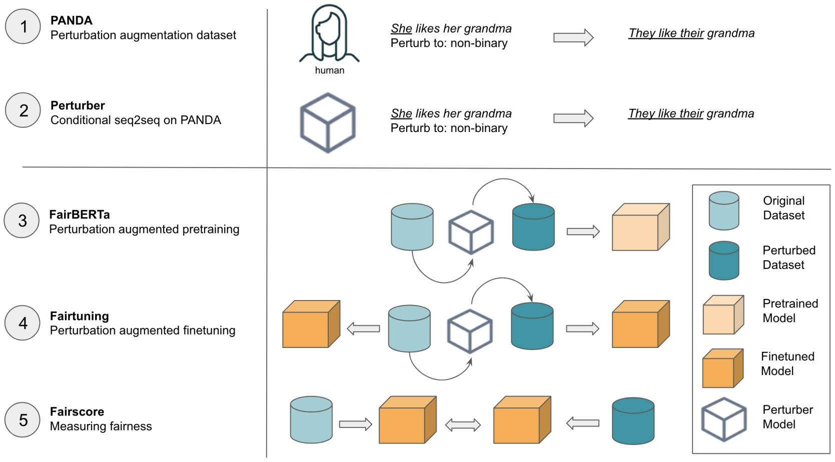

Figure 1: Our contributions.

higher quality text rewrites, there are currently no annotated datasets large enough for training neural models (Sun et al., 2021).

We train a perturber model on the **P**erturbation **A**ugmentation **N**LP **DA**taset (**PANDA**)[1], a novel dataset of 98K human-generated demographic text perturbations. The perturber takes in (i) a source text snippet, (ii) a word in the snippet referring to a demographic group, and (iii) a new target demographic attribute, and generates a perturbed snippet that refers to the target demographic attribute, while preserving overall meaning. We find that the perturber generates high quality perturbations, outperforming heuristic alternatives. We use our neural perturber to augment existing training data with demographically altered examples, weakening unwanted demographic associations.

We explore the effect of demographic perturbation on language model training both during *pretraining* and *finetuning* stages. We pretrain **FairBERTa**, the first large language model trained on demographically perturbed corpora, and show that its fairness is improved, without degrading performance on downstream tasks.

We also investigate the effect of **fairtuning**, i.e. finetuning models on perturbation augmented datasets, on model fairness. We find that fairtuned models perform well on a variety of natural language understanding (NLU) tasks while also being fairer on average than models finetuned on the original, unperturbed datasets.

Finally, we propose **fairscore**, an extrinsic fairness metric that uses the perturber to measure fairness as robustness to demographic perturbation. Given an NLU classification task, we define the fairscore as the change in model predictions between the original evaluation dataset and the perturbation augmented version. Prior approaches to measuring fairness in classifiers often rely on "challenge datasets" to measure how predictions differ in response to demographic changes in inputs (Zhao et al., 2018; Rudinger et al., 2018; De-Arteaga et al., 2019; Parrish et al., 2021). However, collecting human annotations can be costly, and task specific evaluation sets do not always generalize across NLU tasks. The fairscore is a versatile, complementary method to challenge datasets that can be easily applied to any NLP dataset. We see significant improvements in the fairscore from fairtuning on a range of GLUE tasks.

Our main contributions are summarized in Figure 1. Using a neural perturber to demographically augment model training data is a promising direction for lessening bias in large language models. To enable more exploration and improvement upon the present work, we will release PANDA, our controllable perturber, FairBERTa, and all other trained models and code artifacts under a permissive license.

[1] For more details on how PANDA was collected, see the full paper to appear in EMNLP proceedings 2022.

### 1.1 Training the Demographic Perturber

We frame training a demographic perturber as a conditional sequence-to-sequence task. Given input snippet $s$, perturbable word $w$ and target attribute $a_t$, we seek to learn $P(\tilde{s}|s, w, a_t)$, where $w$ and $a_t$ are discrete control variables that we prepend to perturber inputs. The perturber inputs take the form `[perturbable word] [target attribute] <PERT_SEP> [input]`. The perturber is a finetuned BART model (Lewis et al., 2020) with 24 layers, 1024 hidden size, 406M parameters, and 16 attention heads. To train the perturber, we finetune BART on PANDA using the ParlAI library[2] (Miller et al., 2017), with training parameters provided in Table 5. We achieve a BLEU score of 88.0 (measured against the source) on the validation set, and perplexity of 1.06, which is likely low because perturbation preserves the majority of tokens.

Perturbing large ML training datasets is an important application of perturbation augmentation. Therefore, it is crucial that generation is fast and scalable to large text corpora. We experimented with different architectures and generation techniques to optimize for both quality and efficiency. Notably, T5 (Raffel et al., 2020) performed slightly better on certain NLP metrics (such as BLEU-4), but used much more memory during training and inference, resulting in *16x* slower generations in a distributed setting. We also explored different ways of decoding, and surprisingly, found that greedy decoding performs as well as beam search in our setting. We therefore use greedy decoding in our perturbation augmentation applications, which is also memory efficient.

**Comparison to Heuristics.** Is it necessary to train a perturber, or can we just use heuristics? Previous approaches relied on word lists (Zhao et al., 2019) or designing handcrafted grammars to generate perturbations (Zmigrod et al., 2019; Ma et al., 2021; Renduchintala and Williams, 2022; Papakipos and Bitton, 2022). However, word list approaches are necessarily limited (Dinan et al., 2020) and which words are included can really matter (Sedoc and Ungar, 2019). For instance, attributes are often excluded for being hard to automate: e.g., *Black, white* have been excluded because they often denote colors in general (Ma et al., 2021). Grammar-based approaches also require ad hoc solutions for phonological alternations (***a*** *banana* v. ***an*** *apple*), and struggle with one-to-many-mappings for pronouns (Sun et al., 2021), often incompletely handling pronoun coreference chains. We find that a neural perturber trained on high quality human annotations can correctly identify perturbable words and their coreference chains, and then generate rewritten text that is grammatical, fluent and preserves overall meaning.

### 1.2 Results

We present results showing that using the perturber leads to fairer models during pretraining (subsection 1.3) and to fairer models during finetuning without sacrificing accuracy (subsection 1.4).

### 1.3 FairBERTa: Perturbation Augmented Pretraining

***Setting:*** We train FairBERTa with the RoBERTa$_{\mathrm{BASE}}$ architecture (Liu et al., 2019) using 256 32GB V100 GPUs for 500k steps. To generate training data for FairBERTa, we apply the perturber to the RoBERTa training corpus (Liu et al., 2019) to help balance the representation of underrepresented groups and thereby reduce the prevalence and severity of unwanted demographic associations. During perturbation augmentation, we sample contiguous sequences of 256 tokens and select a demographic word and target attribute with uniform probability, which are provided as inputs to the perturber. Although it would be in principle straightforward to upsample the training data size appreciably, keeping data size fixed allows us to make a direct comparison between FairBERTa and RoBERTa on a variety of fairness metrics and downstream tasks. We train FairBERTa and RoBERTa on the full RoBERTa training corpus (160GB) and the BookWiki subset (16GB), and show that our observations on fairness and accuracy are consistent.

***Fairness Evaluations:*** We compare FairBERTa to RoBERTa trained with the same settings according to their performance on three fairness evaluation datasets. For CrowS-Pairs (Nangia et al., 2020), we report the percentage of examples for which a model assigns a higher (pseudo-)likelihood to the stereotyping sentence over the less stereotyping sentence. For the template-based Word Embedding Association Test (WEAT, Caliskan et al. 2017) and Sentence Encoder Association Test (SEAT, May

[2] `github.com/facebookresearch/ParlAI`

|  |  | **RoBERTa** | **FairBERTa** | **RoBERTa**$^{\dagger}$ | **FairBERTa** |
|---|---|---|---|---|---|
|  |  | *16GB of training data* |  | *160GB of training data* |  |
| HolisticBias | *gender* | 36.1 | **19.9** | 40.6 | **35.7** |
|  | *race* | 27.3 | **23.8** | 28.4 | **27.6** |
|  | *age* | 42.9 | **38.9** | **36.4** | 41.7 |
| WEAT/SEAT | *% sig. tests* | 53.5 | **40.0** | 60.0 | **36.7** |
| CrowS-Pairs | *gender* | 52.3 | **51.9** | 55.0 | **51.5** |
|  | *race* | **55.0** | **55.0** | **53.9** | 57.6 |
|  | *age* | **50.6** | 63.2 | 66.7 | **63.2** |

Table 1: Results of FairBERTa and RoBERTa on 3 fairness metrics across varying training dataset sizes. Numbers are percentages of metric tests revealing bias. RoBERTa$^{\dagger}$ refers to the model from liu-etal-2019-roberta; all other models were trained from scratch. For CrowS-Pairs, closer to 50 means a more fair model; for WEAT/SEAT & HolisticBias, lower means more fair. See subsection 1.3 for more details.

| **Model** | **Tuning** | **Size** | **CoLA** | **SST-2** | **STS-B** | **QQP** | **RTE** | **QNLI** | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|
| FairBERTa | orig. | 16GB | 62.81 | 92.66 | 88.37 | 91.22 | 72.75 | 92.13 | 83.32 |
| RoBERTa | orig. | 16GB | 59.81 | 93.92 | 89.87 | 91.17 | 72.92 | 91.89 | 83.26 |
| FairBERTa | orig. | 160GB | 61.57 | 94.61 | 90.40 | 91.42 | 76.90 | 92.99 | 84.65 |
| RoBERTa$^{\dagger}$ | orig. | 160GB | 61.36 | 93.50 | 90.90 | 91.77 | 75.50 | 92.70 | 84.29 |
| FairBERTa | fair | 16GB | 61.37 | 92.20 | 87.64 | 90.93 | 70.03 | 92.13 | 82.38 |
| RoBERTa | fair | 16GB | 58.09 | 93.58 | 88.66 | 91.04 | 71.12 | 91.73 | 82.37 |
| FairBERTa | fair | 160GB | 60.60 | 94.95 | 89.63 | 91.49 | 75.09 | 92.77 | 84.09 |
| RoBERTa$^{\dagger}$ | fair | 160GB | 59.71 | 93.50 | 90.20 | 91.56 | 75.80 | 92.70 | 83.91 |

Table 2: FairBERTa matches RoBERTa in Downstream Task Accuracy (GLUE Benchmark). Tuning refers to whether models are finetuned on original datasets or "fairtuned" on perturbed ones (denoted with 'fair'). RoBERTa and FairBERTa models report similar accuracy regardless of training size and tuning approach. We report Matthew's correlation for CoLA, Pearson's correlation for STS-B, and accuracy for all other tasks. Results are the median of 5 seeded runs. A dagger marks the Liu et al. model.

et al. 2019), we report the percentage of statistically significant tests and their average effect size. Lastly, for HolisticBias (HB, Smith et al. 2022), we measure the percentage of pairs of descriptors by axis for which the distribution of pseudo-log-likelihoods (Nangia et al., 2020) in templated sentences significantly differs.

***FairBERTa is more fair:*** Overall, FairBERTa shows improvements in fairness scores over training-size-matched RoBERTa models across our evaluations, and across two training dataset sizes (see Table 1). FairBERTa models show reduced demographic associations overall across HB templates, and have notably fewer statistically significant associations on WEAT/SEAT. CrowS-Pairs is more equivocal: e.g., FairBERTa (16GB) is closer than RoBERTA (16GB) to the desired score of 50% (demographic parity) for gender, but not for age. Worse performance on the age category is possibly due to the varied ways in which age is conveyed in language, e.g., *I was born 25 years ago* vs. *I am a child*. While the perturber is capable of perturbing phrases with numbers such as *eleven years old*, general issues with numerical reasoning (Dua et al., 2019; Geva et al., 2020; Lin et al., 2020) may still be present.

We find that fairness metrics sometimes report conflicting results, corroborating other recent findings (Delobelle et al., 2021; Goldfarb-Tarrant et al., 2021). While WEAT/SEAT tests and HB evaluation find FairBERTa (160GB) to be more fair along the race axis, CrowS-Pairs reported a better score for RoBERTa (160GB). Inconsistencies may be partly explained by data noise in CrowS-Pairs Blodgett et al. (2021), but we believe that the agreement (or lack thereof) of different NLP bias measurements warrants further exploration, and closer examinations of fairness evaluation datasets.

***FairBERTa has no Fairness-Accuracy Tradeoff:*** Previously, a fairer model often meant accepting lower task performance (Zliobaite, 2015; Menon and Williamson, 2018) or seeking a Pareto optimal solution (Berk et al., 2017; Zhao and Gordon, 2019). To determine whether there is a tradeoff between downstream task accuracy and fairness in our setting, we evaluate on 6 GLUE benchmark tasks Wang et al. (2018): sentence acceptability (Warstadt et al., 2019, CoLA), sentiment analysis (Socher et al., 2013, SST-2), text similarity (Cer et al., 2017, STS-B), textual entailment (Dagan et al., 2005; Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009, RTE), and question answering (Rajpurkar et al., 2016) recast to textual entailment (QNLI).[3]

FairBERTa models match the performance of RoBERTa models trained under the same setting to within 0.40% accuracy on average (see top half of Table 2). For some tasks (CoLA, SST-2, RTE and QNLI), FairBERTa (160GB) also slightly outperforms RoBERTa (160GB) and averages 0.75% higher overall accuracy on these tasks.

| Model | Tuning | Size | CoLA | SST2 | QQP | RTE | QNLI | Avg. |
|---|---|---|---|---|---|---|---|---|
| FairBERTa | orig. | 16GB | 5.46 | 2.04 | 5.61 | **6.45** | **1.70** | 4.25 |
| FairBERTa | fair | 16GB | **4.20** | **1.02** | **3.34** | **6.45** | 1.94 | **3.39** |
| FairBERTa | orig. | 160GB | 5.88 | 1.02 | 5.56 | **3.23** | 2.17 | 3.57 |
| FairBERTa | fair | 160GB | **4.41** | **0.51** | **2.86** | 6.45 | **1.70** | **3.19** |
| RoBERTa | orig. | 16GB | 6.51 | **1.02** | 6.89 | **6.45** | 2.88 | 4.75 |
| RoBERTa | fair | 16GB | **5.46** | 3.06 | **3.43** | 6.86 | **1.58** | **4.08** |
| RoBERTa† | orig. | 160GB | 6.93 | 2.55 | 7.60 | **4.03** | 2.17 | 4.66 |
| RoBERTa† | fair | 160GB | **3.78** | **1.02** | **3.22** | 6.45 | **1.67** | **3.23** |

Table 3: The fairscore for fairtuned models is lower in general. A lower fairscore, i.e., the percentage of classifier predictions that change during inference for a single model between the original evaluation set and the same evaluation set after perturbation augmentation, corresponds to a fairer model. The lowest fairscore for each task and setting is bolded. RoBERTa† is the model from liu-etal-2019-roberta.

### 1.4 Fairtuning: Finetuning on Perturbed Data

***Setting:*** In addition to comparing downstream performance in a traditional finetuning setting, we also compare performance and fairness during **fairtuning**, where models are finetuned on demographically perturbed downstream datasets. The number of perturbable examples and the proportions of demographic axes varies across fairtuning data by task (see statistics in Table 8, and examples in Table 9).

***Fairtuning does not degrade downstream task accuracy:*** Fairtuned models match their finetuned counterparts in accuracy on the original (unperturbed) GLUE validation sets (compare the top half of Table 2 to the bottom). Surprisingly, for some tasks (SST-2, QQP and RTE), fairtuning resulted in slightly higher original validation set performance than finetuning does for some model configurations. The largest drop in performance from fairtuning occurs for RTE, where FairBERTa trained on BookWiki (16GB) shows a decrease of 2.72% in accuracy. Swings on RTE may be due to its smaller size (see Table 8), as we observe more variance across finetuning runs as well. Finetuning or fairtuning from an existing NLI checkpoint, as in Liu et al. 2019, might result in more stability.

### 1.5 Measuring Fairness with the Fairscore

***Setting:*** Finally, we compute the **fairscore** as an extrinsic fairness evaluation metric. Recall that, given a classifier and evaluation set, the fairscore of the classifier is the percentage of predictions that change when the input is demographically altered with the perturber.

[3] We exclude several GLUE tasks for which the number of demographically perturbable examples was too low to draw meaningful conclusions. We follow liu-etal-2019-roberta's training procedure, conducting a limited hyperparameter sweep for each task varying only learning rate and batch size. For each task, we finetune for 10 epochs and report the median development set results from five random initializations.

***Fairscore is best for Fairtuned Models:*** Fairtuned models have lower (i.e., better) fairscores on average[4], meaning that their predictions change the least from perturbation (see Table 3). On average, fairtuned models saw a 0.84 point reduction in the fairscore as compared to models finetuned on unperturbed data; this is true for both RoBERTa and FairBERTa and across training data sizes. We also find that FairBERTa models are more robust to demographic perturbation on downstream tasks, even when finetuned on the original datasets (Table 3). FairBERTa models have lower fairscores than RoBERTa models pretrained on similar sized datasets.

We also observe an additive effect where models that are both pretrained *and* finetuned on demographically perturbed data show more robustness to demographic perturbation on downstream tasks. Notably, the fairtuned versions of FairBERTa (16BG) and FairBERTa (160GB) have better average fairscores in general. The fairtuned FairBERTa (160GB) model reports the lowest average fairscore across all tasks (3.19). In our setting, we do not observe any relationship between demographic bias and data size in downstream tasks, suggesting that models of any size can learn demographic biases.

Overall, we find that perturbation augmentation can mitigate demographic bias during classification without any serious degradation to task performance for most tasks on the GLUE benchmark (see Table 2). While we do observe an interesting additive effect where LMs are more robust to demographic differences when they are pretrained on demographically altered datasets then fairtuned, we believe that further work is needed to better understand exactly how bias is learned and propagated during different stages of language model training.

## 2 Conclusion

As language models become more powerful and more popular, more attention should be paid to the demographic biases that they can exhibit. Models trained on datasets with imbalanced demographic representation can learn stereotypes such as *women like shopping*. While recent works have exposed the biases of LMs using a variety of techniques, the path to mitigating bias in large scale training datasets is not always clear. Many approaches to correct imbalances in the dataset have used heuristic rules to identify and swap demographic terms. We propose a novel method that perturbs text by changing the demographic identity of a highlighted word, while keeping the rest of the text the same. We find that our perturber model creates more fluent and humanlike rewrites than heuristics-based alternatives. We also show that training on demographically perturbed data results in more fair language models, in both pretrained language models and in downstream measurements, without affecting accuracy on NLP benchmarks. We hope our contributions will help drive exciting future research directions in fairer NLP.

## References

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: from allocative to representational harms in machine learning. In *Special Interest Group for Computing, Information and Society (SIGCIS)*, volume 2.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

[4] We report on all tasks except STS-B, a regression task, because the fairscore is defined for classification tasks.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Kai-Wei Chang, Vinodkumar Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, Hong Kong, China. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Maria De-Arteaga, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2021. Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models. *arXiv preprint arXiv:2112.07447*.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6342–6348, Hong Kong, China. Association for Computational Linguistics.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958, Online. Association for Computational Linguistics.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, volume 7.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. *arXiv preprint arXiv:2202.11923*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. *Advances in Neural Information Processing Systems*, 34.

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

Ninareh Mehrabi, Pei Zhou, Fred Morstatter, Jay Pujara, Xiang Ren, and Aram Galstyan. 2021. Lawyers are dishonest? quantifying representational harms in commonsense knowledge resources. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5033, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 107–118. PMLR.

Jack Merullo, Luke Yeh, Abram Handler, Alvin Grissom II, Brendan O'Connor, and Mohit Iyyer. 2019. Investigating sports commentator bias within a large corpus of American football broadcasts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6354–6360, Hong Kong, China. Association for Computational Linguistics.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Zoe Papakipos and Joanna Bitton. 2022. Augly: Data augmentations for robustness. *arXiv preprint arXiv:2201.06494*.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2021. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5740–5745, Hong Kong, China. Association for Computational Linguistics.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32(10):6363–6381.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Adi Renduchintala and Adina Williams. 2022. Investigating failures of automatic translationin the case of unambiguous gender. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3454–3469, Dublin, Ireland. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14.

João Sedoc and Lyle Ungar. 2019. The role of protected class word lists in bias identification of contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 55–61, Florence, Italy. Association for Computational Linguistics.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": finding bias in language models with a holistic descriptor dataset. *arXiv preprint arXiv:2205.09209*.

Eric Michael Smith and Adina Williams. 2021. Hi, my name is Martha: Using names to measure and mitigate bias in generative dialogue models. *arXiv preprint arXiv:2109.03300*.

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

Pierre Stock and Moustapha Cissé. 2018. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, volume 11210 of *Lecture Notes in Computer Science*, pages 504–519. Springer.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral english. *arXiv preprint arXiv:2102.06788*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Han Zhao and Geoffrey J. Gordon. 2019. Inherent tradeoffs in learning fair representations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 15649–15659.

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

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Indre Zliobaite. 2015. On the relation between accuracy and fairness in binary classification. *CoRR*, abs/1505.05723.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. 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. Association for Computational Linguistics.

# A Appendix

## A Problems with Perturbation Augmentation

While heuristic approaches have been widely used, they suffer from quality issues, which in turn result in particular demographic attributes being excluded in general. Three axis-attributes are most affected, and we will point to them as exemplars of the general issue: non-binary/underspecified, race/ethnicity-african-american, race/ethnicity-white.

To take an obvious example, English language heuristic demographic perturbation systems have to somehow handle the linguistic fact that gendered pronouns have different forms for each grammatical role in so-called "standard" English: both the feminine and the masculine pronouns use the same form for two grammatical functions, but not for the same two: *she, **her, her**, hers* v. *he, him, **his, his***. It is not straightforward for a heuristic system given *her* to determine whether to replace it with *his* or *him*. Put simply, a heuristic system that always maps *her* → *him* would fail for an example with a possessive (*unfortunately for her, I recently changed **her** schedule* → *unfortunately for him, I recently changed **him** schedule*) and one that maps *her* → *his* would fail for an example with an accusative (*unfortunately for **her**, I recently changed **her** schedule* → *unfortunately for **his**, I recently changed **his** schedule*). One might hope that a random selection of mappings could help, but since pronouns are highly frequent in natural language, even that sort of noisy approach would lead to a lot of ungrammatical examples.

The pronoun situation becomes even more complicated when including non-binary gender, since the most frequent pronoun for non-binary gender affects the verb form as well. For example, if we wanted to replace *he*→ *they* in the following example, *the owner came to our table and told us **he** already **is** thinking about starting a Turkish breakfast*, this would result in another grammatically incorrect sentence, *the owner came to our table and told us **they** already **is** thinking about starting a Turkish breakfast*. One might hope that one could just add bigrams to the word lists containing pronouns and all verb forms, but that doesn't straightforwardly work, as other words (sometimes several of them) can appear between the pronoun and the verb, and thus not be caught by a heuristic system. Although this particular issue only occurs (in English) in the context of singular *they*, it would be counter to the goals of a responsible AI work such as this one to accept higher noise for underserved identities like non-binary that are often ignored or overlooked in NLP tasks (Sun et al., 2021; Lauscher et al., 2022).

As if the situation with pronouns weren't complicated enough, often context is needed to determine whether particular words should be perturbed at all. For example, "Black" and "white" are polysemous adjectives that can be used not only as demographic terms but also as color terms. Despite the fact that these references aren't demographic, they would get perturbed by nearly every heuristic demographic perturbation system (*the person was wearing **a white** shirt* → *the person was wearing an Asian shirt* or *the **white** pawn attacked the black bishop* → *the black pawn attacked the black bishop*), altering the meaning significantly. If a heuristic system like this were used to measure model robustness to demographic perturbation say in an NLU classification task like natural language inference, it would be hard to determine whether the model failed to be robust to demographic changes (and hence should be deemed unfair) or if the textual changes had altered the meaning too much and that affected the label.

## B Perturber Human Evaluation

We conduct a human evaluation of the perturber outputs. We randomly selected 200 examples from the PANDA validation set to be perturbed by the perturber. The perturber outputs are annotated by 2 expert annotators; each example was annotated by both experts to calculate interannotator agreement. We use the same categorization of errors for perturber outputs as we did for the PANDA audit. We will release anonymized annotations along with the other artifacts from this work. The results of the perturber human evaluation are reported in Table 4.

Compared to the PANDA audit, the perturber human evaluation shows lower incidence of factuality change, fewer incomplete/incorrect perturbations, and fewer typos and unnatural examples. On the other hand, the perturber evaluation found higher occurrence of stage 1 errors and incorrectly unperturbed examples. From inspection, we observe that the perturber often fixes typos and grammatical issues in the input, likely an artifact of BART pretraining. The perturber is also successful at identifying complex coreference entity chains, even in long passages, resulting in fewer instances of incomplete perturbations. However, the perturber leaves more examples unperturbed, which may also reflect in the lower incidence of factuality issues. Interannotator agreement is similar to the PANDA audit.

We aim to be transparent about limitations of the perturber to inform downstream applications of the perturber, such as training data augmentation or model evaluation. While we found that most perturber outputs that are flagged under our annotation scheme are useable and inoffensive, our analysis is constrained to a small sample of PANDA. We encourage researchers to examine the perturber in other domains, and to make informed decisions around using the perturber.

| *Tag* | *% occurrence* | *% agreement* |
|---|---|---|
| factuality change | 28.5 | 77.5 |
| incomplete/incorrect | 20.0 | 84.5 |
| Stage 1 errors | 24.5 | 91.0 |
| typos and naturalness | 11.5 | 91.5 |
| incorrectly unperturbed | 19.5 | 84.0 |

Table 4: Perturber quality audit. Under each tag, we report the rate at which it occurs, as well as how often two annotators agreed. If either annotator included a tag for an example, that tag was aggregated as % occurrence.

## C Perturber Training Parameters

In this section, we describe hyperparameters for training the perturber. Table 5 describes the hyperparameters for finetuning BART-Large (Lewis et al., 2020) on PANDA, with 24 layers, 1024 hidden size, 16 attention heads and 406M parameters. Validation patience refers to the number of epochs where validation loss does not improve, used for early stopping. All perturber training and evaluation runs are conducted using the ParlAI library (Miller et al., 2017).[5] We trained the perturber using 8 × 16GB Nvidia V100 GPUs for approximately 4 hours.

| **Hyperparam** | **PANDA** |
|---|---|
| Learning Rate | 1e-5 |
| Batch Size | 64 |
| Weight Decay | 0.01 |
| Validation Patience | 10 |
| Learning Rate Decay | 0.01 |
| Warmup Updates | 1200 |
| Adam $\epsilon$ | 1e-8 |
| Adam $\beta_1$ | 0.9 |
| Adam $\beta_2$ | 0.999 |
| Gradient Clipping | 0.1 |
| Decoding Strategy | greedy |

Table 5: Hyperparameters for training the perturber by finetuning BART on PANDA.

## D FairBERTa Training Parameters

Table 6 contains hyperparameters for pretraining FairBERTa. FairBERTa is trained with the RoBERTa$_{\text{BASE}}$ Liu et al. (2019) architecture on 32GB Nvidia V100 GPUs with mixed precision using the Fairseq library Ott et al. (2019). We pretrain FairBERTa on 160GB perturbed data using 256 V100 GPUs for approximately three days. For RoBERTa and FairBERTa models trained on the 16GB BookWiki corpus (and perturbed BookWiki corpus), we use the same training settings, but use 100K max steps.

## E Downstream Task Training Parameters

Table 7 describes hyperparameters for finetuning and fairtuning RoBERTa and FairBERTa on GLUE tasks and the RACE (Lai et al., 2017) reading comprehension dataset. We conducted a basic hyperparameter exploration sweeping over learning rate and batch size, and select the best hyperparameter values based on the median validation accuracy of 3 runs for each task. Configurations for individual models, tuning approach and GLUE task will be released in our GitHub repository. Training runs on downstream tasks are done using HuggingFace. Models are trained on 8 × 32GB Nvidia V100 machines, with runtime ranging from 5 minutes for the smallest dataset (RTE) to 45 minutes for the largest dataset (QQP).

[5] https://parl.ai

| **Hyperparam** | **FairBERTa** |
|---|---|
| # Layers | 12 |
| Hidden Size | 768 |
| FFN inner Hidden Size | 3072 |
| # Attention Heads | 12 |
| Attention Head Size | 64 |
| Hidden Dropout | 0.1 |
| Attention Dropout | 0.1 |
| # Warmup Steps | 24k |
| Peak Learning Rate | 6e-4 |
| Batch Size | 8k |
| Weight Decay | 0.01 |
| Sequence Length | 512 |
| Max Steps | 500k |
| Learning Rate Decay | Linear |
| Adam $\epsilon$ | 1e-8 |
| Adam $\beta_1$ | 0.9 |
| Adam $\beta_2$ | 0.999 |
| Gradient Clipping | 0.0 |

Table 6: Hyperparameters for pretraining FairBERTa.

| **Hyperparam** | **GLUE** | **RACE** |
|---|---|---|
| Learning Rate | {1e-5, 2e-5, 3e-5} | 1e-5 |
| Batch Size | {16, 32} | 16 |
| Weight Decay | 0.1 | 0.1 |
| Max # Epochs | 10 | 3 |
| Learning Rate Decay | Linear | Linear |
| Warmup Ratio | 0.06 | 0.06 |

Table 7: Hyperparameters for finetuning RoBERTa and FairBERTa on GLUE and RACE.

## F Additional GLUE Statistics

We provide the percentage of examples in the validation set (used for reporting accuracy as test sets are hidden) that were perturbed across six tasks from the GLUE benchmark in Table 8. CoLA and RTE had the highest percentage of perturbable examples, followed by QNLI and STS-B, with SST-2 having the fewest.

width=

| | **CoLA** | **SST-2** | **STS-B** | **QQP** | **RTE** | **QNLI** |
|---|---|---|---|---|---|---|
| *age* | 9.2 | 7.5 | 12.2 | 6.4 | 13.2 | 6.5 |
| *gender* | 32.3 | 9.9 | 20.2 | 8.3 | 32.9 | 18.4 |
| *race* | 4.1 | 5.2 | 4.5 | 5.8 | 0.8 | 6.3 |
| *total* | 45.6 | 22.5 | 36.9 | 20.5 | 47 | 31.2 |

Table 8: The percentage of examples perturbed by demographic axis for each fairtuning task.

## G Preserving Classification Labels After Perturbation

We have assumed for the purposes of the fairscore that perturbing word axes and attributes should not affect the gold classification label. In general, this is a reasonable assumption, but there are edge cases, in particular, for examples that rely on human-denoting references as part of their meaning. Consider for example the hypothetical textual entailment example {P: *John saw his aunt*, H: *John saw his uncle*, gold-label: `not-entailment`}. If *aunt* is the chosen word, and the target attribute is `gender:man`, we have an issue: the new example will be {P: *John saw his uncle*, H: *John saw his uncle*, gold-label: `entailment`}. The entailment label will have changed, because the original

| Dataset | Input | Label | Perturbed |
|---|---|---|---|
| RTE | **premise**: *Swansea striker Lee Trundle has negotiated a lucrative image-rights deal with the League One club.* **hypothesis**: *Lee Trundle is in business with the League One club.* | entailment | No |
| RTE | **premise**: *Swansea striker Lisa Trundle has negotiated a lucrative image-rights deal with the League One club.* **hypothesis**: *Lisa Trundle is in business with the League One club.* | entailment | Yes |
| SST-2 | *his healthy sense of satire is light and fun ...* | positive | No |
| SST-2 | *their healthy sense of satire is light and fun ...* | positive | Yes |
| QNLI | **question**: *How many people lived in Warsaw in 1939?* **sentence**: *Unfortunately this belief still lives on in Poland (although not as much as it used to be)* | not entailment | No |
| QNLI | **question**: *How many women lived in Warsaw in 1939?* **sentence**: *Unfortunately this belief still lives on in Poland (although not as much as it used to be)* | not entailment | Yes |
| QQP | **question 1**: *Do women cheat more than men?* **question 2**: *Do more women cheat than men?* | not duplicate | No |
| QQP | **question 1**: *Do middle-aged women cheat more than men?* **question 2**: *Do more middle-aged women cheat than men?* | not duplicate | Yes |
| CoLA | *John arranged for himself to get the prize.* | acceptable | No |
| CoLA | *Joanne arranged for herself to get the prize.* | acceptable | Yes |
| STSB | **sentence 1**: *Senate confirms Janet Yellen as chair of US Federal Reserve* **sentence 2**: *US Senate Confirms Janet Yellen as New Central Bank Chief* | 4.2 | No |
| STSB | **sentence 1**: *Senate confirms John Yellen as chair of US Federal Reserve* **sentence 2**: *US Senate Confirms John Yellen as New Central Bank Chief* | 4.2 | Yes |

Table 9: Original and perturbed examples from the GLUE tasks.

example relied on the contrast of *aunt* and *uncle*, and even though we concatenated the premise and the hypothesis so coreference across them would be clear, the perturbation still changed the gold label in this hypothetical example.

To get an estimate of how much perturbation actually altered the ground truth classification for our investigated tasks, we ran a pilot hand-validation of a subset of perturber perturbed examples from RTE, CoLA, SST-2, QNLI, QQP.[6] We enlisted one expert annotator and instructed them to label, or validate 25 randomly selected perturbed examples per task, for a total of 125 examples. See Table 9 for examples. The validator labels agreed with the original gold labels for the majority of the examples: 25/25 RTE examples, 25/25 CoLA examples, 25/25 SST-2 examples, 21/25 QNLI examples, and 20/25 QQP examples.

Generally, when the validator label didn't agree with the gold, there was noise in the source data. For example, in QNLI, *In which year did Alexander Dyce bequeathed his books to the museum?* was listed as entailing *These were bequeathed with over 18,000 books to the museum in 1876 by John Forster.*, although the bequeather of the books differs across the two sentences in the source (the perturber only changed "John" to "Jay"). QQP was somewhat of an outlier in our pilot validation, because it has a unexpectedly high proportion of explicit sexual content, which resulted in more drastic semantic changes for the 5 examples the validator disagreed on.

In short, the methodological assumption that demographic perturbation shouldn't alter the gold label seems largely warranted, although we might take the QQP results with a grain of salt. A more in-depth validation round could be performed to confirm our pilot findings.

[6] STS-B was excluded because it is on a 5 point Likert scale that was averaged over several annotators such that many examples have fractional scores. We found it hard with only a single pilot annotation to determine how close was close enough to count as gold label agreement.