Mitigating Label Bias via Decoupled Confident Learning

Yunyi Li [1] **Maria De-Arteaga** [1] **Maytal Saar-Tsechansky** [1]

Abstract

Growing concerns regarding algorithmic fairness have led to a surge in methodologies to mitigate algorithmic bias. However, such methodologies largely assume that observed labels in training data are correct. This is problematic because bias in labels is pervasive across important domains, including healthcare, hiring, and content moderation. In particular, human-generated labels are prone to encoding societal biases. While the presence of labeling bias has been discussed conceptually, there is a lack of methodologies to address this problem. We propose a pruning method—**De**coupled **Co**nfident **Le**arning (DeCoLe)—specifically designed to mitigate label bias. After illustrating its performance on a synthetic dataset, we apply DeCoLe in the context of hate speech detection, where label bias has been recognized as an important challenge, and show that it successfully identifies biased labels and outperforms competing approaches.

1. Introduction

The rapid advancement of AI technologies has sparked important discussions on fairness, equity, and ethics, as AI becomes more integrated into our daily lives. In particular, there has been a surge in the development and application of methodologies to mitigate algorithmic bias. However, most of the methods developed to address algorithmic bias often rely on a critical assumption: the observed labels used to train these systems are accurate. This assumption, while convenient, is often violated, given that bias is often embedded in the training labels, which can result in undesirable consequences that further exacerbate bias (Li et al., 2022).

Label bias refers to a systematic disparity between the ground truth labels intended to train an AI system and the observed labels, such that the relationship underlying the mismatch differs across groups (Li et al., 2022). This type of bias is particularly prominent when labels are generated by humans. For example, African American English (AAE) Twitter posts are disproportionately labeled as toxic by crowd-sourcing annotators, despite those posts being understood as non-toxic by AAE speakers (Sap et al., 2019). In general, such systematic disparity between ground truth and observed labels can result from a wide range of causes, including historical discrimination, human cognitive bias, or social inequalities. For these reasons, label bias has been a growing concern across domains, including but not limited to healthcare (Obermeyer et al., 2019; Philbin & DiSalvo, 1998), criminal justice (Fogliato et al., 2021), and hiring (Hunter et al., 1979).

Exiting methods for mitigating algorithmic bias often operate under the assumption that the observed labels are accurate and reliable for training and evaluation. Particularly, this is true for measures of bias that are derived from the confusion matrix, such as equalized odds (Hardt et al., 2016). However, this assumption can be highly problematic since machine learning models can inadvertently perpetuate existing biases encoded in the biased labels. Procedures for mitigating bias and evaluating model performance with respect to biased observed labels can result in misleadingly positive performance on testing data, thereby exacerbating the existing bias without being aware of it (Li et al., 2022).

The presence of label bias in many datasets across different important domains introduces a new challenge: How can we identify instances in the data that are highly likely to be mislabeled, such that removing those instances can decrease label bias in the data?

A large body of work dealing with noisy labels addresses different challenges related to uncertainty in labels. Such works typically make assumptions about the noise form, such as uniform or class-conditional noise (Angluin & Laird, 1988). The common assumption of class-conditional noise is that label noise depends only on the latent true class. This assumption is commonly used (Goldberger & Ben-Reuven, 2017; Sukhbaatar et al., 2014), and may be reasonable under some circumstances. However, the noise structure can be closely related to group membership, which can lead to label bias. For instance, in the context of toxicity detection, false positive labels may be more common among posts written

[1] University of Texas at Austin, Austin, Texas. Correspondence to: Yunyi Li <Yunyi.Li@mccombs.utexas.edu>.

AI & HCI Workshop at the 40th International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023. Copyright by the author(s).

in AAE. Moreover, instances from different groups can exhibit distinct relationships between covariates and labels—a phenomenon sometimes referred to as differential subgroup validity—, and under such circumstances, predictive models may be dominated by the relationships that hold true for the majority (Chouldechova & Roth, 2020). For example, hateful posts targeting different groups often take very different forms (Gupta et al., 2023). Under such circumstances, there is a potential risk that models that mitigate label noise through a "one size fits all" approach may fail to correctly identify erroneous labels affecting minority groups.

We propose a novel pruning method—Decoupled Confident Learning (DeCoLe)—specifically designed to mitigate label bias. The goal is to identify instances for which the label is likely to be mistaken, so that such instances may be pruned. We estimate the group and class-conditional label uncertainty by training decoupled classifiers (Dwork et al., 2018), and perform group-specific pruning based on the principles of estimating incorrect labels in classification (Northcutt et al., 2021; Elkan, 2001; Forman, 2005). Notably, DeCoLe is a model-agnostic, **data-centric** algorithm, making it a general-purpose framework that can be applied to identify erroneous labels and generate a dataset with reduced bias.

The remaining sections are structured as follows: we briefly review the extant research on algorithmic fairness and label bias, bias in human-generated labels, and noise mitigation in Section 2. We then formalize the problem of label bias mitigation with the presence of asymmetric label errors in Section 3.1, and propose Decoupled Confident Learning (DeCoLe) to prune biased labels in Section 3.2. Next, we illustrate and validate the performance of DeCoLe on a synthetic dataset in Section 4.1, followed by a real-world evaluation in the context of hate speech label bias mitigation in Section 4.2. We conclude the paper with a discussion of future research directions in Section 5.

2. Related literature

Algorithmic fairness and label bias While algorithmic fairness has garnered substantial attention, most bias mitigation strategies and metrics of fairness are primarily concerned with inductive bias, and assume that labels available for training are reliable (Chouldechova & Roth, 2020). For instance, approaches that equalize errors or other metrics derived from the confusion matrix, do so by evaluating performance with respect to the observed label (Mitchell et al., 2018). *Label bias* has increasingly emerged as a concern. However, most work has centered on characterizing or conceptualizing it. Li et al. (2022) provide an overview of different types of label bias within supervised learning systems and empirically demonstrates that collecting more data can exacerbate bias if label bias is overlooked. Fogliato et al. (2020) find that even small biases in observed labels can produce disparate performance across races of recidivism risk assessment tools, and Akpinar et al. (2021) show that differential rates in crime reporting can lead to bias in predictive policing systems. Label bias has also been identified as a potential problem in other contexts such as healthcare (Obermeyer et al., 2019), child maltreatment hotline screenings (De-Arteaga et al., 2021), hiring (Hunter et al., 1979), and offensive language detection (Sap et al., 2019).

Bias in human-generated labels With the rise of crowdsourcing services (Howe, 2008), such as Amazon Mechanical Turk, researchers have noted the risks of annotator cognitive biases (Eickhoff, 2018; Draws et al., 2021) and stereotyping in annotator judgments (Otterbacher, 2015).Expert-generated labels can also reflect biases. For example, in healthcare, the quality of pain assessment and treatment recommendations can be undermined by provider biases (Hoffman et al., 2016). For a comprehensive review of label bias and bias in human-generated labels, please refer to (Li et al., 2022).

Noise mitigation A large stream of work on dealing with noisy labels has proposed aggregating multiple noisy labelers' opinions to reduce the noise in labels (Zhang et al., 2016), as well as learning probabilistic models to jointly estimate labelers' quality and gold standard labels (Snow et al., 2008; Smyth et al., 1994; Dawid & Skene, 1979; Whitehill et al., 2009; Welinder et al., 2010; Yan et al., 2010). Another branch of work focuses on *learning from noisy labels* that do not require labelers' information. Such approaches have investigated training models on noisy datasets through loss reweighting (Shu et al., 2019), surrogate loss (Natarajan et al., 2013), co-teaching (Han et al., 2018) and normalized loss functions (Ma et al., 2020). These approaches tackle the issue of noisy labels by proposing novel model architectures or modifications to the loss function during training. Importantly, methods for learning from noisy labels often assume a particular learning framework and noise structure and cannot be directly adapted to learn any arbitrary model from the data. In particular, they do not consider label bias and assume the noise is either random or solely conditioned on the class.

DeCoLe builds upon a stream of work that estimates incorrect labels in binary classification (Elkan, 2001; Forman, 2005), and closely related to the work by Northcutt et al. (2021), which focuses on estimating label uncertainty using confident learning (CL) (Northcutt et al., 2021). Both CL and DeCoLe are model-agnostic and data-centric, and focus on generating cleaner data. However, it is important to note that the aforementioned works and approaches assume constrained forms of noise such as uniform or class-conditional noise (Angluin & Laird, 1988), excluding the shared societal biases and disregarding fairness considerations during model evaluation. In contrast, DeCoLe specifically addresses the issue of bias in labels, and relaxes the class conditional noise, motivated by the fact that in many cases the noise structure is conditioned on both group and class, as we discussed in Section 1. To the best of our knowledge, DeCoLe is the pioneering pruning method designed to address the group and class-conditioned label noise. By acknowledging this, DeCoLe aims to mitigate label bias and improve the fairness characteristics of the data post-pruning.

3. Methodology

In this section, we propose Decoupled Confident Learning (DeCoLe), a pruning approach to mitigate label bias. We first formally introduce the problem of group and class conditioned noise in section Section 3.1. In section Section 3.2, we describe key steps in DeCoLe, and present the proposed algorithm.

3.1. Preliminaries

In the context of binary classification with possible biased labels, let $\boldsymbol{D} := (\boldsymbol{x}, \tilde{y})^n$ denote the dataset of n examples $\boldsymbol{x}$ with associated observed labels $\tilde{y} \in \{0, 1\}$. We assume there is a group membership indicator $g \in \boldsymbol{x}$, which is typically a pre-defined categorical attribute based on sensitive feature(s). For example, group membership may denote attributes such as age, gender, race, or an intersection of multiple of these.

Let $\boldsymbol{D} := (\boldsymbol{x}, \tilde{y})^n$ be an observed dataset. We suppose there exists a group and class conditional noisy labeling process that results in bias in observed labels $\tilde{y}$. Let y^* be the latent, unbiased labels ("ground truth labels"). For each group g_i, $i \in \{0, ...k\}$ where g_i refers to a specific value of g, let $\pi_{1\text{-}g_i}$ be the fraction of positive instances in group g_i that has been mislabeled as negative, and $\pi_{0\text{-}g_i}$ be the fraction of negative instances in group g_i that has been mislabeled as positive. Formally:

$$\begin{aligned}\pi_{1\text{-}g_i} &= P(\tilde{y} = 0 | y^* = 1, g = g_i)\\ \pi_{0\text{-}g_i} &= P(\tilde{y} = 1 | y^* = 0, g = g_i)\end{aligned}$$

Label bias occurs when there is a disparity in either $\pi_{1\text{-}g_i}$ or $\pi_{0\text{-}g_i}$, or in both, accross different groups g_i. We formulate the problem using a binary classification setting, while allowing multi-categorical group memberships. However, the proposed approach can be extended to a multi-class classification setting. Our goal is to provide a generic approach that can be used to prune erroneous labels in a training dataset, whenever noise is both group and class conditioned. Additionally, we account for the fact that there may be differential subgroup validity, i.e. relationships between the covariates $\boldsymbol{x}$ and the target label y^* may vary across groups. By tackling this, we provide a methodology to mitigate the label bias problem and prevent ML from propagating existing prejudice and inequities present in labels.

Algorithm 1 Decoupled Confident Learning (DeCoLe)

```
Input: Noisy dataset $\boldsymbol{D} := (\boldsymbol{x}, \tilde{y})^n$, group indicator $g$,
initialize a set of classifiers $\{\mathrm{C}_1, ..., \mathrm{C}_k\}$
for $i = 1$ to $k$ do
    Part 1: Estimating $p(\boldsymbol{x})$
    $C_i$.fit($\boldsymbol{x}_{g_i}, \tilde{y}$) where $\boldsymbol{x} \in g_i$
    $\hat{p}(\boldsymbol{x}_{gi}) \leftarrow C_i$.predict_crossval_prob ($\tilde{y} = 1 | \boldsymbol{x}_{g_i}$)
    Part 2: Estimating the thresholds
    $\mathrm{LB}_{g_i} = \mathrm{LB}(y^* = 1, g = g_i) = E_{\boldsymbol{x} \in \tilde{y}=1, g=g_i}[\hat{p}(x)]$
    $\mathrm{UB}_{g_i} = \mathrm{UB}(y^* = 0, g = g_i) = E_{\boldsymbol{x} \in \tilde{y}=0, g=g_i}[\hat{p}(x)]$
    Part 3: Pruning
    Remove $(\boldsymbol{x}_{g_i}, \tilde{y}) \in \boldsymbol{D}$ where $\tilde{y} = 1, \hat{p}(\boldsymbol{x}_{g_i}) < \mathrm{UB}_{g_i}$
    Remove $(\boldsymbol{x}_{g_i}, \tilde{y}) \in \boldsymbol{D}$ where $\tilde{y} = 0, \hat{p}(\boldsymbol{x}_{g_i}) > \mathrm{LB}_{g_i}$
end for
```

3.2. Decoupled Confident Learning (DeCoLe)

We propose Decoupled Confident Learning (DeCoLe). At a high level, DeCoLe tackles the goal of pruning biased labels by training decoupled classifiers for each group $g_i, i \in \{0, .., k\}$, and applying a series of confident learning procedures in parallel, in order to separately identify noise for each group. For each group, the algorithm identifies *pruning thresholds*, which are a set of lower bounds (LB_{g_i}) and upper bounds (UB_{g_i}) of predicted probabilities. The key challenge is to find out the regions where we can confidently determine that any instances within group g_i with predicted probabilities above the LB_{g_i} belong to the positive class, while any instances with predicted probabilities below the UB_{g_i} belong to the negative class. Consequently, we prune instances that are confidently predicted to belong to the positive class, yet have an observed negative label, and vice versa. The main pruning procedure consists of three steps:

1. Train a separate predictive model C_i for each group g_i, and obtain out-of-sample predicted probabilities $\hat{p}(\boldsymbol{x}_{g_i}) = \hat{p}(\tilde{y} = 1; \boldsymbol{x}_{g_i}, C_i)$, where $\boldsymbol{x}_{gi}$ denotes the instances $\boldsymbol{x}$ that belong to group g_i, and $\hat{p}(\tilde{y} = 1; \boldsymbol{x}, C_i)$ is the estimated probability of instance $\boldsymbol{x}$ belonging to class 1 according to the classifier C_i.

2. For each group g_i, estimate the upper bound (UB_{g_i}) threshold and lower bound (LB_{g_i}) threshold that can be used to identify instances that are inferred to have an erroneous observed label. Formally,

$$\begin{aligned}UB_{g_i} &= UB(y^* = 0, g = g_i) = E_{\boldsymbol{x} \in \tilde{y}=0, g=g_i}[\hat{p}(\boldsymbol{x})]\\ LB_{g_i} &= LB(y^* = 1, g = g_i) = E_{\boldsymbol{x} \in \tilde{y}=1, g=g_i}[\hat{p}(\boldsymbol{x})]\end{aligned}$$

3. Prune instances whose observed label $\tilde{y} = 0$, and pre-

dicted probability $\hat{p}(\boldsymbol{x}_{g_i}) \geq \text{LB}_{g_i}$. Analogously, prune instances whose observed label $\tilde{y} = 1$ and predicted probability $\hat{p}(\boldsymbol{x}_{g_i}) \leq \text{UB}_{g_i}$.

The detailed algorithm can be found in Algorithm 1. When there is group and class-conditional noise, decoupling—training individual models for each group—disentangles the noise structure. For each predictive model for g_i, the noise becomes class conditional. At that stage, the theoretical guarantees of confident learning provided by Northcutt et al. (2021) are inherited by each classifier C_i.

4. Experiments

We first test the efficacy of Decoupled Confident Learning (DeCoLe) on a synthetic setting, and show that it improves pruning recall and precision across groups. This allows us to illustrate how the method works and validate its performance in a context where we have full control over the relationship between y^* and $\tilde{y}$. We then apply DeCoLe in the context of hate speech detection, a domain where label bias has been recognized to be an important problem. We compare our approach with the classical Confident Learning (CL) algorithm (Northcutt et al., 2021), and with random sampling (Random). In Section 4.1.1, we motivate and describe in detail the synthetic dataset, and present results on this dataset in Section 4.1.2. In Section 4.2, we present the details of the empirical dataset of hate speech detection, and demonstrate that DeCoLe effectively prunes erroneous instances and mitigates false negatives in hate speech labels, yielding better results than classical CL.

We assess the quality of pruning and the fairness metrics of a cleaned dataset post-pruning. To evaluate the quality of pruning, we focus on measuring the recall and precision of label error detection. Additionally, we examine the remaining label bias in the cleaned dataset by measuring false positive and false negative rates of the observed labels $\tilde{y}$ vs. the ground truth labels y^*.

4.1. Synthetic Experiments

In this section, we present a preliminary validation, showing that DeCoLe effectively identifies erroneous labels and mitigates label bias in a synthetic dataset. We illustrate how, compared to classical confident learning, DeCoLe significantly improves pruning recall, pruning precision, as well as fairness metrics of the cleaned dataset.

4.1.1. Data Generation

We create a synthetic dataset with group and class-conditional noise rates, which allows us to have full control of the relationship between observed labels $\tilde{y}$ and latent ground truth labels y^*.

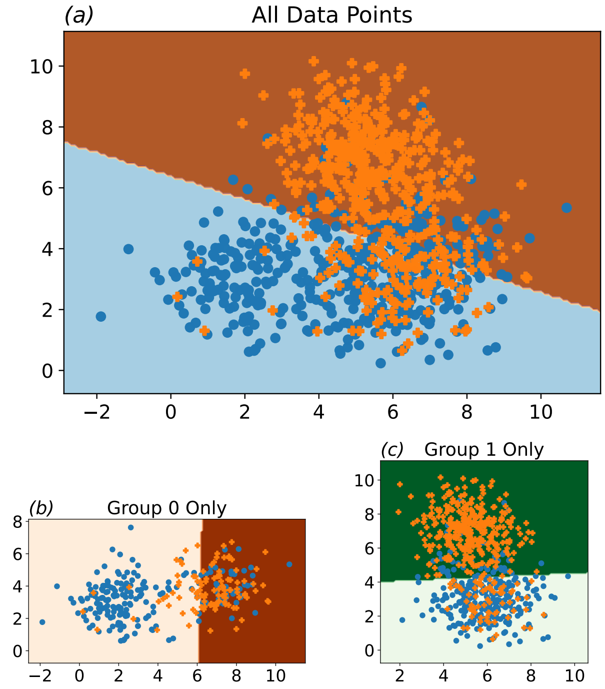

Figure 1. Dataset generated with the group and class-conditional noise. y^* are represented by orange-thickened plus signs and negative instances are denoted by blue-filled circles. Figure (a) encompasses all data points, while (b) only includes instances belonging to g_0, and (c) only includes instances belonging to g_1. Group g_1, which constitutes 70% of the instances, is the majority group. Observed labels for g_0 suffer from a high false negative rate, while those for g_1 have a high rate of false positives.

We introduce a synthetic population consisting of $N = 10000$ instances, each associated with covariates $\boldsymbol{X} \in \mathbb{R}^2$, a binary group membership $g \in \{0, 1\}$, an outcome of interest $y^* \in \{0, 1\}$, and an observed label $\tilde{y} \in \{0, 1\}$. We consider group imbalance, a widely recognized issue in algorithmic fairness (Mitchell et al., 2018), by creating a predominant group ($g = 1$) representing 70% of the total population. To account for differential subgroup validity (Hunter et al., 1979; De-Arteaga et al., 2022), which denotes differences in the relationship between covariates and target labels across groups, we draw instances for different group and class combinations from bi-dimensional normal distributions with different means. We use the same standard deviation for all normal distributions. Further details about the sampling of $\boldsymbol{X}$ and labels $y^* \in \{0, 1\}$ can be found in Appendix A.

We generate observed labels $\tilde{y}$ with group and class-conditional noise, i.e. different error types for different groups. Suppose the positive class represents opportunities or goods, such as job offers. We assume group g_0, the minority group, is more likely to be affected by false negative labels, and group g_1, the majority group, benefits from false

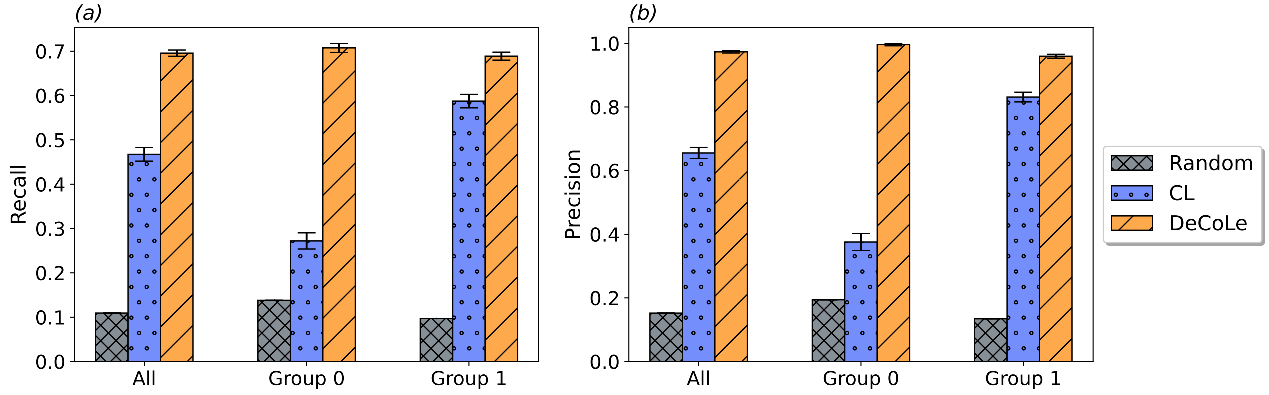

Figure 2. Pruning recall (a) and pruning precision (b) over all instances, group g_0 instances, group g_1 instances. Striped orange bars represent DeCoLe, while dotted blue bars and grid grey bars represent classical confident learning (CL) and random sampling, respectively. DeCoLe significantly outperforms CL in all scenarios, with particularly higher recall and precision for identifying erroneous labels of group g_0, the disadvantaged group.

positive labels. We set $\pi_{1\text{-}g_0} = 0.4$ and $\pi_{0\text{-}g_1} = 0.2$. Additionally, we assume some level of noise for the remaining instances, and set $\pi_{0\text{-}g_0} = 0.05$ and $\pi_{1\text{-}g_1} = 0.05$.

This simulation therefore consists of four clusters depicted in Figure 1 (b) and (c), where positive observed labels are represented by orange thickened plus signs and negative observed labels are denoted by blue filled circles. Figure 1 (b) corresponds to group g_0 instances, while Figure 1 (c) corresponds to group g_1 instances. Combining Figure 1 (b) and (c) together, we get the full picture of the four clusters in Figure 1 (a). Assuming we were to use a linear classifier to differentiate between the two classes, it is not hard to find that linear classifiers for group g_0 only (Figure 1 (b)) and for group g_1 only (Figure 1 (c)) exhibit fundamental dissimilarity, being nearly orthogonal. Additionally, when we fit one linear classifier for both groups, as depicted in Figure 1 (a), it demonstrates differential subgroup validity, wherein its predictive accuracy is notably higher for the majority group compared to the minority group. Furthermore, the linear classifier in Figure 1 (a) tends to misclassify positive group g_0 instances as negative and negative group g_1 instances as positive, reflecting how the predictive model may lean and amplify bias in the data labels. In the same way, this affects models used for pruning, and, as we show, results in poor pruning performance when we do not consider the group-specific nature of predictive relationships and label errors.

4.1.2. Results and Analysis

We apply DeCoLe framework described in Algorithm 1 on the dataset generated in Section 4.1.1. We also consider the CL alternative as a baseline and include the performance of random pruning as reference. We use logistic regression as a base model, and generate 95% confidence bounds via 5 runs on different seeds.

Figure 2 (a) and (b) illustrate the pruning recall and precision, respectively. For each, we can assess these metrics overall, for group g_0, and for group g_1. Striped orange bars represent DeCoLe framework, while dotted blue bars and grid grey bars represent classical confident learning (CL) and random sampling, respectively. Figure 2 clearly demonstrates that *DeCoLe significantly outperforms CL in all scenarios, with particularly remarkable higher accuracy in correctly identifying erroneous labels of group g_0, the disadvantaged group.*

For a more nuanced view, Figure 3 shows the pruning precision and recall for different error types over the two groups. As we described in Section 4.1.1, group g_0 mainly suffers from false negatives, while the main error for group g_1 labels is false positives. According to Figure 3 (a) and (b), DeCoLe significantly improves recall and precision on identifying false negatives (FN) for group g_0. Additionally, according to Figure 3 (c), DeCoLe significantly improves recall rate on identifying false positives (FP) for group g_1. The improvement on FN recall and precision for group g_0 is especially large, which is important from a fairness perspective. Furthermore, according to Figure 3 (d), CL produces extraordinarily low FP precision for group g_0. There are very few (only 5%) false positives in group g_0; such a low precision means that the CL algorithm pruned many positive instances in group g_0 even though those instances are true positives, which would further exacerbate the existing bias in the data.

The central goal of pruning is to effectively detect inaccurate labels in order to yield a much cleaner dataset, preventing bias propagation through ML models trained on data with different error types for different groups. Figure 4 depicts the quality of the label $\tilde{y}$ with respect to the label y^* in the

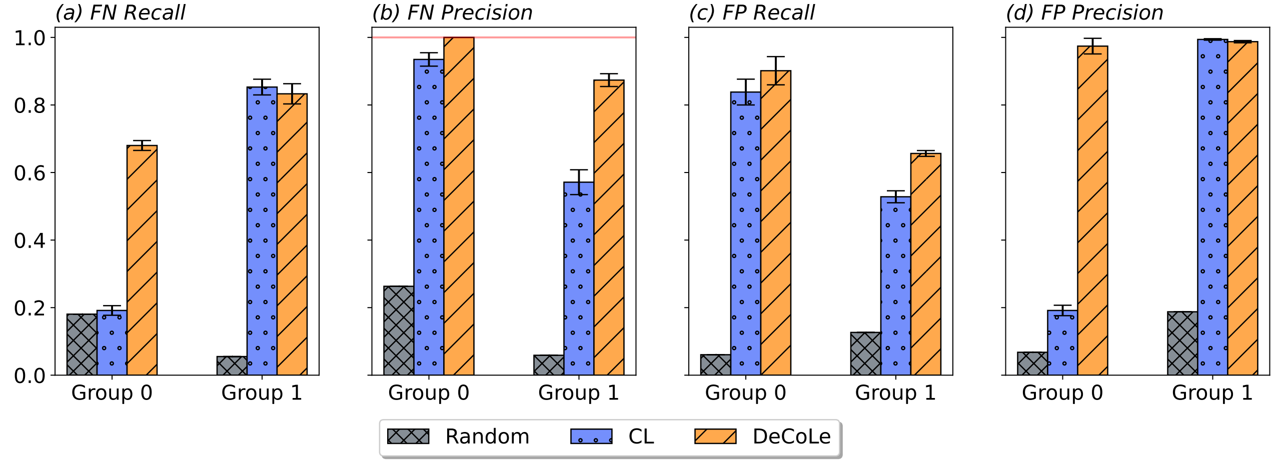

Figure 3. Pruning recall and precision of different error types over group g_0 instances and group g_1 instances: (a) false negatives recall; (b) false negatives precision; (c) false positives recall; (d) false positives precision. Striped orange bars represent DeCoLe, while dotted blue bars and grid grey bars represent classical confident learning (CL) and random sampling, respectively. According to (a) and (b), DeCoLe significantly improves recall and precision on identifying false negatives for group g_0. According to (c), DeCoLe significantly improves recall rate on identifying false positives for group g_1. The improvement on FN recall and FP precision for group g_0 is especially large.

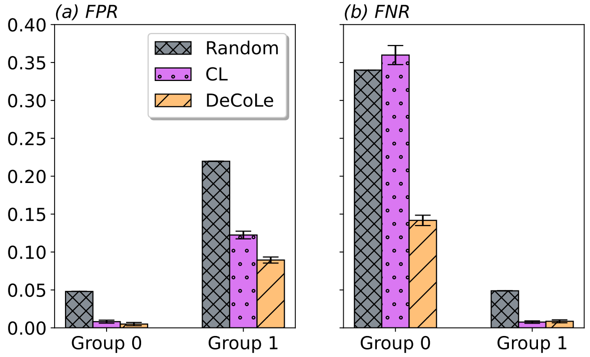

Figure 4. (a) False Positive Rates (FPR) and (b) False Negative Rates (FNR) over group 0 and group 1 instances of data after pruning employing random sampling (grid grey bars), CL algorithm (dotted orchid vars), and DeCoLe (striped pastel orange bars). DeCoLe substantially more capably mitigates both error types (false positives for group 1 and false negatives for group 0).

cleaned data. Figure 4 (a) shows the FPR, the proportion of actual negatives (y^*) incorrectly labeled as positives ($\tilde{y}$); Figure 4 (b) shows the FNR, the proportion of actual positives (y^*) incorrectly labeled as negatives ($\tilde{y}$). The results show the errors in the data post-pruning, when pruning is done employing random sampling (grid grey bars), CL algorithm (dotted orchid bars), and DeCoLe (striped pastel orange bars). As in the other cases, the results are shown for both group g_0 and group g_1 instances. As shown in Figure 4 (b), CL yields worse-than-random performance for the disadvantaged group, g_0. Across both plots, it is evident that *DeCoLe is substantially more capable of mitigating the two most prominent error types (false positives for group g_1 and false negatives for group g_0) and thus is more suitable for preventing systematic bias present in labels.*

4.2. Hate Speech Label Bias Mitigation

Hate speech causes significant harm. It is used to radicalize and recruit within extremist groups, incite violence, and even genocide (Kennedy et al., 2020). However, labeling hate speech is challenging given that judgments of offensiveness depend on societal circumstances (Sap et al., 2019). Extremist groups also intentionally make their hate speech obscure to evade detection (Kennedy et al., 2020). Crowdsourced annotations used for training automatic hate speech detection systems are prone to bias from varied annotator knowledge and perspectives in what constitutes hate speech (Davani et al., 2022). What one person labels as hateful, another may see as benign, yielding conflicting labels for the same data (Kennedy et al., 2020; Davani et al., 2022). Although majority votes are commonly used to aggregate multiple opinions, normative stereotypes embedded in society and homogeneity of the annotators' bias can easily lead to systematic labeling errors (Davani et al., 2023; 2022).

Recently, Kennedy et al. (2020) proposed a novel method based on Rasch Measurement Theory (RMT) to construct a less biased hate speech measure. Their measure articulates hate speech theoretically across eight dimensions (incite violence, humiliate, etc.), capturing the complexity of hate speech and limiting bias from oversimplification. Furthermore, by evaluating inter-rater reliability, they are able to remove inconsistent raters, correcting human judgment biases and promoting reliability. The researchers assessed

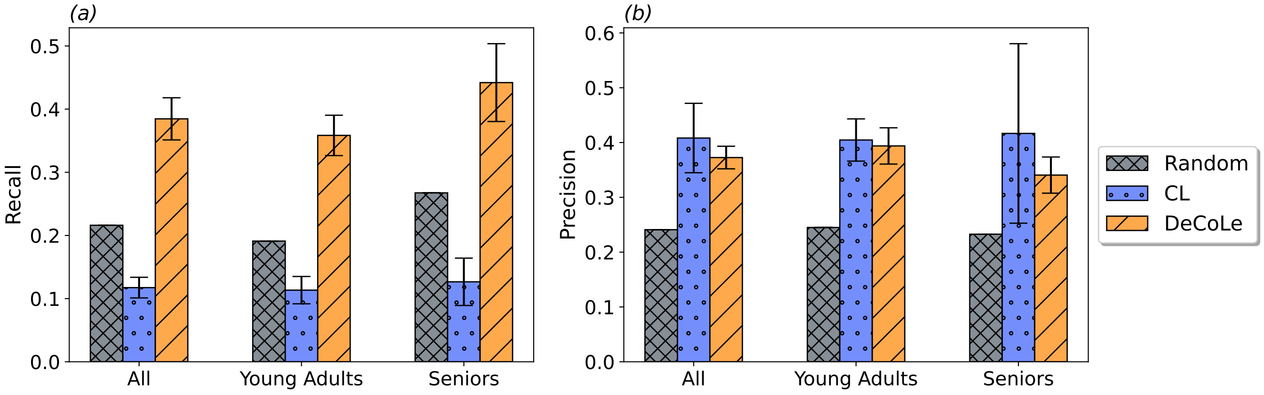

Figure 5. In the context of hate speech detection, pruning (a) recall rates and (b) precision rates over all instances, posts targeting young adults, and posts targeting seniors. Striped orange bars represent DeCoLe, while dotted blue bars and grid grey bars represent classical confident learning (CL) and random sampling, respectively. DeCoLe significantly outperforms CL in recall rates while yielding comparable precision rates.

the validity and reliability of their proposed measurement approach and found that it demonstrates high internal consistency, test-retest reliability, and construct validity. In summary, Kennedy et al. (2020) limits bias in labels for hate speech, but it involves a significantly more costly labeling process, as it requires labels for each instance across eight different dimensions, rather than one. The dataset also has the advantage of collecting the more common labels used for hate speech (directly asking if a post constitutes hate speech), and includes demographic information about the target group of the posts. Thus, the data contains a label $\tilde{y}$ and an improved label y^* that we leverage to assess the performance of DeCoLe in a real-world dataset from an impactful domain.

4.2.1. Empirical Results

To empirically validate that DeCoLe effectively identifies erroneous labels and mitigates label bias in data used to train hate speech detection systems, we utilized the newly generated RMT-based hate speech measure as our gold standard labels, y^*. The observed labels $\tilde{y}$, which may contain biases, were obtained from a single hate speech survey item (Kennedy et al., 2020; Sachdeva et al., 2022). We conducted this validation specifically in the context of hate speech targeting two groups: young adults and seniors. We use random forest as the base model, and to ensure robustness, we performed five runs with different random seeds to obtain a 95% confidence bound. Our empirical findings demonstrate that DeCoLe outperforms CL algorithm by significantly improving the recall rate of erroneous labels and more effectively mitigating false negatives for both groups.

Figure 5 (a) and (b) show the pruning recall (a) and pruning precision (b) for all posts, posts targeting young adults, and posts targeting seniors. The striped orange bars represent the results from DeCoLe, while the dotted blue bars and grid grey bars represent CL approach and random sampling, respectively. It can be observed that while both CL and DeCoLe yield comparable precision, DeCoLe demonstrates significant improvement in recall compared to both CL and random sampling.

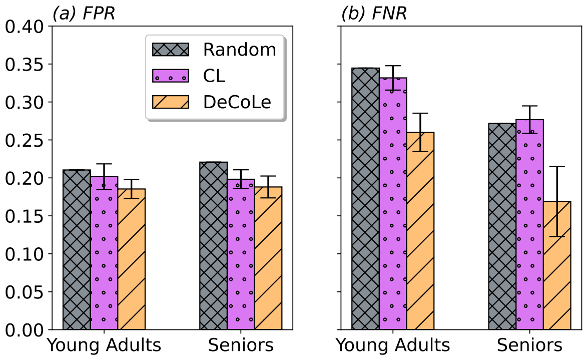

Figure 6. (a) False positive rates and (b) false negative rates of the dataset after pruning using random sampling (grid grey bars), CL algorithm (dotted orchid bars), and DeCoLe (striped pastel orange bars) framework. The grid grey bars could be understood as representing the error rates of the dataset before pruning. According to (b), DeCoLe significantly reduces false negatives for both posts targeting young adults and posts targeting Seniors, surpassing the performance of CL algorithm.

Furthermore, Figure 6 contains information about the quality and fairness of the cleaned dataset after pruning with different methods. Figure 6 (a) displays the false positive rates and Figure 6 (b) the false negative rates of observed labels in the cleaned dataset after pruning using random sampling (grid grey bars), CL (dotted orchid bars), and DeCoLe algorithm (striped pastel orange bars). The results are disaggregated for posts targeting young adults, and posts targeting seniors. Note that random sampling maintains the same rate of errors, thus, the grid grey bars can also be understood as representing the error rates of the dataset before pruning. From Figure 6 (b), it can be observed that DeCoLe significantly reduces false negatives for both posts targeting young adults and posts targeting seniors, surpassing the performance of CL algorithm.

5. Conclusion

While there is a growing awareness of the presence of label bias in supervised learning systems, particularly those used to guide high-stake decisions, methods that are specifically designed for mitigating label bias remain insufficient. To address this pressing issue, we propose a novel approach called Decoupled Confident Learning (DeCoLe), a pruning method that mitigates label bias. Specifically, DeCoLe improves upon existing noise-mitigation alternatives by accounting for the fact that noise may be group- and class-conditioned. This type of label bias arises when the likelihood that a label is incorrect is influenced by both the group membership and the ground truth class. For instance, in the context of hate speech, it has been shown that labelers' assessment of hate speech depends on the stereotypes they have about a given group (Davani et al., 2023). Our experimental results, which focus on the hate speech domain, validate the effectiveness of DeCoLe in pruning erroneous instances and mitigating group-specific false negatives associated with hate speech labels.

Future research endeavors should focus on the development of methodologies capable of handling other forms of label bias structures. While DeCoLe has primarily focused on mitigating group and class-conditional noise, there is a need for novel methodologies that can address label noise patterns conditioned on other covariates. By extending existing approaches to encompass diverse sources of label noise, researchers can advance the field's understanding and ability to mitigate biases arising from a wider range of factors.

6. Acknowledgements

This research was supported by the Machine Learning Laboratory[1], and Good Systems[2], a UT Austin Grand Challenge to develop responsible AI technologies.

References

Akpinar, N.-J., De-Arteaga, M., and Chouldechova, A. The effect of differential victim crime reporting on predictive policing systems. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 838–849, 2021.

Angluin, D. and Laird, P. Learning from noisy examples. *Machine Learning*, 2:343–370, 1988.

Chouldechova, A. and Roth, A. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

Davani, A. M., Díaz, M., and Prabhakaran, V. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.

Davani, A. M., Atari, M., Kennedy, B., and Dehghani, M. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319, 2023.

Dawid, A. P. and Skene, A. M. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

De-Arteaga, M., Jeanselme, V., Dubrawski, A., and Chouldechova, A. Leveraging expert consistency to improve algorithmic decision support. *arXiv preprint arXiv:2101.09648*, 2021.

De-Arteaga, M., Feuerriegel, S., and Saar-Tsechansky, M. Algorithmic fairness in business analytics: Directions for research and practice. *Production and Operations Management*, 31(10):3749–3770, 2022.

Draws, T., Rieger, A., Inel, O., Gadiraju, U., and Tintarev, N. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, pp. 48–59, 2021.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*, pp. 119–133. PMLR, 2018.

Eickhoff, C. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pp. 162–170, 2018.

Elkan, C. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pp. 973–978. Lawrence Erlbaum Associates Ltd, 2001.

Fogliato, R., Chouldechova, A., and G'Sell, M. Fairness evaluation in presence of biased noisy labels. In *International Conference on Artificial Intelligence and Statistics*, pp. 2325–2336. PMLR, 2020.

[1] https://ml.utexas.edu/

[2] http://goodsystems.utexas.edu/

Fogliato, R., Xiang, A., Lipton, Z., Nagin, D., and Chouldechova, A. On the validity of arrest as a proxy for offense: Race and the likelihood of arrest for violent crimes. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 100–111, 2021.

Forman, G. Counting positives accurately despite inaccurate classification. In *Machine Learning: ECML 2005: 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005. Proceedings 16*, pp. 564–575. Springer, 2005.

Goldberger, J. and Ben-Reuven, E. Training deep neural-networks using a noise adaptation layer. In *International conference on learning representations*, 2017.

Gupta, S., Lee, S., De-Arteaga, M., and Lease, M. Same same, but different: Conditional multi-task learning for demographic-specific toxicity detection. In *Proceedings of the ACM Web Conference 2023*, pp. 3689–3700, 2023.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Hoffman, K. M., Trawalter, S., Axt, J. R., and Oliver, M. N. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences*, 113(16):4296–4301, 2016.

Howe, J. *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House, 2008.

Hunter, J. E., Schmidt, F. L., and Hunter, R. Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin*, 86(4):721, 1979.

Kennedy, C. J., Bacon, G., Sahn, A., and von Vacano, C. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*, 2020.

Li, Y., De-Arteaga, M., and Saar-Tsechansky, M. More data can lead us astray: Active data acquisition in the presence of label bias. *arXiv preprint arXiv:2207.07723*, 2022.

Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., and Bailey, J. Normalized loss functions for deep learning with noisy labels. In *International conference on machine learning*, pp. 6543–6553. PMLR, 2020.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

Northcutt, C., Jiang, L., and Chuang, I. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Otterbacher, J. Crowdsourcing stereotypes: Linguistic bias in metadata generated via gwap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1955–1964, 2015.

Philbin, E. F. and DiSalvo, T. G. Influence of race and gender on care process, resource use, and hospital-based outcomes in congestive heart failure. *The American journal of cardiology*, 82(1):76–81, 1998.

Sachdeva, P., Barreto, R., Bacon, G., Sahn, A., Von Vacano, C., and Kennedy, C. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pp. 83–94, 2022.

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1668–1678, 2019.

Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., and Meng, D. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32, 2019.

Smyth, P., Fayyad, U., Burl, M., Perona, P., and Baldi, P. Inferring ground truth from subjective labelling of venus images. *Advances in neural information processing systems*, 7, 1994.

Snow, R., O'connor, B., Jurafsky, D., and Ng, A. Y. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pp. 254–263, 2008.

Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., and Fergus, R. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.

Welinder, P., Branson, S., Perona, P., and Belongie, S. The multidimensional wisdom of crowds. *Advances in neural information processing systems*, 23, 2010.

Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J., and Ruvolo, P. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22, 2009.

Yan, Y., Rosales, R., Fung, G., Schmidt, M., Hermosillo, G., Bogoni, L., Moy, L., and Dy, J. Modeling annotator expertise: Learning when everybody knows a bit of something. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 932–939. JMLR Workshop and Conference Proceedings, 2010.

Zhang, J., Wu, X., and Sheng, V. S. Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review*, 46:543–576, 2016.

A. Appendix: Synthetic Data Generation

We describe the details of the simulated 4 clusters for different class and group combinations here.

We introduce a bi-dimensional ($\boldsymbol{X} \in \mathbb{R}^2$) synthetic population ($N$ = 10000) divided into two groups ($g \in \{0, 1\}$), and assume group 1 is the majority group that account for 70% of the population. Specifically, for group 0, we sample the attributes of class 0 instances $\boldsymbol{X}_{g=0,y^*=1} \in \mathbb{R}^2$ from normal distribution $\mathcal{N}((\mu_{x_1} = 2, \mu_{x_2} = 3), \sigma^2)$, and sample the attributes of class 1 instance $\boldsymbol{X}_{g=0,y^*=0} \in \mathbb{R}^2$ from normal distribution $\mathcal{N}((\mu_{x_1} = 7, \mu_{x_2} = 4), \sigma^2)$. Similarly, we sample the attributes of group 1 class 0 instance $\boldsymbol{X}_{g=1,y=0} \in \mathbb{R}^2$ from $\mathcal{N}((\mu_{x_1} = 6, \mu_{x_2} = 3), \sigma^2)$, and sample group 1 class 1 instance $\boldsymbol{X}_{g=1,y=1} \in \mathbb{R}^2$ from normal distribution $\mathcal{N}((\mu_{x_1} = 5, \mu_{x_2} = 7), \sigma^2)$. We set $\sigma = 1.2$ for all distributions.