Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/2025.29219

A Comparative Analysis of Pre-processed Data Debiasing of a Mathematical and SimSeed Approach

Claire Zhou[1], Xiangyu Yang[2*], Jake Hu[3]

[1]Torrey Pines High School, San Diego, USA
[2]ShenZhen RDF School, Shenzhen, China
[3]Aragon High School, San Mateo, USA
**Corresponding Author. Email: 32761344@qq.com*

Abstract. Bias in machine learning datasets and models can pose significant challenges to achieving fairness in real-world applications. In this paper, we look over two methods aimed at mitigating bias in machine learning datasets: “Identifying and Correcting Label Bias in Machine Learning” and “Debiasing made state-of-the-art Revisiting the Simple Seed-based Weak Supervision for Text Classification”. The first method utilizes a mathematical approach that re-weights training samples, addressing label bias by integrating fairness constraints directly for the optimization process. Such examples include demographic parity and equalized odds; iterative training and adjustments with fairness violation penalties establish a balance between accuracy and fairness. The second method presents seed deletion in weak supervision as a way to minimize bias in text classification tasks. By removing specific seed words from pseudo-labeled texts, and data augmentation via random deletion, the model reduces the overreliance on biased features, which improves robustness and generalization. Overall, we evaluated that these methods can achieve improvements in fairness and accuracy across diverse data sets and domains which include: crime prediction, credit scoring, and text classification. Our paper highlights the potential of combining advanced mathematical techniques with preprocessing to mitigate bias in machine learning.

Keywords: Weak Supervision, Label Bias, SimSeed, Mathematical Approach, Fairness

1. Introduction

In a world where bias is a growing concern, datasets, too, also have their own prejudice. In this review paper, we aim to compare the methods and results of two papers that preprocess data to mitigate bias: Identifying and Correcting Label Bias in Machine Learning—which re-weights labels or implicit weak supervision—and: Debiasing Made State-of-the-art Revisiting the Simple Seed-based Weak Supervision for Text Classification—which uses weak supervision and seed deletion [1].

Weak supervision is a machine-learning technique designed to address labeled data by assembling noisy estimates of labels from multiple sources, modeling inaccuracies, and combining them into high-quality pseudo labels for downstream tasks. However, despite success in various applications, weak supervision needs the perspective of fairness. It has been demonstrated through past research that weak supervision can produce unfair and biased outcomes, even when perfectly fair labels are achieved with ground truth labels. This motivates the need for more fairly specific approaches in weak supervision. As a result, this work creates a counterfactual, fairness-based technique that mitigates biases inherent in weak supervision by addressing unequal accuracy of label sources in groups (leading to improvements in fairness and accuracy). In addition, weak supervision models have raised concerns within real-world applications. Past studies exemplify the biases in past training data that often become reinforced with models that lead to unfair outcomes.

2. Methodology

2.1. Mathematical approach to reduce bias from Identifying and Correcting Label Bias in Machine Learning

The paper Identifying and Correcting Label Bias in Machine Learning employs a mathematical approach to reduce the likelihood of bias when handling data. Its core method—re-weighting data—works to correct biases by approximating what the true labels would be if there were no inherent bias in the data [1].

Here are the primary steps of the process:

1. Mathematical Formulation of Label Bias: The authors propose a model in which biased labels result from an agent aiming for accuracy but affected by inherent biases. They demonstrate that this bias can be expressed mathematically, which allows for adjustments to the weights of training samples without altering the labels themselves. Also, the author assumes there is a true label in the dataset, and it can be represented from a transformation of this function. The Kullback-Leibler divergence can measure the difference between the observed error labels and the assumed real labels.

2. Re-weighting Mechanism: This key methodology applies derived formulas to re-weight data points, thereby approximating training on an unbiased version of the dataset. This ensures that models trained on these re-weighted samples align more closely with an unbiased classifier. When it comes to the advantages of using this method, it is compatible with the common algorithm we use daily.

3. Fairness Constraints: The approach incorporates various fairness concepts, such as demographic parity and equal opportunity, directly into the optimization process through a penalty system. The authors also compare this method to traditional post-processing fairness adjustments and alternative techniques like the Lagrangian approach. Specifically, the fairness constraints achieve its objective by introducing a penalty system. This can make sure the model can take fairness as a factor while training. The article also compares the traditional method and the one we are using now, showing how this approach can fulfill the requirements without sacrificing accuracy.

4. Iterative Training: The model iteratively updates both the weights and parameters based on any fairness constraint violations observed, continuing until a more balanced model is achieved. In this case, the experiments on several commonly used datasets, result shows that this technique can achieve a better balance between fairness and accuracy.

5. Empirical Evaluation: The method is tested on standard fairness benchmark datasets, with results showing that it often strikes a better balance between fairness and accuracy than other techniques.

In essence, this approach involves re-weighting data points through a derived formula that approximates training on unbiased labels. By directly integrating fairness constraints into the optimization process, this method enables the model to better balance fairness and accuracy through iterative adjustments based on fairness violations.

2.2. Simple Seed-based (SimSeed) approach to minimizing bias with seed removal

The methodology used in the essay Debiasing Made State-of-the-art: Revisiting the Simple Seed-based Weak Supervision for Text Classification focuses on debiasing the model for simple seed-based weak supervision for text classification. The core of its methodology is identifying the seeds and the deletion of the texts [2].

2.2.1. Seed matching

Seed matching is a simple form of weak supervision for text classification. There are pseudo-labels generated based on specific seed words within the text, and the texts are automatically labeled if they contain the seed words associated with that label. Though this method seems to be simple, the most disadvantage is the overreliance on the seed words, which makes the method susceptible to label bias.

2.2.2. Seed deletion

Seed deletion is the method introduced by the paper to mitigate the disadvantage of seed matching. Here, once a text is pseudo-labeled using seed words, those seed words are deliberately removed from the text before it is used for training the classifier. Thus, when the classifier is trained, it will not be affected by the seed words but will be encouraged to make generalizations of the text from features other than those specific seed words.

2.2.3. Random deletion

Besides seed deletion, random deletion is also implemented to further improve the model’s robustness. Since we cannot always know which words are seed words, randomly deleting words from the texts at a specified deletion ratio can help improve the model’s performance. The model will not only reduce its dependence on any specific words but will also learn to handle inputs with missing information, thereby improving robustness and reducing overfitting.

By integrating these methods, the paper aims to show that even relatively simple adjustments like deleting words can significantly improve the effectiveness and reduce the bias of models.

3. Results

3.1. Experiment results based on fairness tasks from Identifying and Correcting Label Bias in Machine Learning

The authors apply their method to real-world datasets and scenarios. To begin, they use datasets from Bank Marketing to predict if a person will buy a bank account with five protected groups; Communities and Crime to predict if a community has above 70-th percentile crime rate with four race features; ProPublicas COMPAS to predict recidivism with criminal history, jail and prison time, demographic, and risk scores; German Statlog Credit Data to predict if a person is a well or poor credit risk based on financial situation; and Adult to predict if a person’s income is greater or less than 50k every year based on four protected groups [3-7] . The mathematical approach targets fairness in the areas of including demographic parity (Dem. Par.), equal opportunity (Eq. Opp.), equalized odds (Eq. Odds), and disparate impact (Disp. Imp.).

In a nutshell demographic parity means that positive outcomes must be equal, while equal opportunity means that false negative is equal and equalized odds means that both false positive and false negative rates are equal. These three notions of fairness must apply to all demographics. On the other hand, disparate impact is a test to determine if an event has a disproportionate effect on one group.

Table 1. Experiment results. The authors demonstrate the usefulness of their method (Our) compared to more commonly used approaches of debiasing: Lagranian (Lagr.), post-processing calibration (Cal.), and a control group with no constraints (Unc.). The bolded numbers denote the method that has the fairness violation (Vio.)

Dataset	Metric	Unc. Err.	Unc. Vio.	Cal. Err.	Cal. Vio.	Lagr. Err.	Lagr. Vio.	Our Err.	Our Vio.
Bank	Dem. Par.	9.41%	.0349	9.70%	.0068	10.46%	.0126	9.63%	.0056
	Eq. Opp.	9.41%	.1452	9.55%	.0506	9.86%	.1237	9.48%	.0431
	Eq. Odds	9.41%	.1452	N/A	N/A	9.61%	.0879	9.50%	.0376
	Disp. Imp.	9.41%	.0304	N/A	N/A	10.44%	.0135	9.89%	.0063
COMPAS	Dem. Par.	31.49%	.2045	32.53%	.0201	40.16%	.0495	35.44%	.0155
	Eq. Opp.	31.49%	.2373	31.63%	.0256	36.92%	.1141	33.63%	.0774
	Eq. Odds	31.49%	.2373	N/A	N/A	42.69%	.0566	35.06%	.0663
	Disp. Imp.	31.21%	.1362	N/A	N/A	40.35%	0.499	42.64%	.0256
Communities	Dem. Par.	11.62%	.4211	32.06%	.0653	28.46%	.0519	30.06%	.0107
	Eq. Opp.	11.62%	.5513	17.64%	.0584	28.45%	.0897	26.85%	.0833
	Eq. Odds	11.62%	.5513	N/A	N/A	28.46%	.0962	26.65%	.0769
	Disp. Imp.	14.83%	.3960	N/A	N/A	28.26%	0.557	30.26%	.0073
German Stat.	Dem. Par.	24.85%	.0766	24.85%	.0346	25.45%	.0410	25.15%	.0137
	Eq. Opp.	24.85%	.1120	24.54%	.0922	27.27%	.0757	25.45%	.0662
	Eq. Odds	24.85%	.1120	N/A	N/A	34.24%	.1318	25.45%	.1099
	Disp. Imp.	24.85%	.0608	N/A	N/A	27.57%	.0468	25.15%	.0156
Adult	Dem. Par.	14.15%	.1173	16.60%	.0129	20.47%	.0198	16.51%	.0037
	Eq. Opp.	14.15%	.1195	14.43%	.0170	19.67%	.0374	14.46%	.0092
	Eq. Odds	14.15%	.1195	N/A	N/A	19.04%	.0160	14.58%	.0221
	Disp. Imp.	14.19%	.1108	N/A	N/A	20.48%	.0199	17.37%	.0334

While other methods outperformed the paper's approach, a majority of lowest fairness violation is attributed to the mathematical model for decreasing labeling bias. The authors found that Lagrangain often succumbed to overfitting, causing unfairness, oftentimes having the largest fairness violation.

Separately, the authors compare a method trained on true labels, Unc., Cal., Lagr., and their own method of finding suitable weights for a classifier to predict the digit 2 with the dataset MNIST.

Table 2. Comparing accuracy of MNIST with label bias

Method	Test Accuracy
Trained on True Labels	97.85%
Unconstrained	88.18%
Calibration	89.79%
Lagrangian	94.05%
The author’s method	96.16%

In this experiment, the authors find that their method is the closest to training on true labels, where the second-best method—Lagrangian—improves the error rate by 30%.

3.2. Experiment results from SimSeed deletion from debiasing made state-of-the-art: revisiting the simple seed-based weak supervision for text classification [2]

As the authors focused on text classification, they used datasets from the New York Times (NYT), 20 Newsgroups (20News), AGNews, Rotten tomatoes as well as other papers [2, 3].

The authors compared their methods of SimSeed using the following benchmarks, while train for four epochs to follow LOPS:

- Vanilla: a base model using SimSeed without any other bias minimizing approaches, which can also be used for image detection.
- Standard confidence: assigns a confidence score to each predicted label based on true label
- O2U-Net: a neural net architecture designed to detect out-of-distribution and bias [4].
- LOPS: a structure to assess label quality and identify label bias based on label reliability [1].
- Oracle: a maximum, or upper bound, for performance in a hypothetical situation where labels are perfect, accurate, and unbiased.

Another option for seed debiasing is using MLM-replace—by randomly replacing words with BERT predictions—or Paraphrase—which paraphrases a whole document.

Table 3. Main results. This demonstrates different pseudo-label methods compared to one another over multiple datasets. The authors show Macro-F1 and Micro-F1 to evaluate accuracy

	AG News		20News-Coarse		NYT-Coarse		20News-Fine		NYT-Fine		Rotten-Tomatoes	
Method	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1
Oracle	86.3(0.3)	86.2(0.3)	90.6(0.4)	90.6(0.4)	97.0(0.2)	93.9(0.4)	81.1(0.2)	81.1(0.3)	96.5(0.1)	92.0(0.4)	79.6(1.0)	79.6(1.0)
Vanilla	83.9(0.6)	83.9(0.6)	80.5(0.6)	80.1(0.6)	87.8(0.2)	78.4(0.4)	68.2(0.6)	69.0(0.7)	73.0(0.7)	68.7(0.8)	71.4(1.2)	71.3(1.2)
Standard Confidence	82.2(2.1)	82.0(2.1)	78.9(1.8)	80.0(1.5)	88.3(4.1)	79.1(2.8)	64.4(2.2)	66.6(1.8)	45.8(0.5)	58.6(0.3)	72.2(2.4)	72.0(2.4)
O2U-Net	79.8(0.5)	79.8(0.5)	80.9(0.3)	78.5(0.2)	92.9(0.4)	85.9(0.7)	71.1(0.4)	71.2(0.8)	14.7(10.2)	8.70(7.3)	74.1(1.4)	74.0(1.4)
LOPS	79.5(0.9)	79.5(0.6)	81.7(1.0)	80.7(0.4)	94.6(0.4)	88.4(0.5)	73.8(0.6)	72.7(1.0)	84.3(0.5)	81.6(0.3)	70.4(0.4)	70.4(0.4)
Seed-Deletion	84.3(0.7)	84.2(0.7)	86.4(0.9)	86.1(0.8)	92.4(1.3)	85.0(2.0)	73.7(0.7)	75.0(0.5)	81.7(1.5)	79.4(1.1)	70.4(1.3)	70.3(1.3)

Random-Deletion	86.2(0.5)	86.1(0.5)	84.4(0.9)	84.8(0.8)	91.7(1.3)	83.3(1.8)	76.3(0.8)	76.8(0.7)	84.6(1.4)	79.6(1.1)	73.6(4.3)	73.4(4.5)
Paraphrase	85.4(0.3)	85.4(0.3)	86.9(1.2)	86.7(1.2)	94.0(0.8)	88.5(0.7)	75.6(0.6)	76.8(0.5)	76.1(4.0)	74.8(2.4)	75.4(0.1)	75.3(0.1)
MLM-Replace	85.8(0.1)	85.8(0.1)	87.4(0.1)	87.5(0.2)	94.5(0.1)	88.9(0.2)	73.6(1.0)	74.8(0.8)	84.1(0.6)	80.0(0.4)	76.7(1.5)	76.7(1.5)

As shown in Table 3, random deletion and seed deletion often have the highest F1 scores; for example, in some cases such as AGNews, they are almost equal to the F1 score of Oracle. They infer that seed deletion is more successful, in part, due to heavy data augmentation.

The authors also used weak supervision text classifiers and reported its Micro and Macro F1 scores. They used ConWea, X-Class, LOTClass, ClassKG, and LIME [3].

Table 4. F1 accuracy of text classification using weak supervision models compared to seed deletion and random seed deletion. The authors chose not to use the Rotten Tomatoes dataset since it was not used in the cited papers

	AGNews		20News-Coarse		NYT-Coarse		20News-Fine		NYT-Fine	
Method	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1
ConWea	73.4	73.4	74.3	74.6	93.1	87.2	68.7	68.7	87.4	77.4
X-Class	82.4	82.3	58.2	61.1	96.3	93.3	70.4	70.4	86.6	74.7
LOTClass	84.9	84.7	47.0	35.0	70.1	30.3	12.3	10.6	5.3	4.1
ClassKG	99.9	-	80	75	96	83	78	77	92	80
LIME	87.2	87.2	79.7	79.6	-	-	-	-	-	-
Seed Deletion	84.3	84.2	86.4	86.1	92.4	85.0	73.7	75.0	81.7	79.4
Random Deletion	86.2	86.1	84.4	84.8	91.7	83.3	76.3	76.8	84.6	79.6

While most weak supervision models performed better than seed deletion, the author suggests that SimSeed is still a competitor, especially in 20News, is oftentimes almost as good as other models.

4. Conclusion

As fairness becomes a growing concern in our society today, the authors of both papers set forth solutions to debias data. In Identifying and Correcting Labeling Bias in Machine Learning, they suggest a mathematical, multifaceted approach to mitigate bias through preprocessing techniques, such as data cleaning, augmentation, and weighting to address fairness. In this paper, the model focuses on race, gender, socioeconomic, and age demographics whereas the other paper—Debiasing Made State-of-the-art: Revisiting the Simple Seed-based Weak Supervision for Text Classification—focuses on text classification and preprocesses the data through random word deletion and augmentation to enhance the quality of pseudo-labels [2].

While neither paper is perfect, the approaches are competitive in comparison to other methods, as proven in the results. Nonetheless, they are reliant on the assumption that there is truly an underlying, unbiased label for every set of data. In the real world, this may not be the case. Furthermore, both papers bring up ethical questions of augmenting data, either through creating synthetic points or deleting seeds. Even so, they take significant steps towards addressing label bias ensuring fairness in artificial intelligence [8,9].

Acknowledgments

We thank Professor Pradeep Ravikumar and teaching assistant Yunshan Guo for guidance, teaching, and instruction. We also thank Neoscholar Education Group for giving us this opportunity. Without them, we would not have been able to complete this paper.

References

[1] Mekala, Chengyu Dong, and Jingbo Shang. 2022. Lops: Learning order inspired pseudo-label selection for weakly supervised text classification. ArXiv, abs/2205.12528.

[2] Mekala and Jingbo Shang. 2020. Contextualized weak supervision for text classification. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 323– 333.

[3] C., Wang, Z., & Shang, J. (2023, May 24). [2305.14794] Debiasing Made State-of-the-art: Revisiting the Simple Seed-based Weak Supervision for Text Classification. arXiv. Retrieved October 10, 2024

[4] J., Qu, L., Jia, R., & Zhao, B. (2019). O2U-Net: A Simple Noisy Label Detection Approach for Deep Neural Networks - Jinchi Huang Alibaba Group Hangzhou, China. CVF Open Access. Retrieved October 11, 2024

[5] H., & Nachum, O. (2019, January 15). [1901.04966] Identifying and Correcting Label Bias in Machine Learning. arXiv. Retrieved October 10, 2024

[6] Lichman et al. Uci machine learning repository, 2013.

[7] Compas recidivism risk score data and analysis, Mar 2018. URL https: //www. propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis.

[8] Park and Jihwa Lee. 2022. Lime: Weaklysupervised text classification without seeds. ArXiv, abs/2210.06720.

[9] Wang, Dheeraj Mekala, and Jingbo Shang. 2021. X-class: Text classification with extremely weak supervision. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3043–3053, Online. Association for Computational Linguistics.