An Empirical Investigation of Learning from Biased Toxicity Labels

Anonymous Authors[1]

Abstract

Collecting annotations from human raters often results in a trade-off between the quantity of labels one wishes to gather and the quality of these labels. As such, it is only possible to gather a small amount of high-quality labels. In this paper, we study how different training strategies can leverage a small dataset of human-annotated labels and a large but noisy dataset of synthetically generated labels (which exhibit bias against identity groups) for predicting toxicity of online comments. We evaluate the accuracy and fairness properties of these approaches, and whether there is a trade-off. While we find that pre-training on all of the data and fine-tuning on clean data produces the most accurate models, we could not determine a single strategy that was better across all fairness metrics considered.

1. Introduction

Supervised learning requires large amounts of labeled data, often human-annotated. This creates a trade-off. Human raters are imperfect and introduce bias and variance into their labels (Geva et al., 2019). When given enough time and resources, the quality of such labels can improve dramatically (Stiennon et al., 2020). Hence, given a fixed budget, there is a trade-off between label quality and quantity. One possible solution to this trade-off is to create a large amount of cheap, low-quality labels and a small amount of expensive, high-quality labels. This enables novel training approaches that use high-quality labels to minimise biases learnt from low-quality labels (Xiao et al., 2015; Ren et al., 2018; Zhang et al., 2020; Song et al., 2020).

In this work, we explore different ways to train a fair textual toxicity (Wulczyn et al., 2017; Dixon et al., 2018; Borkan et al., 2019) classifier in this regime. We have access to a small amount of high-quality labels and a large amount of low-quality labels. Our low-quality labels exhibit fairness-relevant biases, in particular, systematic differences in accuracy and predicted toxicity rate for different identity groups. Natural language is a particularly compelling context to study, as the field has seen recent rapid progress (Devlin et al., 2018; Brown et al., 2020) and models are becoming increasingly widely deployed, yet often exhibit bias (Dixon et al., 2018; Kurita et al., 2019; Sap et al., 2019).

We formalise this as a noisy labels problem, where we have a dataset of noisy labels (low-quality) and of clean labels (high-quality). To study this problem, we build a setup with the following key properties:

Labeler type The training data is annotated with labeler type - we know whether each data point is clean or noisy, and we have data of each type. This is in contrast to work that assumes we can only train on the noisy data (Jiang & Nachum, 2019).

Imbalance We have significantly more noisy data than clean data

Complex bias The biases are difficult to model precisely and often qualitative, as they emerge from human judgement. This is in contrast to prior work that models noisy labels as flipping labels between classes according to a transition matrix, independent of the input (Hendrycks et al., 2018; Lamy et al., 2019)

We focus on the Civil Comments dataset (Borkan et al., 2019), a collection of online comments annotated as toxic or non-toxic. This is suitable for a study of fairness as comments are annotated by identity references, enabling measurement of unintended bias against protected groups.

Similarly to Gu et al. (2021), we synthetically generate noisy labels, as there is no preexisting clean/noisy label split in the Civil Comments dataset. We treat the human labels as clean and train models on the human labels to generate synthetic labels. We explore several standard approaches to training models from imperfect data. We evaluate their accuracy and bias, and whether there is a trade-off between the two.

We find that pre-training on all of the data and then fine-tuning on the clean data is the best way to train an accurate model. Measuring fairness is more complex, and the right approach depends on the specific context where a model will be applied (Barocas et al., 2017). Accordingly, we

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

use the fairness-relevant metrics introduced by Borkan et al. (2019) to evaluate a range of possible biases. We focus on metrics that measure systemic differences in accuracy and systematic differences in predicted toxicity rate for different demographic groups. We find that no single model performs best on all metrics.

2. Methods

In this section we detail our experimental setup and baselines. In Section 3 we discuss the measurement of the accuracy and fairness properties of our baseline and robustness checkss. In Section 4 we summarise our findings.

2.1. Data

For our investigation, we create noisy (biased) and clean (less biased) datasets based on the Civil Comments dataset, a collection of almost two million online comments labeled as toxic and non-toxic. We follow the approach set out in Gu et al. (2021) to synthetically generate noisy labels for a dataset without a well-defined clean/noisy label split. Our models are based on a pre-trained BERT (Devlin et al., 2018) encoder, followed by a 2 layer MLP. We train them on both the clean and noisy datasets, validate on clean data only.

The original human labels are our clean labels and to create the noisy labels, we train networks to imitate these human-annotated labels. We then use these synthetic raters to generate synthetic labels for each comment, our noisy labels. To ensure a suitable level of noise, we stop training the networks before convergence, attaining a validation set accuracy of 95%. To avoid the synthetic labels being memorised from the training data, we hold-out half of the dataset when training the synthetic raters and only generate synthetic labels for the held-out portion.

We create our clean and noisy datasets for training our baselines from the held-out portion. We re-label 95% with noisy labels (i.e. discard the original clean label for that subset of data points), and the other 5% retains the original human label. This ensures an imbalance between clean and noisy dataset size, as desired.

Prior work has shown that networks trained on this dataset develop biases for or against identity groups, where different groups have systematic differences in accuracy and predicted toxicity rates (Dixon et al., 2018; Borkan et al., 2019). This is in part because the dataset contains correlations where comments mentioning certain identity groups are more or less likely to be toxic, and models tend to exaggerate this bias (Borkan et al., 2019). Thus our noisy labels exhibit bias relative to the human labels, as required for our analysis. While the original human labels may also exhibit bias, we refer to them as clean to indicate that they are *less* biased, not that they are unbiased.

Naturally, this approach has the key limitation that our noisy labels are synthetically generated, rather than being generated by true human labelers. We are limited by the lack of publicly available datasets with well-defined clean/noisy splits, and which allow us to measure fairness properties. As argued in Gu et al. (2021), we consider our approach a useful simulation of human bias. Neural network errors are complex and difficult to model, and share similarities with human error that simpler synthetic methods miss, such as having a higher error rate on harder examples.

See Appendix A for a more detailed discussion of how we generate this data, the properties of our noisy dataset and the limitations of this approach.

2.2. Baselines

We train several baselines on this synthetic dataset. All models are based on a pre-trained BERT (Devlin et al., 2018) encoder, followed by a 2-layer MLP. When training, we update the weights of both BERT and the MLP. All data points are of the form (k, x, y), where x is the comment text, the labeler type $k \in \{C, N\}$ represents whether the label is clean or noisy, and y is the comment label. We train on 878,620 noisy and 46,232 clean data points.

We evaluate the following strategies:

Clean The model just trains on the clean data (5% of the total training data)
Naive The model trains on both clean and noisy data, and ignores the labeler type
Multi-head The model has two heads, and uses one for clean data points, one for noisy data points. Parameters in all prior layers are shared
One-hot The labeler type is one-hot encoded and appended to the BERT output before entering the MLP. A variant of multi-head.
Loss correction (Patrini et al., 2017) The noisy data is modelled as a corrupted version of the clean data, where for each pair of classes there is a certain fixed probability that each element of the first is corrupted to the second. The parameters of the corruption matrix are estimated from the available clean data, and applied to the model outputs when predicting noisy labels. No corruption is applied when predicting clean labels.

We further fine-tune each baselines (except Clean) on clean data. We denote this by appending the suffix **FT** to name of the baseline.

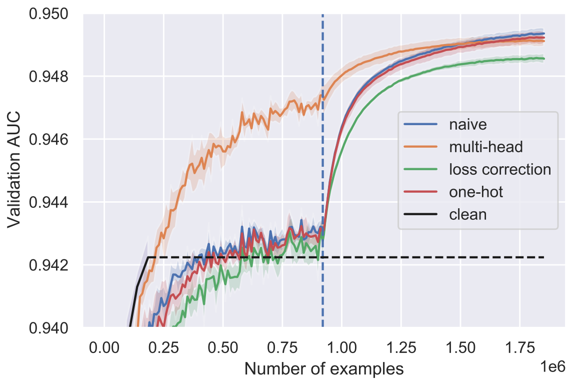

Figure 1. AUC for each baseline. The vertical line is the start of fine-tuning. Before fine-tuning, multi-head performs best. After fine-tuning all baselines improve and AUC difference become smaller.

3. Experiments

3.1. Accuracy

We first measure the performance of each baseline, as measured by Area Under the ROC Curve (AUC) with respect to the clean labels. The results for each baseline can be seen in Figure 1. This is calculated as AUC on the validation set, which has only clean labels. We observe that fine-tuning performs best (with a final AUC of 94.9%), then multi-head (with 94.7%), and then clean, naive and one-hot all perform similarly (between 94.2% and 94.3%). Notably, after fine-tuning all methods obtain similar performance (between 94.86% and 94.94%) despite there being significant variation in performance before fine-tuning. While we primarily use AUC to measure accuracy due to significant class imbalance, we note that our reported ordering between baselines is robust to alternate metrics such as binary accuracy and cross-entropy loss.

3.2. Fairness

3.2.1. Metrics

To measure fairness we use the fairness-relevant metrics introduced by Borkan et al. (2019), a common method for measuring bias in textual toxicity classification tasks (Conversation AI, 2019; Nozza et al., 2019; Zorian & Bikkanur, 2019). The Civil Comments Identities dataset is a subset of Civil Comments with annotations for whether each comment is a member of 13 identity groups, covering a range of race, religion, sexuality and gender considerations, allowing us to evaluate these metrics for each identity group. In particular, we focus on three of the metrics:

Subgroup AUC Evaluate the AUC of the model on each subgroup.

Background Positive, Subgroup Negative AUC (BPSN AUC) Evaluate the AUC of the model on the non-toxic data points of the subgroup and toxic data points not of the subgroup.

Negative Average Equality Gap (Negative AEG) Randomly select a non-toxic data point from the subgroup and a non-toxic data point not of the subgroup. Evaluate the proportion of the time that the model's predicted toxicity is higher for the subgroup data point. We subtract 0.5, so that an unbiased model has 0 Negative AEG.

We focus on these metrics as they measure two biases exhibited in our noisy data: systematic differences in accuracy between different identity groups and systematic differences in predicted toxicity rate between different identity groups. Subgroup AUC measures differences in performance, Negative AEG measures differences in predicted toxicity rate, and BPSN AUC measures both (Borkan et al., 2019).

We distinguish between **accuracy-based** metrics which correlate with overall AUC, and **accuracy-agnostic** metrics which do not. Subgroup AUC and BPSN AUC are accuracy-based as they measure model AUC on subsets of the data. Negative AEG is accuracy-agnostic, as a uniformly random classifier has a perfect Negative AEG of 0.

3.2.2. Results

We measure the Subgroup AUC, BPSN AUC and Negative AEG for each baseline, for each of 13 identity groups. The results are displayed in Figure 2. We aggregate the metrics across the 13 identity groups by taking the arithmetic mean. Alternate approaches such as weighting by identity group size give similar results, and ordering is consistent across subgroups.

For the Subgroup AUC and BPSN AUC metrics, the fine-tuned baselines exhibit least bias, followed by multi-head. However, this is the same ordering as overall AUC, as shown in Figure 1. As these metrics are accuracy-based and correlate with overall AUC, it is difficult to determine whether this effect is due to lower bias, or a consequence of higher overall AUC.

For the Negative AEG metric, the clean baseline exhibits the least bias. We find that all algorithms leveraging the noisy data introduce bias. Negative AEG is an accuracy-agnostic metric, suggesting that the performance of fine-tuning on Subgroup AUC and BPSN AUC may be attributed to higher overall AUC rather than decreased bias.

However, the difference in bias, though statistically significant, is slight and all baselines exhibit notable bias. The probability of classifying a subgroup data point as more toxic increases from 17.7% for clean to 18.2% for one-hot.

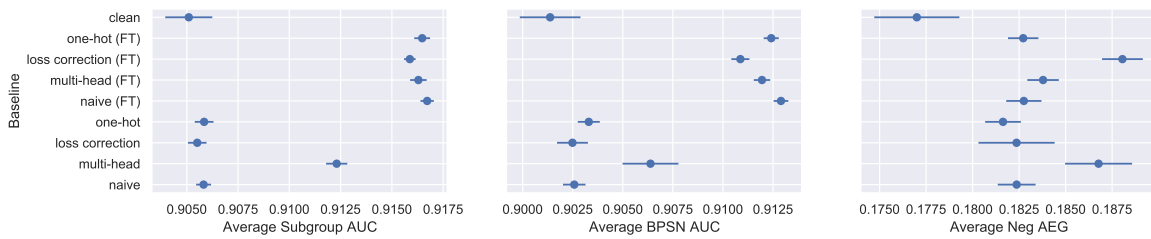

Figure 2. The Subgroup AUC, Background Positive Subgroup Negative AUC (BPSN AUC) and Negative Average Equality Gap (Neg AEG) for each baseline, averaged over the 13 identity groups. Each baseline was run 5 times with different seeds, and the mean and standard deviation of the aggregated metric are plotted. Low Subgroup AUC and BPSN AUC and high Neg AEG indicate bias.

3.3. Robustness Checks

To investigate the robustness of our results, we explore the sensitivity to the level of noise. We generate a new synthetic dataset of higher quality noisy labels, reproduce our baselines and measurements and compare the results to our results for lower quality noisy labels. We produce higher-quality noisy labels by training our synthetic raters on more data points than before: 880,000 data points with batch size 16, in comparison to 220,000 data points with batch size 4. The accuracy of the noisy labels relative to the clean labels increases from 95% to 95.5% and the AUC increases from 96% to 97%.

Overall AUC for each baseline on the higher quality noisy labels is shown in Figure 3. Fine-tuning has highest overall AUC, though is an improvement on naive of 0.1%, in comparison to an improvement of 0.6% before. Multi-head AUC is similar to naive and one-hot. The orderings for fairness metrics are similar to before: the accuracy-based metrics, Subgroup AUC and BPSN AUC, have the same ordering as overall AUC and clean remains least biased under Negative AEG. We discuss this experiment further in Appendix C.

It is clear that performance will depend on the degree of difference between clean and noisy data. However, this was an unexpected level of sensitivity and suggests that the results of this paper may be fairly context-specific.

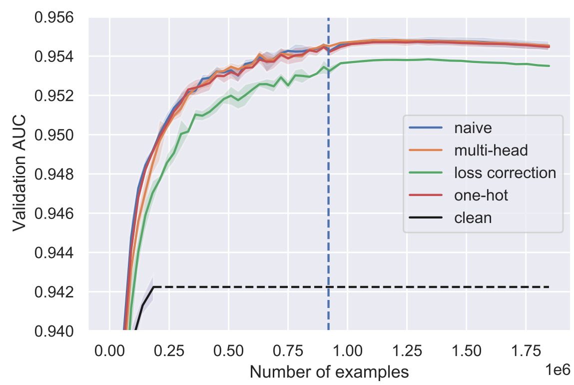

Figure 3. Baseline AUC on higher-quality noisy data. The vertical dashed line shows the switch to fine-tuning. The clean baseline uses early stopping. Each baseline was run 5 times with different seeds.

4. Conclusion

In this work, we conducted an empirical investigation into learning from biased labels for toxicity prediction, using synthetic labels from a neural network as a proxy for noisy human labels. With respect to AUC, fine-tuned models performed best on our original dataset. With respect to fairness metrics, no single model performed best for all metrics – while fine-tuning exhibited the least bias on the accuracy-based metrics of Subgroup AUC and BPSN AUC, the approach of ignoring noisy labels entirely exhibited the least bias on the accuracy-agnostic Negative AEG metric.

As training machine learning models on large amounts of loosely curated data becomes commonplace, it is essential that we understand the effects of imperfect labels on accuracy and fairness of the resulting models. We recommend caution in extrapolating to other contexts based on these results – we only study a single dataset with synthetically generated labels, and different comparisons may result from different noise characteristics. Nevertheless, we hope this work provides a useful set of empirical observations towards this important question.

References

Barocas, S., Hardt, M., and Narayanan, A. *Fairness in machine learning*, chapter Audit studies, pp. 124. 2017.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561, 2019. URL http://arxiv.org/abs/1903.04561.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Conversation AI. Jigsaw unintended bias in toxicity classification, 2019. URL https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification.

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

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73, 2018.

Geva, M., Goldberg, Y., and Berant, J. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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)*, pp. 1161–1166, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1107. URL https://www.aclweb.org/anthology/D19-1107.

Gu, K., Masotto, X., Bachani, V., Lakshminarayanan, B., Nikodem, J., and Yin, D. A realistic simulation framework for learning with label noise. *Under Review*, 2021.

Hendrycks, D., Mazeika, M., Wilson, D., and Gimpel, K. Using trusted data to train deep networks on labels corrupted by severe noise. *CoRR*, abs/1802.05300, 2018. URL http://arxiv.org/abs/1802.05300.

Jiang, H. and Nachum, O. Identifying and correcting label bias in machine learning. *CoRR*, abs/1901.04966, 2019. URL http://arxiv.org/abs/1901.04966.

Kurita, K., Vyas, N., Pareek, A., Black, A. W., and Tsvetkov, Y. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*, 2019.

Lamy, A., Zhong, Z., Menon, A. K., and Verma, N. Noise-tolerant fair classification. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/8d5e957f297893487bd98fa830fa6413-Paper.pdf.

Nozza, D., Volpetti, C., and Fersini, E. Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 149–155, 2019.

Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1944–1952, 2017.

Ren, M., Zeng, W., Yang, B., and Urtasun, R. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pp. 4334–4343. PMLR, 2018.

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.

Song, H., Kim, M., Park, D., Shin, Y., and Lee, J.-G. Learning from noisy labels with deep neural networks: A survey. *arXiv preprint arXiv:2007.08199*, 2020.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. Learning to summarize from human feedback. *arXiv preprint arXiv:2009.01325*, 2020.

Wulczyn, E., Thain, N., and Dixon, L. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pp. 1391–1399, 2017.

Xiao, T., Xia, T., Yang, Y., Huang, C., and Wang, X. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2691–2699, 2015.

Zhang, Z., Zhang, H., Arik, S. O., Lee, H., and Pfister, T. Distilling effective supervision from severe label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9294–9303, 2020.

Zorian, A. A. and Bikkanur, C. S. Debiasing personal identities in toxicity classification. *arXiv preprint arXiv:1908.05757*, 2019.