Is the Lottery Fair? Evaluating Winning Tickets Across Demographics

Victor Petrén Bach Hansen,[1 2] **Anders Søgaard**[1]
[1]Department of Computer Science, University of Copenhagen, Denmark
[2]Topdanmark A/S, Denmark
{victor.petren,soegaard}@di.ku.dk

Abstract

Recent studies have suggested that weight pruning, e.g. using lottery ticket extraction techniques (Frankle and Carbin, 2018), comes at the risk of compromising the group fairness of machine learning models (Paganini, 2020; Hooker et al., 2020), but to the best of our knowledge, no one has empirically evaluated this hypothesis at scale in the context of natural language processing. We present experiments with two text classification datasets annotated with demographic information: the Trustpilot Corpus (sentiment) and CivilComments (toxicity). We evaluate the fairness of lottery ticket extraction through layer-wise and global weight pruning across three languages and two tasks. Our results suggest that there is a small increase in group disparity, which is most pronounced at high pruning rates and correlates with instability. The fairness of models trained with distributionally robust optimization objectives is sometimes less sensitive to pruning, but results are not consistent. The code for our experiments is available at https://github.com/vpetren/fairness_lottery.

1 Introduction

Heavily pruning deep neural network models is a way of reducing inference cost for resource-constrained environments, but does weight-pruning of deep neural networks increase their unfairness? Several recent papers suggest this (Paganini, 2020; Hooker et al., 2020), based on experiments from face and digit recognition, but does this also hold for natural language processing (NLP) models? Systematic biases may easily be exacerbated by pruning interventions in high-dimensional problems because of feature swamping effects (Sutton et al., 2006). Overparameterized deep neural networks generalize well, in part because they can hedge their bets and rely on multitudes of weak evidence rather than the most prominent independent variables. Sparse models do not have that luxury and are therefore more sensitive to shifts (Globerson and Roweis, 2006; Søgaard, 2013).

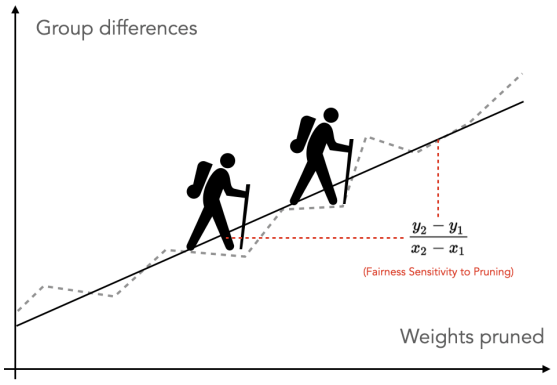

Figure 1: Fairness Sensitivity to Pruning (FSP): the gradient of the linear fit of (the logarithm of) the pruning ratio to min-max group-level disparity. We use this to quantify the sensitivity of Rawlsian min-max fairness to weight pruning across architectures, pruning strategies and datasets.

We introduce a *fairness sensitivity to pruning* metric that measures how Rawlsian min-max fairness across demographic groups changes with weight pruning. We estimate this sensitivity by taking the gradient of the linear fit of the logarithm of the pruning ratio to min-max group-level disparity. We show that across four datasets, fairness sensitivity to pruning is similar for layer-wise and global pruning strategies (Frankle and Carbin, 2018), as well as for text classifiers based on feed-forward and recurrent neural networks. Subsequently, we consider the impact of a popular robust optimization strategy designed to improve the fairness of classification models (Hashimoto et al., 2018; Sagawa et al., 2020b), on the fairness sensitivity of feed-forward networks.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3214–3224
August 1–6, 2021.

Contributions We are, to the best of our knowledge, the first to study the impact of weight pruning on fairness in NLP at scale. We introduce a *fairness sensitivity to pruning* (FSP) metric that measures how Rawlsian min-max fairness across demographic groups decreases with weight pruning. We evaluate FSP across two architectures, two pruning strategies and two datasets, including multilingual sentiment classification and English toxicity classification. Our results suggest that pruning increases group-level performance disparities, but mostly at high pruning rates and with some variance across architectures and pruning strategies. Group-level disparities seem to be in part a result of the instability of weight pruning. We compare FSP between our baseline empirical risk models and robust models induced with Distributional Robust Optimization (DRO) (Hashimoto et al., 2018; Sagawa et al., 2020b). Our results show that weight pruning in combination with DRO can *sometimes* (8/16 cases here) be used to induce fairer, sparse classifiers, but the effect is not significant ($p \sim 0.18$) across our experiments.

2 Related Work

Pruning neural networks The literature on pruning neural networks is decades old (Mozer and Smolensky, 1989; Cun et al., 1990; Hassibi and Stork, 1993), but has recently seen a resurgence with the all-encompassing success of neural networks and the need for small and fast on-device model inference (Han et al., 2015; Sze et al., 2017; Frankle and Carbin, 2018; Frankle et al., 2019). In NLP, specifically, pruning methods have been applied to recurrent neural networks (Desai et al., 2019; Yu et al., 2020), as well as transformers (Gordon et al., 2020; Brix et al., 2020; Prasanna et al., 2020; Chen et al., 2020; Sanh et al., 2020).

Fairness in pruned models Measuring fairness in pruned models is an unexplored area. However, Paganini (2020) evaluates the fairness, i.e., the difference between the best- and worst-case groups, of lottery ticket-style weight pruning for digit recognition problems: Specifically, they retrain models for a fixed number of iterations using global unstructured pruning. In addition, they present a meta-regression study suggesting that underrepresented and more complex classes are most severely affected by pruning procedures. See Hooker et al. (2020) for related work and similar results in face recognition.[1]

Improving fairness Fairness of overparameterized models can be improved by distributionally robust optimization (DRO) (Hashimoto et al., 2018; Levy et al., 2020), or to some extent by simpler post-hoc correction methods such as classifier retraining or group-specific classification thresholds (Menon et al., 2021). DRO minimizes the worst-case expected loss over an uncertainty set of distributions. The uncertainty set represents the distributions we want our model to perform well on. In Sagawa et al. (2020a), the uncertainty set is all possible mixtures of a known set of groups, a variant referred to as Group DRO. Sagawa et al. (2020b) find that subsampling the majority groups can be a way for overparameterized models to achieve both low minority test error as well as low average test error.

3 Pruning methodology

We extract winning lottery tickets from our network according to the iterative procedure outlined in Frankle and Carbin (2018): Given a model $f(x;\theta)$ with initial network parameters θ_0 and mask m_0, for each pruning iteration i, we start by initializing a model $f(x;\theta)$ with initial parameter θ_0 and train it for N epochs, resulting in $f(x;\theta_N)$. After training, we prune a fixed fraction $p \in [0,1]$ from the remaining parameters in θ_N to obtain the mask m_i. The pruned weights are chosen using the L_1 norm, meaning the neurons with the lowest magnitude are masked out. Pruning can either be done w.r.t. individual layers or all of them combined, also referred to as *layer-wise* and *global* pruning. m_i is then carried over to the subsequent pruning iteration $i+1$ with the model $f(x, m_i \odot \theta_0)$ and retrained once again. At iteration i, the fraction of weights pruned is therefore $1-(1-p)^i$.

4 Experiments

4.1 Data

Datasets We examine fairness among heavily pruned models using two text classification datasets: i) The multilingual **Trustpilot** Corpus

[1] Bartoldson et al. (2020) arguably present results from object recognition that show the opposite trend: Generalization increases with (layer-wise) pruning. This seems to be a side effect of overparameterization; interestingly, we see the opposite trend for feed-forward networks and layer-wise pruning.

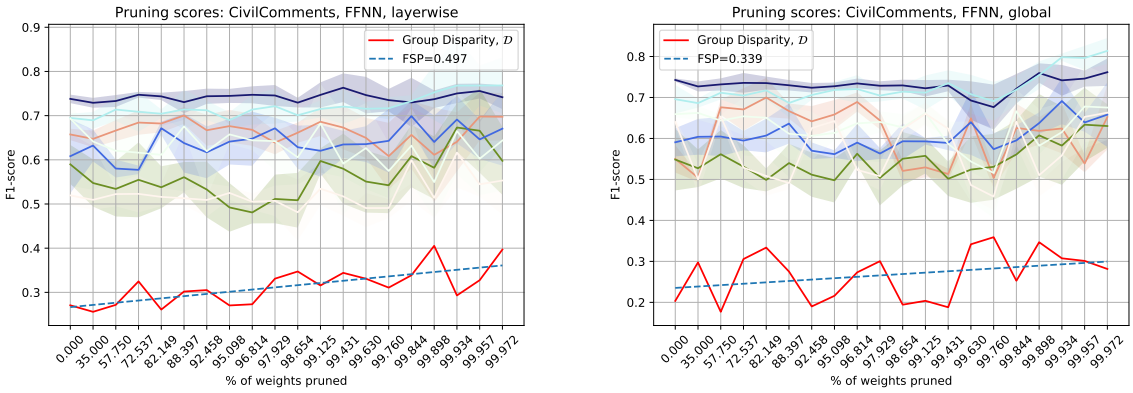

Figure 2: Macro-averaged performance of our feed-forward networks as a function of pruning ratio. Fairness Sensitivity to Pruning (FSP) correspond to the gradient of the linear fit to the min-max differences across individual runs. Results are for CIVILCOMMENTS. The hard line represents the average demographic score over 5 individual runs and the shaded area represents the standard deviation. See the Appendix for similar plots for the Trustpilot Corpus.

(Hovy et al., 2015),[2] which contains user reviews from the Trustpilot website of various companies and services in five different countries (Germany, Denmark, France, United Kingdom and United States). The reviews are based on a one to five star rating scale and some are accompanied by demographic attributes about the author, such as gender, age and location. 2) The **CivilComments** dataset (Borkan et al., 2019),[3] which contains comments annotated for toxicity, for the purpose of hate speech detection. A subset of the comments are also annotated for the protected attributes they address, including gender, race, and religion.

Preprocessing For the Trustpilot Corpus, we divide the data into demographics based on a combination of *gender* (male/female), *age* (young/old) and *location* (NUTS regions). For age, young is defined as being 35 or less. We exclude the French and American parts of the datasets as they do not have properly annotated NUTS regions. For UK and Germany, we use NUTS-1 regions, and for Denmark, where more data is available, we use NUTS-2 regions. We convert the 5-star ratings to binary sentiment labels, grouping 4 and 5 stars as positive, and 1 and 2 as negative. Neutral reviews (three stars) are discarded.[4] Likewise for CivilComments, we threshold comments with a toxicity rating > 0.5 as toxic, and otherwise label them as a non-toxic. This is similar to the binarization performed in Koh et al. (2020). Comments can for each demographic sub-attribute contain multiple partial values (e.g. $asian = 0.3$, $black = 0.4$ for the *race* attribute), so for each annotated attribute we assign it the sub-attribute with the largest value. In our experiments we consider demographics based on combinations of the *race* and *gender* attributes. For each language and dataset we randomly sample 100, 200 or 500 of each demographic as test sets, based on the the amount of annotated datapoints in the dataset, and use a 80-20 split of the remaining data for training and validation. If a demographic contains less than the specified number of datapoints, we disregard it. Due to high class imbalance, the majority class for our train-val data is downsampled to match the minority class. Table 1 shows the statistics for the respective datasets we train and evaluate on.

Dataset	Train	Val	N	S
Trustpilot-DK	222229	55557	20	500
Trustpilot-DE	26146	6536	42	100
Trustpilot-UK	127965	31991	50	200
CivilComments	357602	89400	7	100

Table 1: Detailed dataset statistics. N refers to the number of discrete demographics in the dataset and S is the size of each demographic test set.

4.2 Models

We consider simple **FFNN** (Rumelhart et al., 1986) and **LSTM** (Hochreiter and Schmidhuber, 1997)

[2] https://bitbucket.org/lowlands/release/src/master/WWW2015/data/

[3] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data

[4] This binarization scheme is standard; see, e.g., Gupta et al. (2020) and Desai et al. (2019)

FFNN				
Dataset	E_{dim}	h_{dim}	B	N
Trustpilot-DK	128	256	15	32
Trustpilot-DE	128	256	15	8
Trustpilot-UK	128	256	15	16
CivilComments	128	256	15	32
LSTM				
Dataset	E_{dim}	h_{dim}	B	N
Trustpilot-DK	128	256	10	64
Trustpilot-DE	128	256	15	16
Trustpilot-UK	128	256	10	32
CivilComments	128	256	10	64

Table 2: FFNN and LSTM hyperparameters. E_{dim} is embedding layer size, h_{dim} is hidden layer size, B is batch size and N is number of epochs. Both the layer-wise and global pruning structures use the same set of hyperparameters.

neural networks for text classification.

FFNN The FFNN consists of the following: The embedding layer, which maps every token id in the text to a fixed size vector as a bag-of-embeddings and sums them together, resulting in a single representation $e \in \mathbb{R}^{|E_dim|}$, followed by 3 fully connected layers of size $\mathbb{R}^{|E_dim \times h|}$, $\mathbb{R}^{|h \times h|}$ and $\mathbb{R}^{|h \times 2|}$ respectively. We use the hyperbolic tangent activation between layers and each linear layer is initialized using He initialization (He et al., 2015).

LSTM The LSTM network is a 2-layer bidirectional LSTM (Hochreiter and Schmidhuber, 1997) which encodes our input text, followed by a fully connected layer for classification. The weights are initialized using $\mathcal{U}(-\sqrt{k}, \sqrt{k})$ where $k = \frac{1}{hidden\_size}$ and the final fully connected layer uses He initialization. See all model hyperparameters used in Table 2.

Both the FFNN and LSTM models are trained using the Adam optimizer (Kingma and Ba, 2017) with a learning rate of $1e-3$ and a weight decay of $1e-4$.

Distributionally Robust Optimization Loss Additionally, we also train our models with DRO loss (Levy et al., 2020). We use the implementation provided by Levy et al. (2020) [5]. For our experiments, a χ^2 uncertainty set of size 1 is used.

For all of our experiments, we extract our winning tickets over 20 pruning iterations and use a pruning rate of $p = 0.35$. We run a total of 5 independent runs for each model-dataset combination.

[5] https://github.com/daniellevy/fast-dro/

		Trustpilot			CC	Avg
		da	de	en	en	
FFNN	lw	−0.183	0.281	−0.230	0.497	0.091
	gl	0.227	1.375	1.054	0.339	0.749
FFNN-DRO	lw	−0.044	0.321	0.143	0.089	0.127
	gl	0.351	0.875	−0.040	0.368	0.388
LSTM	lw	0.221	0.411	0.206	0.823	0.415
	gl	1.099	0.198	0.352	0.252	0.475
LSTM-DRO	lw	0.263	−0.282	−0.082	1.335	0.309
	gl	0.262	−0.609	0.544	0.006	0.051

Table 3: FSP values across architectures, layer-wise (lw) and global (gl) pruning, and the four datasets. Our main observation is that FSP values are almost consistently positive, and slightly higher for global pruning. DRO does not consistently reduce FSP; we highlight cases where it does.

4.3 Measuring group disparity

At each pruning step we measure the group disparity $\mathcal{D}$, from a set of demographics D, between repeated runs R, by computing the maximum difference of F_1 scores as follows:[6]

$$\mathcal{D} = \max_{d_m \in D} \max_{d_{n \neq m} \in D} \max_{r_i \in R} \max_{r_{j \neq i} \in R} |F_{1 r_i d_m} - F_{1 r_j d_n}| \quad (1)$$

Intuitively, this corresponds to the difference between the highest scoring run for the highest scoring demographic and the lowest counterpart. We compute FSP by taking the gradient of the linear fit of $\mathcal{D}$ over a P pruning steps multiplied by 100.

5 Results

Main experiments Our first set of results evaluate FSP across architectures, datasets, and pruning techniques. In 14/16 combinations of FFNN and LSTM neural networks, the Trustpilot Corpus and CivilComments, layer-wise and global pruning, we see positive FSP values. In other words, weight pruning leads to higher group-level performance disparities, i.e., less fairness. Comparing layer-wise and global pruning, we note that group disparity is generally higher for global pruning. In Figure 2, we present two plots - for layer-wise and global pruning of a feed-forward network trained on CivilComments. The remaining plots are presented in the Appendix. The FSP values are listed in Table 3. FFNNs exhibit very high FSP values

[6] Maximum discrepancy has also been used as a measure of fairness in Calmon et al. (2017); Alabi et al. (2018). See Williamson and Menon (2019) for discussion.

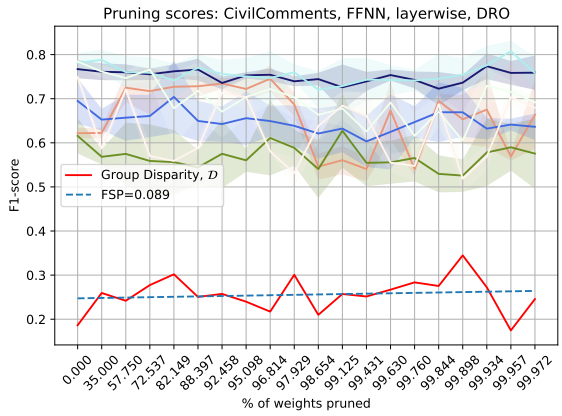

Figure 3: FSP for Distributional Robust Optimization

with global pruning, but while global pruning increases unfairness, layer-wise pruning does not. For LSTMs, the effects of the two pruning strategies are similar: Both lead to moderate increases in group disparities.[7] In a couple of instances we witnessed model degeneration due to heavy pruning resulting in single-class prediction before 20 pruning iterations. The plots and FSP values exclude these datapoints as they are not relevant for our analysis.

Distributionally Robust Optimization We ran comparable experiments using DRO loss (Hashimoto et al., 2018) to see whether the adverse effects of weight pruning on min-max fairness could be reduced by training with a more robust objective. This seems to hold true in some instances. We present a single plot for DRO in Figure 3, for feed-forward networks, layer-wise pruning on CivilComments; see the Appendix for more plots. Comparing with Figure 2 (left) the FSP metric is considerably lower than for baseline empirical risk minimization (0.089 vs. 0.497) while maintaining equal, or even better, performance at high pruning rates; but note from the red numbers in Table 3, that we only see this type of reduction in FSP in 3/8 cases for FFNNs, but DRO does reduce the average FSP for global pruning. In 5/8 cases for the LSTM, however, DRO does improves fairness, reducing the average FSP with both layer-wise and global pruning.

[7] While fairness correlates with stability, the difference between FFNNs and LSTMs is not explained by stability differences (see plots in the Appendix), but should probably be attributed to the general performance differences between FFNNs and LSTMs, as well as relative overparameterization in FFNNs (see Footnote 1).

6 Conclusion

In this work, we take a first step in examining group disparity among heavily pruned models, using lottery ticket extraction, in NLP. We measure group disparity, using *fairness sensitivity to pruning*, on the Trustpilot Corpus, a sentiment classification dataset covering 3 languages, as well as CivilComments, a toxicity classification dataset, for both feed-forward and recurrent neural networks. We find that models subject to heavy pruning are more susceptible to higher levels of group disparity, but that this effect can to some degree be mitigated using distributionally robust optimization objectives.

Acknowledgements

This work was funded by the Innovation Fund Denmark and Topdanmark.

References

Daniel Alabi, Nicole Immorlica, and Adam Kalai. 2018. Unleashing linear optimizers for group-fair learning and optimization. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 2043–2066. PMLR.

Brian Bartoldson, Ari S. Morcos, Adrian Barbu, and Gordon Erlebacher. 2020. The generalization-stability tradeoff in neural network pruning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561.

Christopher Brix, Parnia Bahar, and Hermann Ney. 2020. Successfully applying the stabilized lottery ticket hypothesis to the transformer architecture.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. 2020. The lottery ticket hypothesis for pre-trained bert networks.

Yann Le Cun, John S. Denker, and Sara A. Solla. 1990. *Optimal Brain Damage*, page 598–605. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Shrey Desai, Hongyuan Zhan, and Ahmed Aly. 2019. Evaluating lottery tickets under distributional shifts. *CoRR*, abs/1910.12708.

Jonathan Frankle and Michael Carbin. 2018. The lottery ticket hypothesis: Training pruned neural networks. *CoRR*, abs/1803.03635.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. 2019. Linear mode connectivity and the lottery ticket hypothesis. *CoRR*, abs/1912.05671.

Amir Globerson and Sam Roweis. 2006. Nightmare at test time: Robust learning by feature deletion. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 353–360, New York, NY, USA. Association for Computing Machinery.

Mitchell A Gordon, Kevin Duh, and Nicholas Andrews. 2020. Compressing {bert}: Studying the effects of weight pruning on transfer learning.

Aakriti Gupta, Kapil Thadani, and Neil O'Hare. 2020. Effective few-shot classification with transfer learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1061–1066, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Song Han, Jeff Pool, John Tran, and William J. Dally. 2015. Learning both weights and connections for efficient neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1135–1143, Cambridge, MA, USA. MIT Press.

Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization.

Babak Hassibi and David Stork. 1993. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing Systems*, volume 5, pages 164–171. Morgan-Kaufmann.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. 2020. What do compressed deep neural networks forget?

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. *User Review Sites as a Resource for Large-Scale Sociolinguistic Studies*, page 452–461. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2020. Wilds: A benchmark of in-the-wild distribution shifts.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. 2020. Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems*.

Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. 2021. Overparameterisation and worst-case generalisation: friend or foe? In *ICLR*.

Michael C. Mozer and Paul Smolensky. 1989. *Skeletonization: A Technique for Trimming the Fat from a Network via Relevance Assessment*, page 107–115. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Michela Paganini. 2020. Prune responsibly.

Sai Prasanna, Anna Rogers, and Anna Rumshisky. 2020. When bert plays the lottery, all tickets are winning.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning internal representations by error propagation. In David E. Rumelhart and James L. Mcclelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, pages 318–362. MIT Press, Cambridge, MA.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020a. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020b. An investigation of why overparameterization exacerbates spurious correlations.

Victor Sanh, Thomas Wolf, and Alexander M. Rush. 2020. Movement pruning: Adaptive sparsity by fine-tuning.

Anders Søgaard. 2013. Part-of-speech tagging with antagonistic adversaries. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 640–644, Sofia, Bulgaria. Association for Computational Linguistics.

Charles Sutton, Michael Sindelar, and Andrew McCallum. 2006. Reducing weight undertraining in structured discriminative learning. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 89–95, New York City, USA. Association for Computational Linguistics.

V. Sze, Y. Chen, T. Yang, and J. S. Emer. 2017. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 105(12):2295–2329.

Robert Williamson and Aditya Menon. 2019. Fairness risk measures. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6786–6797. PMLR.

Haonan Yu, Sergey Edunov, Yuandong Tian, and Ari S. Morcos. 2020. Playing the lottery with rewards and multiple languages: lottery tickets in rl and nlp. In *International Conference on Learning Representations*.

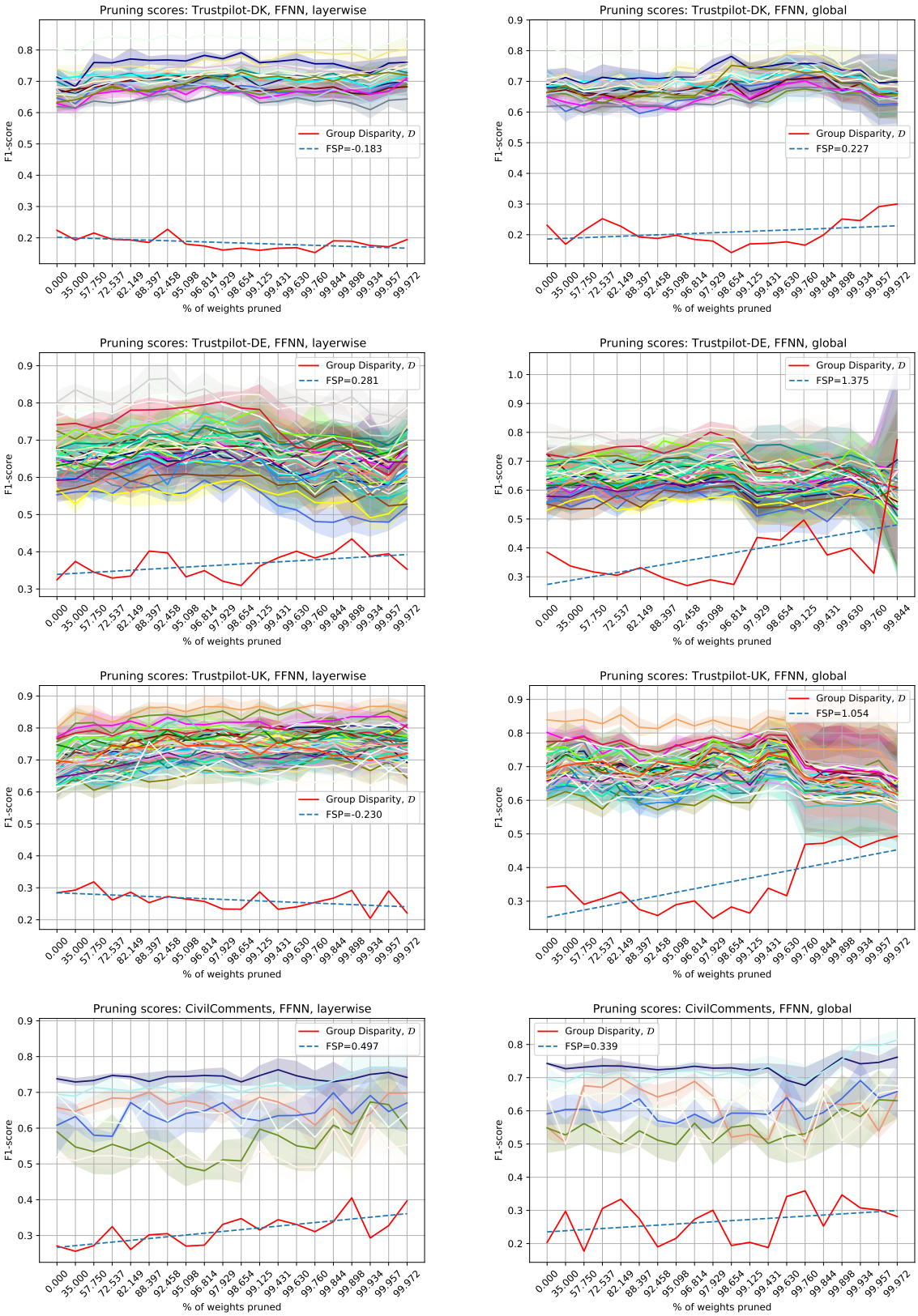

Figure 4: Macro-averaged performance of our feed-forward networks as a function of pruning ratio. The hard line represents the average demographic score over 5 individual runs and the shaded area represents the standard deviation. Fairness Sensitivity as Pruning (FSP) correspond to the gradient of the linear fit to the min-max differences across individual runs.

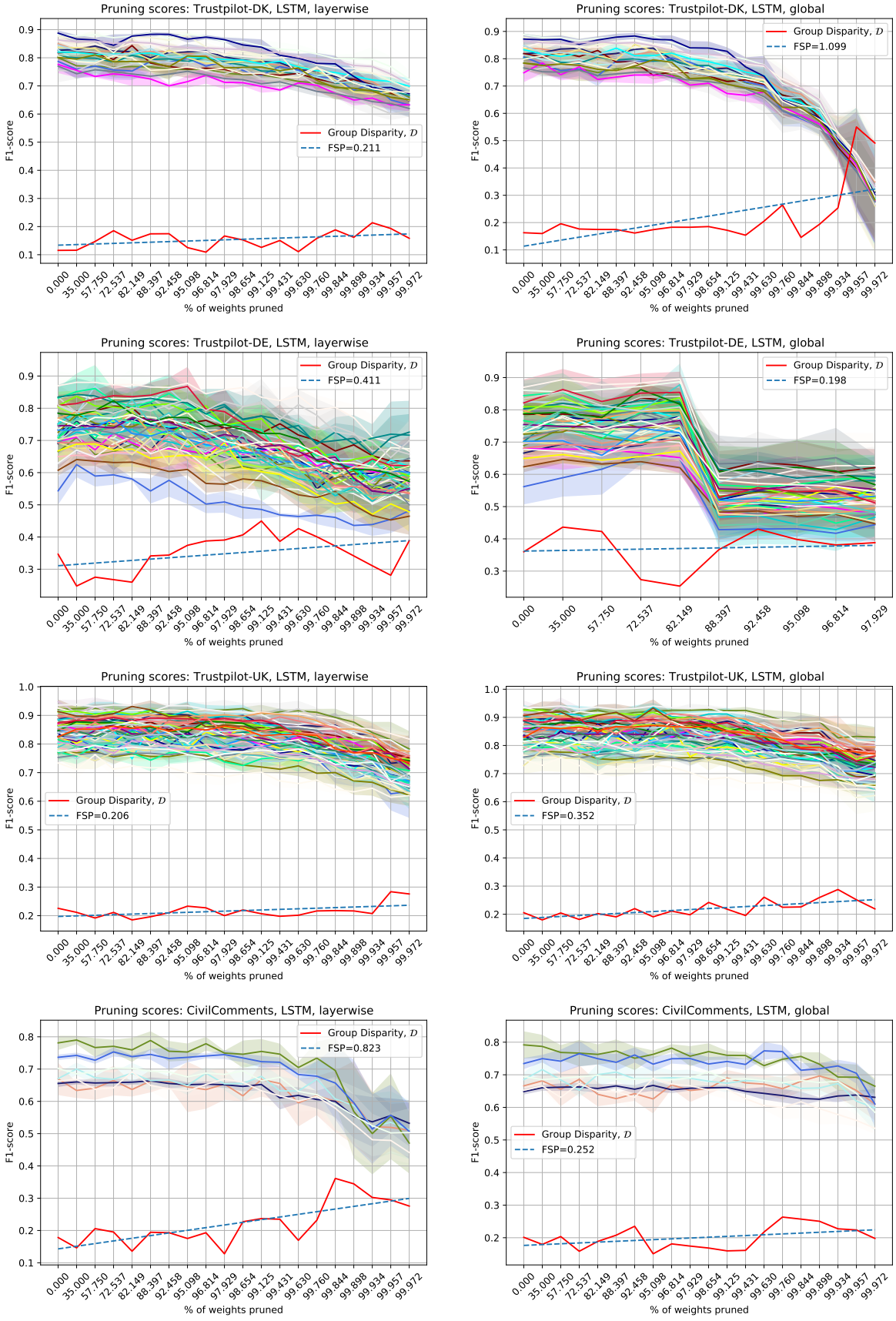

Figure 5: Macro-averaged performance of our LSTMs as a function of pruning ratio. The hard line represents the average demographic score over 5 individual runs and the shaded area represents the standard deviation. Fairness Sensitivity as Pruning (FSP) correspond to the gradient of the linear fit to the min-max differences across individual runs.

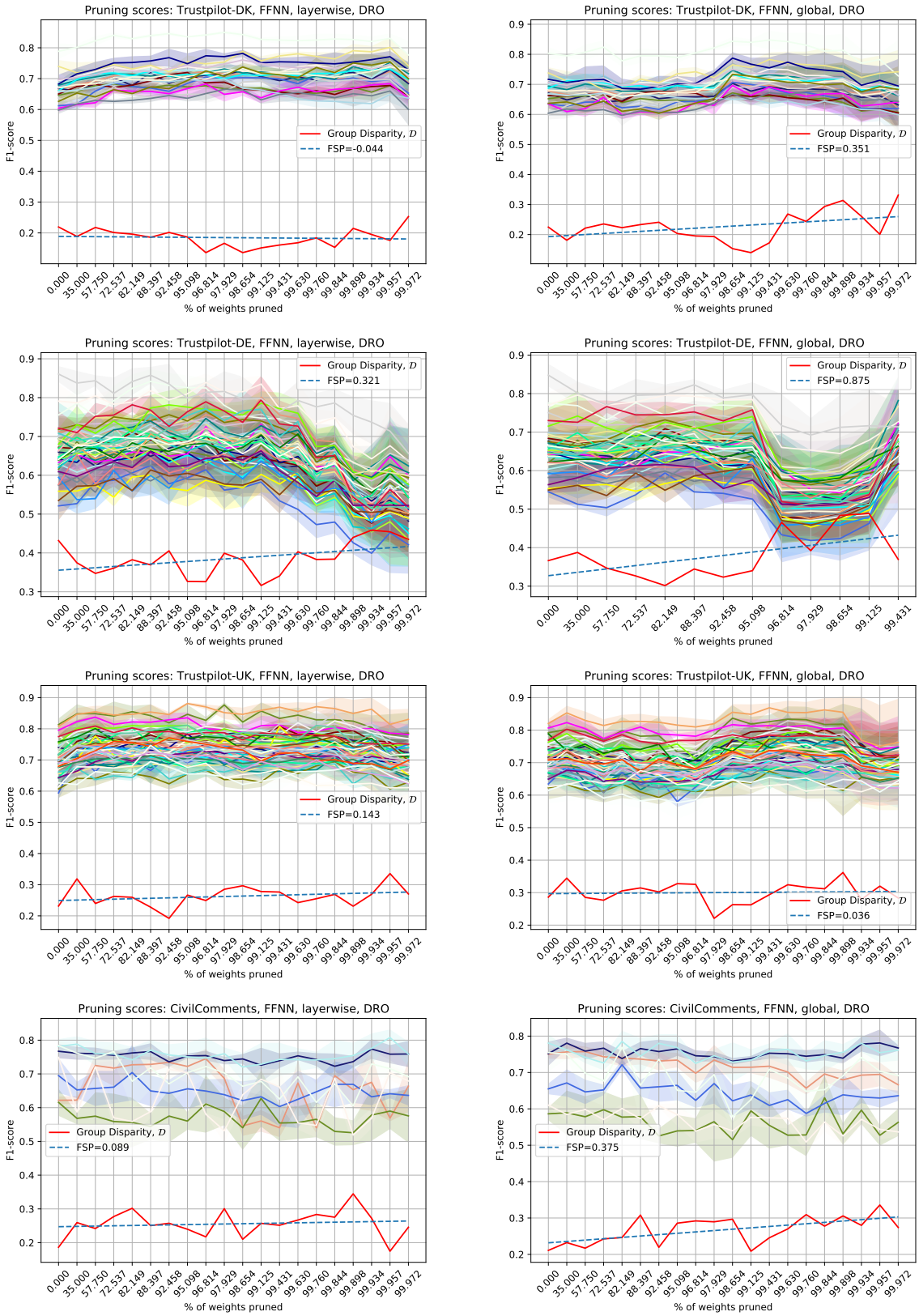

Figure 6: Macro-averaged performance of our layer-wise and globally pruned feed-forward networks trained with DRO as a function of pruning ratio. The hard line represents the average demographic score over 5 individual runs and the shaded area represents the standard deviation. Fairness Sensitivity as Pruning (FSP) correspond to the gradient of the linear fit to the min-max differences across individual runs.

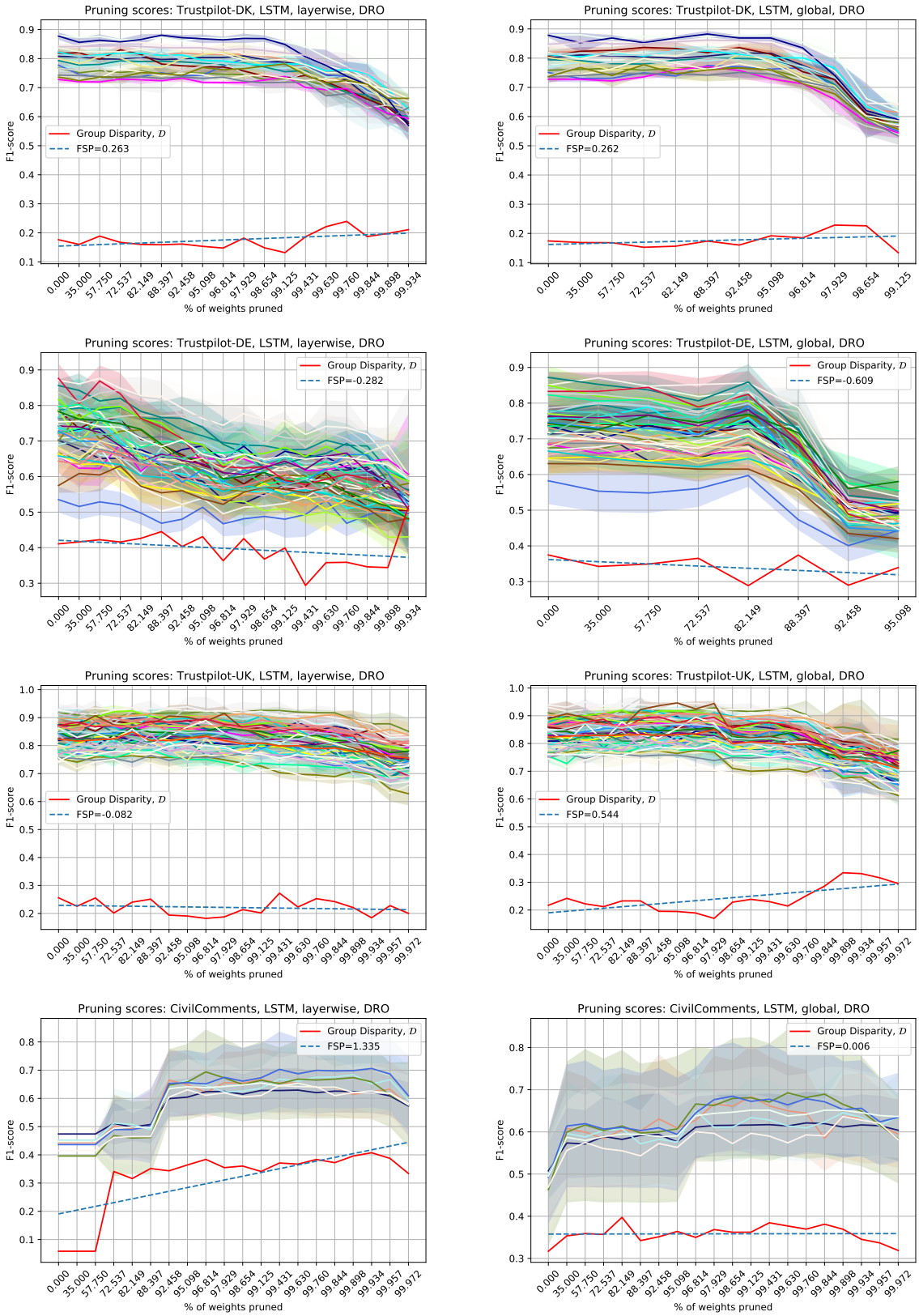

Figure 7: Macro-averaged performance of our layer-wise and globally pruned LSTM networks trained with DRO as a function of pruning ratio. The hard line represents the average demographic score over 5 individual runs and the shaded area represents the standard deviation. Fairness Sensitivity as Pruning (FSP) correspond to the gradient of the linear fit to the min-max differences across individual runs.