TASK-ADAPTIVE PRE-TRAINING FOR BOOSTING LEARNING WITH NOISY LABELS: A STUDY ON TEXT CLASSIFICATION FOR AFRICAN LANGUAGES

Dawei Zhu, Michael A. Hedderich, Fangzhou Zhai, David Ifeoluwa Adelani, Dietrich Klakow
Saarland University, Saarland Informatics Campus, Germany
{dzhu,mhedderich,didelani,dietrich.klakow}@lsv.uni-saarland.de
fzhai@coli.uni-saarland.de

For high-resource languages like English, text classification is a well-studied task. The performance of modern NLP models easily achieves an accuracy of more than 90% in many standard datasets for text classification in English (Xie et al., 2019; Yang et al., 2019; Zaheer et al., 2020). However, text classification in low-resource languages is still challenging due to the lack of annotated data. Although methods like weak supervision and crowdsourcing can help ease the annotation bottleneck, the annotations obtained by these methods contain label noise. Models trained with label noise may not generalize well. To this end, a variety of noise-handling techniques have been proposed to alleviate the negative impact caused by the errors in the annotations (for extensive surveys see (Hedderich et al., 2021; Algan & Ulusoy, 2021)). In this work, we experiment with a group of standard noisy-handling methods on text classification tasks with noisy labels. We study both simulated noise and realistic noise induced by weak supervision. Moreover, we find task-adaptive pre-training techniques (Gururangan et al., 2020) are beneficial for learning with noisy labels.

Problem Settings We consider a k-class text classification problem with noisy labels. In particular, instead of accessing a training set $D = \{(x_i, y_i)_{i=1}^n\}$ sampled from the clean data generation distribution, we work with a noisy version of D, denoted by $\hat{D} = \{(x_i, \hat{y}_i)_{i=1}^n\}$. The goal is to train a model on $\hat{D}$ that generalizes well on the clean distribution.

Noise Simulation Research in this field often constructs a noisy dataset by injecting noise into a clean dataset (Han et al., 2018; Jindal et al., 2019; Tänzer et al., 2021). This simulates annotation scenarios such as crowdsourcing, where some annotators answer randomly or overlook some entries in multiple-choice questions. It also allows controlling the noise level and type. The noise in such a simulation is feature-independent. That is, it assumes $p(\hat{y}|y = i, x) = p(\hat{y}|y = i)$.

Weak Supervision Noise In weak supervision, the labels are annotated in an automatic manner. For example, Hedderich et al. (2020) assign labels to a text document by leveraging word lists and a set of simple rules. Unlike the simulated noise, the noise caused by weak supervision is feature-dependent, meaning that the independence assumption does not hold.

Task-Adaptive Pre-Training (TAPT) The BERT model is pre-trained on a large unlabeled, general-domain corpus. Task-adaptive pre-training further pre-trains BERT on text from the downstream task. In particular, given a task-domain corpus, one continues to train BERT with the Masked Language Model (MLM) task and the Next Sentence Prediction (NSP) task. Previous studies have shown that task-adaptive pre-training is beneficial for downstream NLP tasks with clean training data (Sun et al., 2019; Gururangan et al., 2020). However, it remains unclear whether the noisy settings can inherit this benefit.

Data We simulate the noise on two established text classification dataset: AG-News (Zhang et al., 2015) and IMDB (Maas et al., 2011). We study two simulated noise types, uniform (van Rooyen et al., 2015) and single-flip (Reed et al., 2015) noise. We also evaluate two datasets with weak supervision noise in two low-resource African languages: Hausa and Yorùbá (Hedderich et al., 2020). The weak labels for these two datasets are generated by simple rules. For example, to identify texts for the class "Africa", a labeling rule based on a list of African countries and their capitals is used.

Baselines We compare learning without noise-handling with four popular noise-handling methods. 1) **Co-Teaching**: Han et al. (2018) trains two networks to pick cleaner training subsets for each

Methods	Original	+TAPT
Without Noise-handling	85.49	2.13↑
Co-Teaching	84.74	0.90↑
Noise Matrix	83.90	2.80↑
Noise Matrix with Regularization	84.77	2.77↑
Label Smoothing	86.64	0.59↑

(a) AG-News, 70% uniform noise

Methods	Original	+TAPT
Without Noise-handling	80.12	4.54↑
Co-Teaching	83.77	6.14↑
Noise Matrix	78.82	2.80↑
Noise Matrix with Regularization	80.17	4.15↑
Label Smoothing	80.61	0.10↑

(b) IMDB, 45% uniform noise

Methods	Original	+TAPT
Without Noise-handling	64.72	4.86↑
Co-Teaching	61.37	1.98↑
Noise Matrix	65.96	3.12↑
Noise Matrix with Regularization	61.32	2.14↑
Label Smoothing	65.44	0.59↑

(c) Yorùbá, weak supervision noise

Methods	Original	+TAPT
Without Noise-handling	46.97	0.85↑
Co-Teaching	31.65	4.51↑
Noise Matrix	46.58	0.79↑
Noise Matrix with Regularization	35.36	5.71↑
Label Smoothing	46.44	0.51↑

(d) Hausa, weak supervision noise

Table 1: Performance of the original models and that after applying TAPT to the BERT backbone. TAPT consistently benifits all models in all noise settings. Performance measured in Accuracy (%). The mean performance over five trials is reported.

other. 2) **Noise Matrix**: A noise transition matrix is appended after BERT's prediction to transform the clean label distribution to the noisy one. A variety of methods exists for estimating the noise matrix, i.e. Sukhbaatar et al. (2015); Bekker & Goldberger (2016); Patrini et al. (2017); Hendrycks et al. (2018); Yao et al. (2020). To exclude the effects of estimation errors in the evaluation, we use the ground truth transition matrix as it is the best possible estimation. This matrix is fixed after initialization. 3): **Noise Matrix with Regularization**: Similar to *Noise Matrix*, Jindal et al. (2019) appends a noise matrix after BERT's output. During training, the matrix is learned with an $l2$ regularization and is not necessarily normalized to be a probability matrix. This method is the current state-of-the-art method for text classification with simulated noise 4) **Label Smoothing**: label smoothing Szegedy et al. (2016) is a commonly used method to improve model's generalization and calibration. It mixes the one-hot label with a uniform vector, preventing the model from getting overconfident on the samples. Lukasik et al. (2020) further shows that it improves noise robustness. Finally, for the sake of comparison, we also include a model, denoted by **No Validation**, which neither applies noise-handling nor early-stopping.

Results We evaluate our baselines on four datasets with different noise levels and noise types. We find that BERT is robust to simulated noise, especially under low to mild noise levels. For example, the performance of BERT without noise-handling drops less than 5% under 60% uniform noise in AG-News. In these cases, noise-handling methods rarely outperform the baseline without noise-handling. When the noise levels go higher, noise-handling starts benefiting learning. However, the results are mixed. For example, Co-Teaching outperforms the baseline without noise-handling on IMDB with 45% single-flip noise, yet it underperforms in AG-News with 70% uniform noise. Please refer to Appendix A for detailed performance numbers in each noise setting.

Task-adaptive pre-training offers an implicit way to assist BERT in all noise settings we investigate. Table 1 presents the performance difference before and after adding TAPT. Compared to the mixed results we see from applying noise-handling, TAPT helps consistently. We also notice that the performance difference between different methods is reduced. This could be particularly beneficial if computational resources are limited and trying out all noise-handling methods is not possible. In this case, just using TAPT without a complex noise-handling architecture should already offer a strong baseline.

In summary, we observe that the benefit from noise-handling methods is rather limited and sometimes inconsistent, especially under weak noise. TAPT, on the other hand, is a stable method to assist text classifiers when learning with noisy labels.

ACKNOWLEDGMENTS

This work has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102 and the EU Horizon 2020 projects ROXANNE under grant number 833635 and COMPRISE under grant agreement No. 3081705.

REFERENCES

Görkem Algan and Ilkay Ulusoy. Image classification with deep learning in the presence of noisy labels: A survey. *Knowl. Based Syst.*, 215:106771, 2021. doi: 10.1016/j.knosys.2021.106771. URL https://doi.org/10.1016/j.knosys.2021.106771.

Alan Joseph Bekker and Jacob Goldberger. Training deep neural-networks based on unreliable labels. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pp. 2682–2686. IEEE, 2016. doi: 10.1109/ICASSP.2016.7472164. URL https://doi.org/10.1109/ICASSP.2016.7472164.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 8342–8360. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.740. URL https://doi.org/10.18653/v1/2020.acl-main.740.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8536–8546, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a19744e268754fb0148b017647355b7b-Abstract.html.

Michael A. Hedderich, David Ifeoluwa Adelani, Dawei Zhu, Jesujoba O. Alabi, Udia Markus, and Dietrich Klakow. Transfer learning and distant supervision for multilingual transformer models: A study on african languages. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 2580–2591. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.204. URL https://doi.org/10.18653/v1/2020.emnlp-main.204.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2545–2568, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.201. URL https://aclanthology.org/2021.naacl-main.201.

Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 10477–10486, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/ad554d8c3b06d6b97ee76a2448bd7913-Abstract.html.

Ishan Jindal, Daniel Pressel, Brian Lester, and Matthew S. Nokleby. An effective label noise model for DNN text classification. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 3246–3256. Association for Computational

Linguistics, 2019. doi: 10.18653/v1/n19-1328. URL https://doi.org/10.18653/v1/n19-1328.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pp. 6448–6458. PMLR, 2020.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (eds.), *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pp. 142–150. The Association for Computer Linguistics, 2011. URL https://www.aclweb.org/anthology/P11-1015/.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 2233–2241. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.240. URL https://doi.org/10.1109/CVPR.2017.240.

Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6596.

Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune BERT for text classification? In Maosong Sun, Xuanjing Huang, Heng Ji, Zhiyuan Liu, and Yang Liu (eds.), *Chinese Computational Linguistics - 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019, Proceedings*, volume 11856 of *Lecture Notes in Computer Science*, pp. 194–206. Springer, 2019. doi: 10.1007/978-3-030-32381-3_16. URL https://doi.org/10.1007/978-3-030-32381-3_16.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

Michael Tänzer, Sebastian Ruder, and Marek Rei. BERT memorisation and pitfalls in low-resource scenarios. *CoRR*, abs/2105.00828, 2021. URL https://arxiv.org/abs/2105.00828.

Brendan van Rooyen, Aditya Krishna Menon, and Robert C. Williamson. Learning with symmetric label noise: The importance of being unhinged. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 10–18, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/45c48cce2e2d7fbdea1afc51c7c6ad26-Abstract.html.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual T: reducing estimation error for transition matrix in label-noise learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/512c5cad6c37edb98ae91c8a76c3a291-Abstract.html.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17283–17297. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 649–657, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

A DETAILED ON MODEL PERFORMANCE

We summarize our experimental results on simulated noise and weak supervision noise in Table 2 and 3, respectively.

	AG-News							IMDB			
		uniform			single-flip				single-flip		
	clean	40%	60%	70%	20%	40%	45%	clean	20%	40%	45%
NV	94.07±0.13	84.48±0.78	61.61±3.18	43.78±5.07	90.46±0.37	76.06±0.33	64.74±0.94	94.03±0.13	86.34±0.77	65.05±0.90	58.97±1.26
CT	-	92.18±0.21	89.90±0.38	84.74±2.56	93.33±0.12	90.62±0.53	87.99±1.64	-	92.32±0.27	89.36±0.67	83.77 ±3.88
NMat	-	92.25±0.14	89.91±0.48	83.9±1.87	93.91±0.15	93.13 ±0.31	92.93 ±0.51	-	92.07±0.21	87.13±0.44	78.82±1.37
NMwR	93.64±0.06	92.02±0.20	89.91 ±0.33	84.77±2.24	93.03±0.17	90.23±0.65	88.93±0.68	93.68±0.14	92.12±0.35	85.94±0.86	80.17±2.57
LS	94.43±0.19	92.45 ±0.21	89.79±0.38	86.64±0.78	93.56±0.23	92.40±0.33	90.94±0.86	94.06 ±0.09	92.13±0.43	87.22±1.39	80.61±2.48
WN	94.40±0.13	92.40±0.25	89.53±0.75	85.49±0.76	93.80±0.08	92.33±0.35	88.94±0.92	93.98±0.15	92.13±0.21	85.88±2.78	80.12±4.09
TAPT+NV	94.71±0.12	82.58±1.19	54.84±0.96	36.56±0.89	90.89±0.13	71.51±1.00	61.02±1.75	94.82±0.06	87.03±1.55	64.01±1.52	58.38±0.90
TAPT+CT	-	93.14 ±0.25	90.67±0.27	85.64±2.12	94.09±0.24	91.06±0.51	88.16±2.16	-	93.93 ±0.14	91.08 ±0.87	89.91 ±0.48
TAPT+NMat	-	93.03±0.31	90.79±0.22	86.70±1.27	94.41 ±0.12	94.07 ±0.26	93.64±0.14	-	92.32±0.40	89.69±0.47	83.93±2.99
TAPT+NMwR	94.29±0.11	92.73±0.19	90.43±0.60	87.54±1.19	94.05±0.08	93.50±0.19	93.41±0.16	94.84±0.08	93.58±0.27	89.61±1.60	85.32±3.05
TAPT+LS	94.81±0.04	92.85±0.23	90.60±0.53	87.23±0.97	94.28±0.14	93.03±0.53	90.80±1.12	94.79±0.06	93.21±0.45	90.27±1.16	80.71±4.74
TAPT+WN	94.85 ±0.09	92.94±0.18	90.96±0.82	87.62 ±0.92	94.35±0.08	92.87±0.61	90.94±1.25	94.86 ±0.05	93.51±0.35	90.81±0.69	84.66±3.25

Table 2: Average test accuracy (%) and standard deviation (5 trials) on AG-News and IMDB with uniform and single-flip noise. NV: without noise-handling and no validation set, i.e. train the model without noise-handling and until the training loss converges. CT: Co-Teaching. NMat: Noise Matrix. NMwR: Noise Matrix with Regularization. LS: Label Smoothing. CT and NMat are equivalent to WN in the clean setting. TAPT+[XX]: Task adaptive pre-training followed by the method XX. Note that as IMDB is a binary-classification task, single-flip noise is equivalent to the uniform noise in this case.

	FT		TAPT+FT	
	Yorùbá	Hausa	Yorùbá	Hausa
NV	63.88±1.59	46.98±1.01	66.78±1.38	47.32±0.85
CT	61.37±1.58	31.65±2.71	63.35±1.21	36.16±5.66
NMat	65.96 ±0.81	46.58 ±0.88	69.08±1.56	47.37±0.69
NMwR	61.32±0.71	35.36±3.60	63.46±1.44	41.07±3.75
LS	65.44±1.67	46.36±0.78	69.16±1.35	46.87±1.06
WN	64.72±1.45	46.39±0.81	69.58 ±1.58	46.44±0.60

Table 3: Average test accuracy (%) and standard deviation (10 trials) on Yorùbá and Hausa with noise from weak supervision. FT: direct fine-tuning on text classification task NV: without noise-handling and no validation set, i.e. train the model without noise-handling and until the training loss converges. CT: Co-Teaching. NMat: Noise Matrix. NMwR: Noise Matrix with Regularization. LS: Label Smoothing. CT and NMat are equivalent to WN in the clean setting.