Is BERT Robust to Label Noise? A Study on Learning with Noisy Labels in Text Classification

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

Abstract

Incorrect labels in training data occur when human annotators make mistakes or when the data is generated via weak or distant supervision. It has been shown that complex noise-handling techniques - by modeling, cleaning or filtering the noisy instances - are required to prevent models from fitting this label noise. However, we show in this work that, for text classification tasks with modern NLP models like BERT, over a variety of noise types, existing noise-handling methods do not always improve its performance, and may even deteriorate it, suggesting the need for further investigation. We also back our observations with a comprehensive analysis.

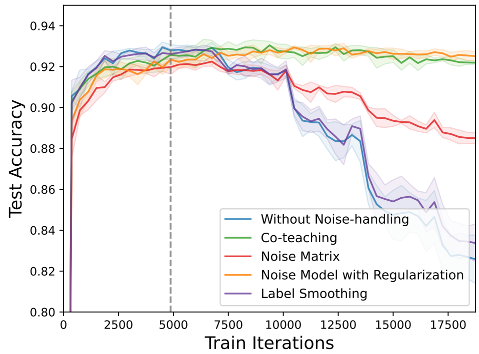


Figure 1: A typical training curve when learning with noise. Learning without noise-handling (blue) will reach a peak accuracy before memorizing the noise. Early-stopping on a noisy validation set (vertical grey line) is often sufficient to find such a peak. Injected uniform noise of 40% on AG-News dataset.

1 Introduction

For many languages, domains and tasks, large datasets with high-quality labels are not available. To tackle this issue, cheaper data acquisition methods have been suggested, such as crowdsourcing or automatic annotation methods like weak and distant supervision. Unfortunately, compared to gold-standard data, these approaches come with more labeling mistakes, which are known as noisy labels. Noise-handling has become an established approach to mitigate the negative impact of learning with noisy labels. A variety of methods have been proposed that model the noise, or clean and filter the noisy instances (Hedderich et al., 2021; Algan and Ulusoy, 2021). Jindal et al. (2019) show e.g. a 30% boost in performance after applying noise-handling techniques on a CNN-based text classifier.

In a recent work, Tänzer et al. (2021) showed that BERT (Devlin et al., 2019) has an inherent robustness against noisy labels. The generalization performance on the clean distribution drops only slowly with the increase of the mislabeled samples. Also, they show that early-stopping is crucial for learning with noisy labels as BERT will eventually memorize all wrong labels when trained long enough. However, their experiments only focus on a single type of noise and a limited range of noise levels. It remains unclear if BERT still performs robustly under a wider range of noise types and with higher fractions of mislabeled samples. Moreover, they perform early-stopping on a clean validation set, which may not be available under low resource settings. Last but not least, they do not compare to any noise-handling methods.

In this work, we investigate the behaviors of BERT on tasks with different noise types and noise levels. We also study the effect of noise-handling methods under these settings. Our main results include **(1)** BERT is robust against injected noise, but could be vulnerable to noise from weak supervision. In fact, the latter, even at a low level, can be more challenging than high injected noise. **(2)** Existing noise-handling methods do not improve the peak performance of BERT under any noise settings we investigated; as is shown with further analysis, noise-handling methods rarely render the correct labels more distinguishable from the incorrect ones. [1]

[1]Our implementation is available on: https://github.com/uds-lsv/BERT-LNL.

Proceedings of the Third Workshop on Insights from Negative Results in NLP, pages 62 - 67
May 26, 2022

2 Learning with Noisy Labels

Problem Settings We consider a k-class classification problem. Let D denote the true data generation distribution over $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is the feature space and $\mathcal{Y} = \{1, ..., k\}$ is the label space. In a typical classification task, we are provided with a training dataset $S = \{(x_i, y_i)_{i=1}^n\}$ sampled from D. In learning with noisy labels, however, we have no access to D. Instead, a noisy training set $\hat{S} = \{(x_i, \hat{y}_i)_{i=1}^n\}$ sampled from a label-corrupted data distribution $\hat{D}$. The goal is to learn a classifier that generalizes well on the clean distribution by only exploiting $\hat{S}$.

Injected Label Noise To rigorously evaluate noise-handling methods at different noise levels, researchers in this area often construct noisy datasets from clean ones by injecting noise. This can, e.g., reflect annotation scenarios such as crowdsourcing, where some annotators answer randomly or prefer an early entry in a list of options. Modeling such noise is achieved by flipping the labels of the clean instances according to a pre-defined noise level $\varepsilon \in [0, 1)$ and a noise type. There are two commonly used noise types: the single-flip noise (Reed et al., 2015):

$$p_{\text{sflip}}(\hat{y} = j | y = i) = \begin{cases} 1 - \varepsilon, & \text{for } i = j \\ \varepsilon, & \text{for one } i \neq j \\ 0, & else \end{cases}$$

and uniform-flip (van Rooyen et al., 2015) noise

$$p_{\text{uni}}(\hat{y} = j | y = i) = \begin{cases} 1 - \varepsilon, & \text{for } i = j \\ \frac{\varepsilon}{k-1}, & \text{for } i \neq j \end{cases} .$$

These noise generation processes are feature-independent, i.e. $p(\cdot|y = i, x) = p(\cdot|y = i)$. Therefore, they can be described by a noise transition matrix T with $T_{ij} := p(\hat{y} = j | y = i)$. It is usually assumed that the noise is diagonally-dominant when generating the noisy labels, i.e. $\forall i, T_{ii} > max_{j \neq i} T_{ij}$.

Label Noise from Weak Supervision Distant and weak supervision (Mintz et al., 2009; Ratner et al., 2016) have become essential methods to acquire labeled data in low-resource scenarios. The resulting noise, unlike injected noise, is often feature-dependent (Lange et al., 2019). We evaluate our methods on two real-world datasets in Hausa and Yorùbá to cover this type of noise.

Dataset	Classes	Average Lengths	Train Samples	Validation Samples	Test Samples	Train Noise Level
IMDB	2	292	21246	3754	25000	various
AG-News	4	44	108000	12000	7600	various
Yorùbá	7	13	1340	189	379	33.28%
Hausa	5	10	2045	290	582	50.37%

Table 1: Statistics of the text classification datasets. The train noise level is the false discovery rate (i.e. 1-precision) of the noisy labels in the training set. The original AG-News has 120k training instances and no validation instances. We therefore held-out 10% of the training samples for validation.

3 Early-Stopping on Noisy Validation Set

When trainied on noisy data without noise-handling, BERT reaches a high generalization performance before it starts fitting the noise. Then it memorizes the noise and the performance on clean distribution drops dramatically (the blue curve in Figure 1). Hence, for models without noise-handling, it is crucial to stop training when the generalization performance reaches its maximum.

Tänzer et al. (2021) use a clean validation set to find this point. However, a clean validation set is often unavailable in realistic low-resource scenarios as it requires manual annotation. Therefore, we use a noisy validation set for early-stopping in all of our experiments and we attain models that generalize well on the clean distribution.

In our example in Figure 1, we see that while most noise-handling methods prevent BERT from fitting the noise in the long run, their peak performance is not significantly higher than a vanilla model without noise-handling.

4 Experiments

Dataset Construction We experiment with four text classification datasets: two benchmarks, AG-News (Zhang et al., 2015) and IMDB (Maas et al., 2011), injected with different levels of single-flip or uniform noise; for the weakly supervised noise, we make use of two news topics datasets in two low-resource languages: Hausa and Yorùbá (Hedderich et al., 2020). Hausa and Yorùbá are the second and the third most spoken indigenous language in Africa, with 40 and 35 million native speakers, respectively (Eberhard et al., 2019). The noisy labels were gazetteered. For example, to identify texts for the class "Africa", a labeling rule based on a list of African countries and their capitals is used. Note that while we can vary the noise levels of injected noise, the amount of weak supervision noise in Hausa and Yorùbá is fixed[2]. We summarize some basic statistics of the datasets in Table 1.

Implementation We use of-the-shelf BERT models for our tasks. Specifically, we apply the BERT-base model for AG-News and IMDB, and the mBERT-base for Yorùbá and Hausa. The fine-tuning approach follows (Devlin et al., 2019). In all settings, we apply early-stopping on a noisy validation set to mimic the realistic low-resource settings, while the test set remains clean. For more implementation details and a discussion on clean and noisy validation sets, see Appendix B and E.

4.1 Baselines

We compare learning without noise-handling with four popular noise-handling methods.[3]

Without Noise-handling Train BERT on the noisy training set as it was clean. A noisy validation set is used for early-stopping.

No Validation For the sake of comparison, we train the model without noise-handling and until the training loss converges.

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 and 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.

Noise Matrix with Regularization The previous state-of-the-art for text classification with noisy labels (Jindal et al., 2019). Similar to *Noise Matrix*, it 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. In the original implementation they use CNN-based models as backbone, we switch it to BERT for fair comparison.

Co-teaching Han et al. (2018) Train two networks to pick cleaner training subsets for each other. The Co-teaching framework requires an estimation of the noise level. Similarly to NMat, we use the ground truth noise level to exclude the performance drop caused by estimation error.

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.

4.2 Experimental Results

We evaluate our baselines on both injected noise (on AG-News and IMDB) and weak supervision noise (on Hausa and Yorùbá). The test accuracy is presented Figure 2. On injected noise, our results match and extend the findings by Tänzer et al. (2021) that BERT is noise robust. For example, the test accuracy drops only about 10% after injecting 70% wrong labels (Figure 2(a)). However, we find that BERT is vulnerable under weak supervision noise. The performance can drop up to 35% in a dataset like Hausa with 50% weak supervision noise compared to training with clean labels (Figure 2(c)). This indicates that the experience on injected noise may not be transferable to weak supervision noise.

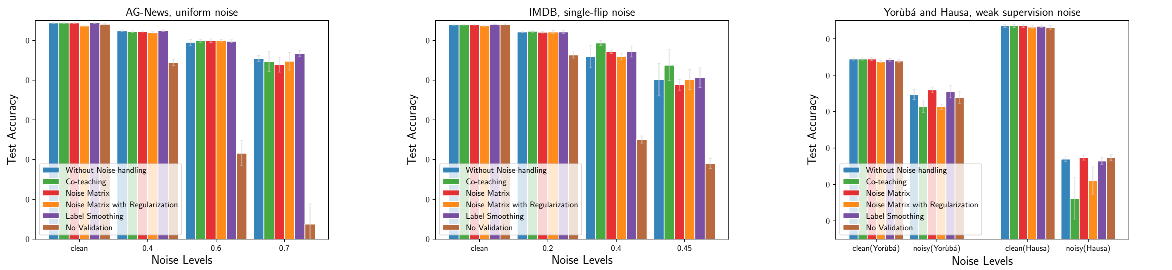
We also observe that noise-handling methods are not always helpful. For injected noise, the benefits from noise-handling become obvious only under high noise levels. But even then, there is no clear winner, meaning that it is hard to decide beforehand which noise method to apply - with the risk that they may even perform worse than BERT without noise-handling. The same applies to weak supervision noise. The maximal performance gap between the best model and BERT without noise-handling is less than 4% and 1.5% under injected noise and weak supervision noise, respectively.

4.3 Analysis of Loss Distributions

To shed some light on why BERT is robust against injected noise but not weak supervision noise, we track the losses on correctly and wrongly labeled samples during training. Figure 4 depicts typical distributions of losses associated with correctly and incorrectly labeled samples, respectively, when early-stopping is triggered. We see that they have minimal overlap, thus different behaviors throughout the training, potentially allowing the model to distinguish correctly and incorrectly labeled sam-

[2]refer to Appendix A for detailed noise distribution.

[3]For a fair comparison, early-stopping on a noisy validation set is applied to all four noise-handling methods.

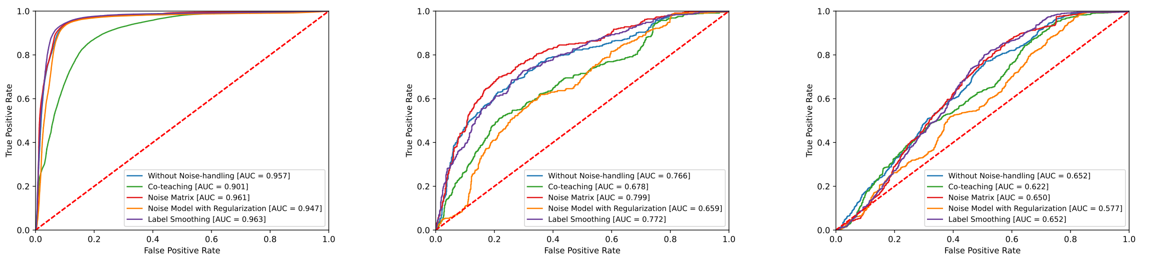


(a) AG-News, uniform noise

(b) IMDB, single-flip noise

(c) Yorùbá & Hausa, weak supervision noise

Figure 2: Test accuracy in different noise settings. a) & b) injected noise with different noise levels c) weak supervision noise, at noise levels of 33.28% and 50.37% in Yorùbá and Hausa, respectively. Noise-handling methods do not always improve peak performances. Further plots in Appendix C.



(a) AG-News - 70% uniform noise

(b) Yorùbá - weak supervision noise

(c) Hausa - weak supervision noise

Figure 3: ROC curves on wrong label detection (binary classification) using the losses. The losses are recorded at the training step when early-stopping is triggered. Noise-handling methods do not make the losses of correct and incorrect labels more distinguishable. Further plots in Appendix D.

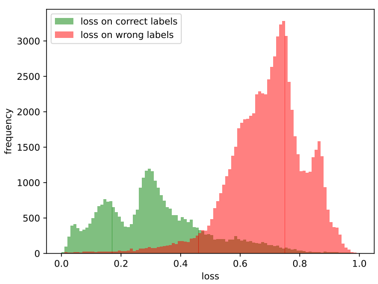


Figure 4: Loss histogram at the training iteration when the early-stopping is triggered. AG-News dataset with 70% uniform noise.

ples from each other. We could further quantify the difference by their separability. Figure 3 presents the receiver operating characteristic (ROC) curves of a thresholds-based classifier. We observe that **(1)** under injected noise, an area under curve (AUC) of more than 90 can be easily achieved without noise-handling (Figure 3(a)), supporting our observation that injected noise has rather a low impact on BERT. **(2)** Under weak-supervision noise, the AUC score is significantly lower, which means the correct and incorrect labels are less distinguishable. Therefore, BERT fits both labels at similar rates. One reason could be that the noise in weak supervision is often feature-dependent, it might become easier for BERT to fit them, which in turn deteriorates the generalization. **(3)** We do not observe a raise in AUC scores when applying noise-handling methods, indicating that noise-handling methods rarely enhance BERT's ability to further avoid the negative impact of wrong labels. This is consistent with the observation in Section 4.2 that noise-handling methods have little impact on BERT's generalization performance.

5 Conclusion

On several text classification datasets and for different noise types, we showed that BERT is noise resistant under injected noise, but not necessarily under weak supervision noise. In both cases, the improvement obtained by applying noise-handling methods are limited. Our analysis on the separability of losses corresponding to correct and incorrect labeled samples provides evidence to this argument. Our analysis offers both motivation and insights to further improve label noise-handling methods and make them useful on more realistic types of noise.

6 Broader Impact Statement and Ethics

Noisy labels are a cheaper source of supervision. This could make it easier to use machine learning for improper use cases. However, it also opens up NLP methods for low-resource settings such as under-resourced languages or applications developed by individuals or small organizations. It can, therefore, be a step towards the democratization of AI.

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. 2021. Image classification with deep learning in the presence of noisy labels: A survey. *Knowl. Based Syst.*, 215:106771.

Alan Joseph Bekker and Jacob Goldberger. 2016. 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*, pages 2682–2686. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). 2019. Ethnologue: Languages of the world. twenty-second edition.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8536–8546.

Michael A. Hedderich, David Ifeoluwa Adelani, Dawei Zhu, Jesujoba O. Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on african languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2580–2591. Association for Computational Linguistics.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. 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*, pages 2545–2568, Online. Association for Computational Linguistics.

Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. 2018. Using trusted data to train deep networks on labels corrupted by severe noise. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 10477–10486.

Ishan Jindal, Daniel Pressel, Brian Lester, and Matthew S. Nokleby. 2019. An effective label noise model for DNN text classification. In *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)*, pages 3246–3256. Association for Computational Linguistics.

Lukas Lange, Michael A. Hedderich, and Dietrich Klakow. 2019. Feature-dependent confusion matrices for low-resource NER labeling with noisy labels. 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)*, pages 3554–3559, Hong Kong, China. Association for Computational Linguistics.

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

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 142–150. The Association for Computer Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 2017. 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*, pages 2233–2241. IEEE Computer Society.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. 2015. Training deep neural networks on noisy labels with bootstrapping. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

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

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

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

Brendan van Rooyen, Aditya Krishna Menon, and Robert C. Williamson. 2015. Learning with symmetric label noise: The importance of being unhinged. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 10–18.

Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. 2020. Dual T: reducing estimation error for transition matrix in label-noise learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.