Handling Realistic Label Noise in BERT Text Classification

Maha Tufail Agro
MBZUAI
maha.agro@mbzuai.ac.ae

Hanan Aldarmaki
MBZUAI
hanan.aldarmaki@mbzuai.ac.ae

Abstract

Label noise refers to errors in training labels caused by cheap data annotation methods, such as web scraping or crowd-sourcing, which can be detrimental to the performance of supervised classifiers. Several methods have been proposed to counteract the effect of random label noise in supervised classification, and some studies have shown that BERT is already robust against high rates of randomly injected label noise. However, real label noise is not random; rather, it is often correlated with input features or other annotator-specific factors. In this paper, we evaluate BERT in the presence of two types of realistic label noise: feature-dependent label noise, and synthetic label noise from annotator disagreements. We show that the presence of these types of noise significantly degrades BERT classification performance. To improve robustness, we evaluate different types of ensembles and noise-cleaning methods and compare their effectiveness against label noise across different datasets.

1 Introduction

Deep learning algorithms have been shown to perform extremely well in supervised classification tasks given high-quality datasets. Unfortunately, obtaining gold-standard labels is often prohibitively expensive with large-scale datasets, leading practitioners to resort to cheaper data collection methods such as crowd-sourcing or automatic annotation methods (Yan et al., 2014). These techniques are known to impart a substantial amount of label noise in the data, which can degrade classification performance (Ji et al., 2021). Label noise refers to errors or inconsistencies within the data labels, such that the prescribed labels do not match the gold labels assigned by experts. Datasets obtained through web scraping often contain label noise given the absence of expert-verified gold labels (Li et al., 2017). Due to a meteoric rise in social media usage, more and more datasets are automatically acquired from online social platforms, and such datasets are likely to contain label noise. Small-scale datasets can also suffer from the same problem if the annotation process is challenging or the annotators have divergent opinions (Ma et al., 2019).

Some prior works have been dedicated to developing and deploying algorithms that combat the effects of label noise in text classification (Han et al., 2018; Sukhbaatar et al.; Zhang and Sabuncu, 2018; Jiang et al., 2018). However, most previous studies simulated label noise by random substitution, and recent research has shown empirically that many methods that successfully handle random noise are ineffective against real-world label noise (Jiang et al., 2020). In the text classification domain, Zhu et al. (2022) explored the robustness of previously proposed methods for handling label noise, including noise matrix with regularization (Jindal et al., 2019), co-teaching (Han et al., 2018), and label smoothing (Szegedy et al., 2016). They concluded that BERT (Devlin et al., 2019) is already robust against randomly injected label noise and these approaches obtain no additional performance gains. On the other hand, they find that feature-dependent label noise, which realistically arises from automatic annotation techniques, degrades BERT performance and these noise handling techniques add little to no robustness at all. This creates a need for a comprehensive evaluation of noise-robust methods in the domain of text classification, considering the presence of realistic labeling errors.

In this paper, we describe methods and experiments for handling realistic label noise in BERT text classification. We use two datasets that contain feature-dependent label noise from automatic annotation, namely Yorùbá and Hausa (Hedderich et al.). These two datasets have been manually cleaned, so a clean version of each exists for evaluation. In addition, we use tweetNLP (Gimpel et al., 2011) and SNLI (Bowman et al., 2015) datasets with syn-

thetic noise that mimics human errors by utilizing multiple crowd-sourced annotations (Chong et al., 2022). This collection of datasets provides a range of noise types and levels that more closely reflect realistic label noise compared to random noise injection. We evaluate the performance of vanilla BERT compared with a subset of noise-handling approaches, namely co-teaching (Han et al., 2018), Consensus Enhanced Training Approach (CETA) (Liu et al., 2022), different types of ensembles (Ganaie et al., 2022), and noise cleaning (Chong et al., 2022; Sluban et al., 2014). We summarize our findings as follows:

1. For datasets with feature-dependent label noise, we find that CETA, some types of ensembles, and noise cleaning, all provide positive performance gains compared to vanilla BERT.

2. For synthetic label noise from multiple annotations, we do not observe significant gains using these approaches. We surmise that this type of noise is more challenging or may even reflect inherently ambiguous labels.

It is worth noting that the noise is qualitatively different in these two categories of label noise as the latter arises from human rather than automatic processes, which could be due to either errors or genuine disagreements. Some recent works attempt to include multiple labels in the training process rather than rely on a single gold label to account for the inherent uncertainty from human disagreements. This may be justified given the nature of some tasks, and the noising scheme performed on tweetNLP and SNLI may warrant that kind of treatment or further scrutiny to identify clear-cut errors. However, as we focus on noise robustness as the scope of this work, we treat the synthetic noise int these datasets as labeling errors and leave any further analysis of this sort for future work.

2 Background & Related Works

2.1 Types of Label Noise

Label noise refers to irregularities or inconsistencies within the data labels, where the prescribed label of a data point does not correspond to the true expert label. In other words, noisy instances in this context specifically pertain to inaccuracies or errors in the labeling of the data, rather than any distortions or imperfections in the input data itself.

When observing the effect of label noise, the majority of existing literature in text classification assumes random injection of label noise (Han et al., 2018; Sukhbaatar et al.; Zhang and Sabuncu, 2018). This type of synthetic noising involves randomly permuting a fixed number of labels according to a pre-defined noise level and noise type. Because the process of simulating such noise is entirely random and does not depend on the input data features in any way, this type of noise is also known as *feature-independent* label noise.

In contrast, *feature-dependent* label noise is correlated with input features (Algan and Ulusoy, 2020). Datasets that use distantly or weakly supervised methods to generate labels are prone to this type of label noise. These approaches are often used in low-resource applications where it is impractical or expensive to manually annotate large amounts of data. Relation extraction is one such application that heavily relies on automatic data generation methods as supervised relation extraction methods necessitate an extensive amount of labeled training data (Mintz et al., 2009). In this area, denoising methods such as the ones proposed in Jia et al. (2019), Qin et al. (2018), Liu et al. (2022) and Ma et al. (2021) are specifically developed to address feature-dependent label noise in relation extraction datasets.

Recently, Chong et al. (2022) developed realistic noising methods that mimic how humans make labeling errors by taking advantage of the multiple rounds of annotation that some datasets undergo. During the annotation process, certain subsets of the data are subjected to rigorous validation schemes, such as gold labels assigned by experts, while others are annotated using less stringent methods, such as crowdsourced evaluations. By incorporating varying annotations generated during this process, their approach produces realistic label noise that reflects how humans make errors. We refer to this noising scheme as *pseudo-real-world label* noise.

2.2 Noise-robust methods

Noise-robust methods in the literature include model enhancements such as **robust loss functions**. Robust loss functions are a class of loss functions used to train models in a way that is more resistant to label noise. One such loss function is the family of generalized cross-entropy loss functions (Zhang and Sabuncu, 2018), which are designed to be more

robust to label noise by penalizing the model less for incorrect predictions that are consistent with noisy labels.

Another class of noise-robust approaches is what we refer to as **multi-netowrk training**. This subcategory of methods introduces multiple networks that learn from each other and as such make more informed decisions regarding which data to use to update the model parameters. For instance, co-teaching (Han et al., 2018) includes two models that are trained in parallel, and each model is presented with examples that incur low loss by the other model. Intuitively, correct labels produce small losses in earlier training epochs and noisy labels produce higher losses. Similarly, the Consensus-Enhanced Training Approach (CETA) proposed in Liu et al. (2022) is a methodology for robust sentence-level relation extraction that emphasizes the selection of clean data points during the training process. The denoising technique is applied to establish a robust boundary for classification, preventing inaccurately labeled data from being assigned to the wrong classification space, and the consensus between two divergent classifiers is used to select clean instances for training.

2.3 Noise cleaning approaches

Noise-cleaning aims to automatically segregate clean data from noisy data in order to train the final classifier using a cleaned subset of the original training set. The "small loss trick" is commonly used to identify potentially noisy or mislabeled data. The intuition behind this approach is that noisy data have comparatively higher loss than clean data (Takeda et al., 2021; Han et al., 2018; Jiang et al., 2018; Ji et al., 2021).

Several approaches have been proposed for automatic noise detection, which can be a first step towards noise-cleaning before training a robust classifier. Wheway (2001) used boosting to detect noisy data instances. The approach involves iteratively re-weighing the data points to emphasize those that are most difficult to classify correctly. The resulting model is then used to identify the noisy data points by measuring their contribution to the final model. Sluban et al. (2014) trained multiple classifiers (ensemble) on different subsets of the data and combined their outputs to obtain a noise ranking for each instance. Similarly, Chong et al. (2022) assessed the performance of pre-trained language models as error detectors using clean held-out data. They experiment with the error detection capabilities of individual pre-trained models and an ensemble of pre-trained language models. They find that an ensemble of pre-trained model losses outperforms individual model loss in error detection.

2.4 Label noise & BERT

BERT (Devlin et al., 2019) is a popular pre-trained language model that is frequently used for text classification by fine-tuning on target labels. Some recent studies have shown that BERT is already robust against randomly injected label noise (Zhu et al., 2022), and early stopping is sufficient to prevent overfitting on noisy labels. Zhu et al. (2022) evaluates popular noise robust approaches in BERT text classification such as appending noise transition matrix after BERT's predictions (Sukhbaatar et al.), acquiring the noise transition matrix with $l2$ regularization (Jindal et al., 2019), and multi-network training via co-teaching (Han et al., 2018). They conclude that while BERT appears to be inherently robust to feature-independent noise, none of the above approaches improves BERT's peak performance in the presence of feature-dependent label nose.

3 Methodology

In this work, we evaluate BERT text classification on datasets containing pseudo-real-world label noise and feature-dependent label noise. We do not consider randomly injected label noise as Zhu et al. (2022) have shown BERT to be already robust to this type of synthetic noise. The scope of this work is limited to text classification with BERT following the baselines established by Zhu et al. (2022).

3.1 Datasets

To study feature-dependent label noise, we use two news-topic categorization datasets from two low-resource African languages: Hausa and Yorùbá (Hedderich et al.). These languages are spoken by large populations in Africa, with Hausa being the second most spoken indigenous language, with 40 million native speakers, and Yorùbá being the third most spoken, with 35 million native speakers[1]. For these datasets, gazetteers were used for automatic labeling, which results in feature-dependent label noise. For instance, when identifying texts for the "Africa" class, a labeling rule based on a list of

[1]https://en.wikipedia.org/wiki/Languages_of_Africa

Dataset	Yorùbá	Hausa	TweetNLP	SNLI
Number of classes	7	5	15	3
Average sentence length	13	10	12	21
Train Samples	1340	2045	11565	363043
Validation Samples	189	290	2874	9831
Test Samples	379	582	-	9815
Train Noise Level	33.28%	50.37%	Various	Various

Table 1: Dataset statistics

African countries and their capitals was employed. These datasets were chosen specifically as they contain automatic annotation label noise i.e., weak-supervision/feature-dependent noise in addition to clean versions of the splits, making it possible to establish ground truth. Note that the amount of label noise in Hausa and Yorùbá is fixed.

Furthermore, we use the noising schemes proposed by Chong et al. (2022) to simulate real-world label noise produced by crowd-sourced labeling. Pseudo-real-world label noise is injected in two benchmark datasets: TweetNLP (Gimpel et al., 2011) and Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015). TweetNLP is a part-of-speech tagging dataset developed by scraping Twitter posts. While TweetNLP already contained crowd-sourced labels, it later received separate crowdsource evaluations, allowing access to multiple annotations from separate annotators. The SNLI dataset is a large Natural Language Inference corpus developed at Stanford. The original corpus consists of 570K sentence pairs, manually labeled by experts. Like TweetNLP, a subset of SNLI later received extensive crowdsource evaluation. We noise both TweetNLP and SNLI to three label noise levels: 10%, 20%, and 30%. Data statistics for all datasets are shown in Table 1.

3.2 Baselines

Zhu et al. (2022) already evaluated the noise matrix approach (Sukhbaatar et al.), label smoothing (Szegedy et al., 2016), and co-teaching (Han et al., 2018) on the feature-dependent datasets, Hausa and Yorùbá, and concluded that no gains are observed using these methods. We use the following as baselines to benchmark our experiments using other approaches:

1. **Vanilla BERT**: BERT trained on noisy training data without a noise-handling mechanism, except early stopping on a noisy validation set, as done in Zhu et al. (2022).
2. **Co-teaching** (Han et al., 2018), which simultaneously trains two networks, with each network independently ranking data points based on their loss to guide the other network on which points to be included for training. In other words, each network independently performs noise-cleaning for the other network.

4 Approaches

We experiment with the following approaches as potential methods for improving performance under realistic label noise conditions:

4.1 Consensus-Enhanced Training Approach (CETA)

CETA (Liu et al., 2022) has been proposed as a noise-robust model for relation extraction and has shown promising results. CETA contains two discrepant classifiers that share a single encoder. The focus of CETA is to train the classifiers only in instances where both classifiers have reached a consensus. Such instances are supposedly deemed clean. To achieve consensus, CETA augments the standard cross entropy loss to include predictions from both classifiers and uses the Wasserstein distance (Kantorovich, 2006) as a secondary criterion. In this manner, CETA can also be considered an ensemble learning approach.

4.2 Deep Ensembles

Deep ensembles have been shown to generally exhibit robustness as compared to singular models and reduce overfitting (Ganaie et al., 2022). To that end, we hypothesize that ensembles may excel in noisy classification tasks due to the presence of label noise in the training data, which can cause individual models to learn false correlations between features and labels. By training multiple classifiers and combining their predictions, each model can develop a unique representation of the input data and filter out spurious information, leading to a

more robust classification boundary. While ensembles have been previously proposed for data and label noise detection (Wheway, 2001; Sluban et al., 2014; Chong et al., 2022), their performance as a method of robust text classification with noisy labels has not been established.

We formally define ensembles as follows: Given m classifiers $C_1, C_2, ..., C_m$, each classifier produces probabilities P_{c_i} on a clean test set T, an ensemble of the predictors averages the probabilities of each predictor such that $P_{ensemble} = \sum_{i=1}^{m} \frac{P_{C_i}}{m}$. It should be noted that each ensemble member is trained on either the same noisy training set or a randomly selected subset of the noisy training set, depending on the employed technique, which is described below. Nevertheless, in all scenarios, each member is evaluated on the same clean test set. We experiment with three types of ensembles:

1. **Homogeneous Ensembles** Ensembles that aggregate predictions from the same type of classifier (i.e. vanilla BERT with early stopping), trained with different initializations and hyperparameters.

2. **Heterogeneous Ensembles** Ensembles that aggregate predictions from different types of classifiers. In our experiments, we use vanilla BERT, co-teaching, and CETA as the heterogeneous classifiers in the ensemble.

3. **Boosting** Ensembles that aggregate predictions from the same type of classifier (i.e. vanilla BERT with early stopping), but each classifier is trained on a different subset of the training data.

4.3 Noise Cleaning Based on Fine-Tuned Model Loss

We use the pre-trained language model's ability to identify noisy labels as a way to clean the training set by removing instances with potential label noise. This involves fine-tuning BERT on noisy task-specific training data and evaluating model loss on each instance. Training instances that have a loss higher than the selected threshold are excluded from the training set used to train the final classifier. We tune the loss threshold on a noisy validation set.

To avoid biasing or overfitting the model when computing loss on the same set used for fine-tuning, we employ an N-fold process to calculate the loss only on held-out data points[2]. The process is outlined in Algorithm 1. In summary, we fine-tune the model using a subset of the noisy training set and use the model to identify and remove noisy samples from the held-out validation set using a fixed loss threshold[3]. The process is repeated separately N times using disjoint validation sets to clean the complete training set.

Algorithm 1 Noise Cleaning Algorithm

Input: Noisy train set T, loss threshold t, number of folds f
Output: Cleaned train set T_{clean}
Divide T into f validation subsets: $V_1, \ldots, V_f$
for $i = 1$ to f **do**
    $T_i = T \setminus V_i$
    Train a fine-tuned model M_i on T_i
    Evaluate the model loss L_{V_i} on V_i
    $T_{\text{clean},i} = V_i[L_{V_i} < t]$
end for
$T_{\text{clean}} = \bigcup_{i=1}^{f} T_{clean,i}$
return T_{clean}

5 Experiments and Results

All of the methods evaluated in these experiments incorporate early stopping on noisy validation set as done by Zhu et al. (2022). We use a noisy validation set because obtaining a clean validation set is often not feasible in practice. Moreover, Zhu et al. (2022) show that using a noisy validation set for early stopping is more or less as effective as using a clean validation set.

5.1 Hyperparameters

The number of training steps is optimally set to 3000[4] unless we are required to vary hyperparameter settings for homogeneous ensembles. For homogeneous ensembles, we cycle through a combination of the following hyperparameters: the number of training steps = [2000, 3000, 4000, 5000, 6000], learning rate = [0.0002, 0.0004, 0.0005, 0.00001, 0.00002, 0.00003, 0.00004, 0.00005], patience (for early stopping) = [25, 30, 40, 50], warm-up steps =

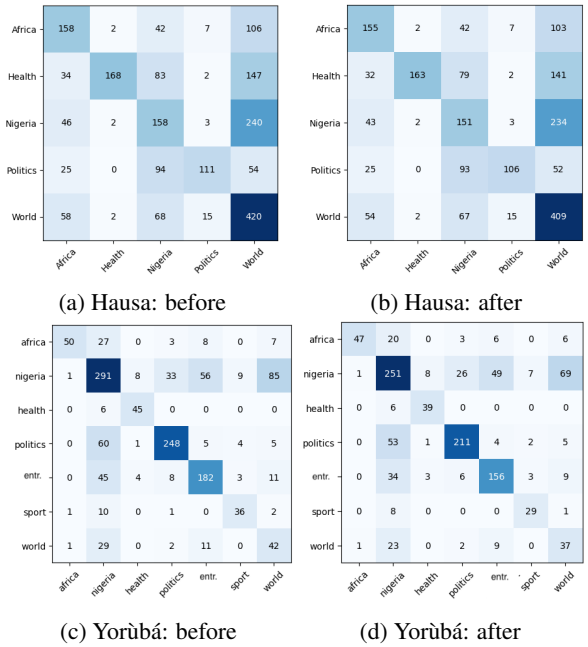
[2]A similar approach is briefly described in (Northcutt et al., 2021) for estimating noise characterization in the confident learning framework.

[3]The loss threshold is a hyperparameter that we tune beforehand.

[4]If the validation accuracy does not improve beyond a certain patience level, we employ early stopping to prematurely halt the training process for all experiments.

	Hausa	Yorùbá
Clean Data		
Vanilla BERT	82.67 ± 0.73	76.23 ± 0.28
Noisy Data		
Vanilla BERT	46.98 ± 1.01	64.72 ± 1.45
Co-Teaching	**48.11** ± 1.71	64.38 ± 0.98
CETA	*__49.31__ ± 0.31	*__68.07__ ± 0.18
HME	46.39 ± 0.21	**67.28** ± 0.81
HTE	**48.28** ± 0.19	**67.81** ± 0.73
Boosting	**47.13** ± 0.42	**67.63** ± 1.26
NC	**47.18** ± 0.22	62.17 ± 0.54

Table 2: A comparison of proposed methods against baselines on datasets with feature-dependent label noise. **HME**: Homogeneous ensembles **HTE**: Heterogeneous ensembles. Boosting: Ensembles of different random subsets from the train set. **NC**: Noise Cleaning. Average accuracy is reported with a standard deviation from 5 runs of each experiment.



(a) Hausa: before (b) Hausa: after

(c) Yorùbá: before (d) Yorùbá: after

Figure 1: Noise matrices for Hausa and Yorùbá showing noise distribution before and after noise cleaning.

[0, 1, 5, 7, 10], weight decay = [0.1, 0.001, 0.0001], and drop rate = [0.1, 0.25, 0.5, 0.8].

For other experiments that do not explicitly require us to vary hyperparameters, we fix the following hyperparameters for the African language datasets, training steps = 3000, learning rate = 0.00005, patience = 25, drop rate = 0.1, warm-up steps = 0, weight decay = 0.1. We fix the following hyperparameters for the English language datasets, training steps = 3000, learning rate = 0.00002, patience = 25, drop rate = 0.25, warm-up steps = 0, weight decay = 0.1. For boosting related experiments, we experiment with two training data subset sizes: 50% of the total training data and 80% of the total training data. For heterogeneous ensembles, we aggregate predictions from the following three classifiers: vanilla BERT, co-teaching, and CETA.

5.2 BERT Models

We use *bert-base-uncased*[5] as the backbone for our English language datasets: TweetNLP and SNLI. We use *bert-base-multilingual-cased*[6] for our African language datasets: Yorùbá and Hausa.

5.3 Loss threshold

To select a loss threshold for noise-cleaning as described in section 4.3, we experiment with different cut-off points in the following interval $[6.0, 8.0]$. We use only a noisy validation set to select the loss threshold. Data points whose loss exceeds the fixed loss threshold are excluded from the training set, effectively 'cleaning' the noisy training set to a certain extent. Note that we only report results on the loss threshold that produces the most optimal accuracy on the noisy validation set. The cleaned training set is once again used to train a vanilla BERT model, at which point we can evaluate how well the noising scheme performed.

5.4 Results

5.5 Feature-dependent label noise

Table 2 shows the results of baseline models and the proposed approaches on datasets containing feature-dependent label noise: Hausa and Yorùbá.

First, we observe that co-teaching and noise cleaning do not consistently improve performance compared to vanilla BERT. CETA, on the other hand, improves performance by around 3 absolute percentage points on both datasets. The homogeneous ensemble method does not consistently improve either, but we do observe consistent gains using heterogeneous ensembles and boosting.

Figure 1 show the noise distribution in the training set before and after applying the noise cleaning procedure in both datasets. Note that the noise-cleaning method results in the removal of both noisy and clean instances, which leads to the total noise level not being considerably reduced. Overall, we we do not observe a larger reduction in noise level in either dataset. After noise cleaning, we

[5]https://huggingface.co/bert-base-uncased

[6]https://huggingface.co/bert-base-multilingual-cased

	TweetNLP			SNLI		
Noise Level	10%	20%	30%	10%	20%	30%
Clean Data						
Vanilla BERT	91.03 ± 0.81			85.03 ± 0.16		
Noisy Data						
Vanilla BERT	82.08 ± 0.03	74.45 ± 0.65	72.96 ± 1.42	84.79 ± 0.87	**83.83** ± 1.01	82.01 ± 0.21
Co-Teaching	81.31 ± 0.11	73.68 ± 0.04	72.41 ± 0.71	84.27 ± 0.15	83.10 ± 1.20	80.99 ± 0.04
CETA	81.00 ± 1.81	72.40 ± 1.01	72.13 ± 0.71	84.24 ± 0.01	82.67 ± 0.21	81.02 ± 0.27
HME	81.81 ± 0.05	74.08 ± 0.03	72.53 ± 0.01	85.02 ± 0.12	83.76 ± 0.10	81.99 ± 0.26
HTE	79.13 ± 0.32	**74.90** ± 0.51	72.32 ± 0.97	84.75 ± 0.34	83.64 ± 1.11	81.16 ± 0.97
Boosting	**82.53** ± 0.01	74.27 ± 0.15	**73.52** ± 3.32	**85.38** ± 0.45	83.80 ± 0.81	**82.06** ± 0.41
NC	80.94 ± 0.09	74.55 ± 0.45	72.65 ± 0.19	85.13 ± 0.05	**84.00** ± 0.01	**82.97** ± 1.09

Table 3: A comparison of proposed methods against baselines on TweetNLP and SNLI datasets noised to various levels. HME: Homogeneous ensembles HTE: Heterogeneous ensembles. Boosting: Ensembles of different random subsets from the training set. Average accuracy is reported with the standard deviation from 5 runs of each experiment.

have 31% label noise in Yorùbá compared to 33% before noise cleaning, with only a 2% reduction in noise. For Hausa, the noise level after cleaning is similarly reduced by 3% (47% compared to 50% before cleaning). In summary, we do not find the noise-cleaning method to be an efficient error detector for feature-dependent label noise, as compared to the other noise-robust we use. This is inconsistent with the result in Chong et al. (2022), where they show that language models are suitable for label error detection. However, they also report that an ***ensemble of large*** pre-trained language models is a better error detector than a smaller individual pre-trained model, and in both cases, while models may have good error detection performance, the performance in the underlying task is not necessarily improved.

5.6 Pseudo-real-world label noise

Table 3 shows the results on datasets containing pseudo-real-world label noise, TweetNLP, and SNLI, with three levels: 10%, 20%, and 30%. In these datasets, we observe that performance drops significantly with increased noise levels in TweetNLP, but only small drops in performance are observed in SNLI. We hypothesize that this potentially reflects the inherent difficulty in the natural language inference task, and the gold labels may already be ambiguous even before applying the noising scheme. Table 4 shows samples from both SNLI and TweetNLP datasets before and after injecting noisy labels. In many cases, particularly in SNLI, the given example is rather ambiguous and both labels can be suitable. These are also cases where there are high inter-annotator disagreements.

In terms of noise handling techniques, we observe that all approaches generally do not produce large gains in performance compared to vanilla BERT. Furthermore, many approaches result in slightly worse performance compared to the baseline. Boosting seems like the most robust technique, as it maintains baseline performance at least, while also being effective against feature-dependent label noise. Noise cleaning in this category obtained mixed results. Surprisingly, CETA does not excel over other methods in this particular category. Although it was specifically designed to address feature-dependent label noise, its performance is somewhat inferior to the vanilla BERT baseline when dealing with realistic label noise. We surmise that this type of artificial noise is more challenging as it's based on actual human errors, and may even reflect intrinsic ambiguities in the task, which makes it harder to detect through automatic approaches.

6 Conclusions

In this paper, we described experiments for evaluating different label noise handling techniques within the framework of BERT text classification. We evaluated some multi-network training approaches (i.e. co-teaching and CETA), different types of ensembles (homogeneous, heterogeneous, and boosting), and a noise cleaning technique and compared them with a vanilla BERT fine-tuned model with early stopping. We used two datasets that contain feature-dependent label noise from automatic labeling, as well as two datasets with synthetic pseudo-real-world label noise obtained by considering multiple

Dataset	Text	Noisy Label	Actual Label
SNLI	(1) Young man wearing a blue jacket, green shirt and denim jeans is photographed by person in beige jacket and burgundy pants while four onlookers watch on an expanse of sand.<!SEP!> The people are ignoring the man getting photographed.	No Relationship	~~Contradiction~~
SNLI	(2) A man wearing a black t-shirt is playing seven string bass a stage.<!SEP!> The man is playing an old guitar.	Contradiction	~~No Relationship~~
SNLI	(3) Many children are sitting in a classroom watching a woman in the front.<!SEP!>The woman is teaching the children	Entailment	~~No Relationship~~
TweetNLP	(1) Reading harry potter in bed! waiting for the new south park to come on	ADJ	~~NOUN~~
TweetNLP	(2) @USER: I'm not insulted, at all, trust me. I'm seeking to understand you and your video. :)	DET	~~ADP~~
TweetNLP	(3) Chicagoan early voters in Uptown even get brownies and entertainment while waiting for a dozen people to do number page ballots.	ADJ	~~NOUN~~

Table 4: Samples from SNLI and TweetNLP with pseudo-real-world noise injection, highlighting the complexity and potential ambiguity of these tasks.

annotations.

For feature-dependent label noise, the recently proposed Consensus Enhanced Training Approach (CETA) shows the most promising results compared to the baselines. Some ensembling techniques, such as boosting, can also improve performance compared to the baselines but do not provide the level of robustness achieved via CETA.

While pre-trained language models have been shown previously to have the potential to detect label errors through out-of-sample loss, our results indicate that using this technique to automatically clean the data does not result in improved performance compared to using the noisy set. This may suggest that removing label errors is not necessarily a good approach for handling label noise; rather, error detection can be used to identify noisy labels for manual correction.

The synthetic pseudo-real-world category of label noise appears to be more challenging as the noise represents actual human errors, which could be an indication of inherent ambiguities in the task itself. Our experiments show that most techniques do not improve performance compared to the baselines. Furthermore, for a dataset like SNLI, which is known to be challenging even for human annotators, the presence of label noise does not reduce the performance to a great extent compared to the other datasets. This may suggest that the noising scheme is compatible with the inherent difficulty or label ambiguity of the task, and any attempts to detect or discard the noise will not necessarily improve the performance using stringent metrics such as accuracy. Recent efforts to embrace annotator disagreements and incorporate them in the training process (Zhang et al., 2021) rather than relying on a single gold label may be more suitable to handle this kind of labeling inconsistencies.

Overall, the results indicate that handling realistic label noise in text classification remains a challenging task, and none of the noise-handling techniques examined so far has shown consistent performance improvements across multiple datasets.

Limitations

The work described in this paper is limited by the small number of datasets that contain both noisy and clean versions in the text classification domain, which are needed for evaluating noise-handling methods. While we observed positive results from at least two approaches, any conclusions we make about their effectiveness are drawn from a sample of two datasets, and may not necessarily generalize to other cases. For the pseudo-real-world label noise category, it is unclear whether the noise represents true errors or inherent ambiguity in the task. The mixed results we observe could also be a result of ambiguities in the presumed 'clean' test set.

References

Gorkem Algan and Ilkay Ulusoy. 2020. Label noise types and their effects on deep learning. *ArXiv*, abs/2003.10471.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page unknown. Association for Computational Linguistics.

Derek Chong, Jenny Hong, and Christopher Manning. 2022. Detecting label errors by using pre-trained

language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9074–9091. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

M.A. Ganaie, Minghui Hu, A.K. Malik, M. Tanveer, and P.N. Suganthan. 2022. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. 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 - Short Papers*, pages 42–47. The Association for Computer Linguistics.

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 *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8536–8546, Red Hook, NY, USA. Curran Associates Inc.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 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)*, pages 2580–2591. Association for Computational Linguistics.

Daehyun Ji, Dokwan Oh, Yoonsuk Hyun, Oh-Min Kwon, and Myeong-Jin Park. 2021. How to handle noisy labels for robust learning from uncertainty. *Neural Networks*, 143:209–217.

Wei Jia, Dai Dai, Xinyan Xiao, and Hua Wu. 2019. ARNOR: Attention regularization based noise reduction for distant supervision relation classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1399–1408, Florence, Italy. Association for Computational Linguistics.

Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. 2020. Beyond synthetic noise: Deep learning on controlled noisy labels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4804–4815. PMLR.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2304–2313. PMLR.

Ishan Jindal, Daniel Pressel, Brian Lester, and Matthew 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, Volume 1 (Long and Short Papers)*, pages 3246–3256, Minneapolis, Minnesota. Association for Computational Linguistics.

Leonid V Kantorovich. 2006. On the translocation of masses. *Journal of mathematical sciences*, 133(4):1381–1382.

Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L. Li. 2017. Learning from noisy labels with distillation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1928–1936. IEEE Computer Society.

Ruri Liu, Shasha Mo, Jianwei Niu, and Shengda Fan. 2022. CETA: A consensus enhanced training approach for denoising in distantly supervised relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2247–2258, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kede Ma, Xuelin Liu, Yuming Fang, and Eero P. Simoncelli. 2019. Blind image quality assessment by learning from multiple annotators. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2344–2348.

Ruotian Ma, Tao Gui, Linyang Li, Qi Zhang, Xuanjing Huang, and Yaqian Zhou. 2021. SENT: Sentence-level distant relation extraction via negative training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6201–6213, Online. Association for Computational 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.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. *CoRR*, abs/2103.14749.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. DSGAN: Generative adversarial training for distant supervision relation extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Melbourne, Australia. Association for Computational Linguistics.

Borut Sluban, Dragan Gamberger, and Nada Lavrač. 2014. Ensemble-based noise detection: noise ranking and visual performance evaluation. *Data Mining and Knowledge Discovery*, pages 265–303.

Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training Convolutional Networks with Noisy Labels. *arXiv e-prints*, page arXiv:1406.2080.

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

Hiroshi Takeda, Soh Yoshida, and Mitsuji Muneyasu. 2021. Training robust deep neural networks on noisy labels using adaptive sample selection with disagreement. *IEEE Access*, pages 141131–141143.

Virginia Wheway. 2001. Using boosting to detect noisy data. In *Advances in Artificial Intelligence. PRICAI 2000 Workshop Reader*, pages 123–130. Springer Berlin Heidelberg.

Yan Yan, Rómer Rosales, Glenn Fung, Ramanathan Subramanian, and Jennifer Dy. 2014. Learning from multiple annotators with varying expertise. pages 291—-327.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Learning with different amounts of annotation: From zero to many labels. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7620–7632.

Zhilu Zhang and Mert R. Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8792–8802, Red Hook, NY, USA. Curran Associates Inc.

Dawei Zhu, Michael A. Hedderich, Fangzhou Zhai, David Ifeoluwa Adelani, and Dietrich Klakow. 2022. Is bert robust to label noise? a study on learning with noisy labels in text classification. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP (Insights at acl), 2022, Dublin, Ireland, May 26*, pages 62–67. Association for Computational Linguistics.