AlleNoise - large-scale text classification benchmark dataset with real-world label noise

Alicja Rączkowska* **Aleksandra Osowska-Kurczab*** **Jacek Szczerbiński***
Kalina Jasinska-Kobus* **Klaudia Nazarko***
Machine Learning Research
Allegro.com
{alicja.raczkowska, aleksandra.kurczab, jacek.szczerbinski,
kalina.kobus, klaudia.nazarko}@allegro.com

Abstract

Label noise remains a challenge for training robust classification models. Most methods for mitigating label noise have been benchmarked using primarily datasets with synthetic noise. While the need for datasets with realistic noise distribution has partially been addressed by web-scraped benchmarks such as WebVision and Clothing1M, those benchmarks are restricted to the computer vision domain. With the growing importance of Transformer-based models, it is crucial to establish text classification benchmarks for learning with noisy labels. In this paper, we present *AlleNoise*, a new curated text classification benchmark dataset with real-world instance-dependent label noise, containing over 500,000 examples across approximately 5,600 classes, complemented with a meaningful, hierarchical taxonomy of categories. The noise distribution comes from actual users of a major e-commerce marketplace, so it realistically reflects the semantics of human mistakes. In addition to the noisy labels, we provide human-verified clean labels, which help to get a deeper insight into the noise distribution, unlike web-scraped datasets typically used in the field. We demonstrate that a representative selection of established methods for learning with noisy labels is inadequate to handle such real-world noise. In addition, we show evidence that these algorithms do not alleviate excessive memorization. As such, with *AlleNoise*, we set the bar high for the development of label noise methods that can handle real-world label noise in text classification tasks. The code and dataset are available for download at `https://github.com/allegro/AlleNoise`.

1 Introduction

The problem of label noise poses a sizeable challenge for classification models [1, 2]. With modern deep neural networks, due to their capacity, it is possible to memorize all labels in a given training dataset [3]. This, effectively, leads to overfitting to noise if the training dataset contains noisy labels, which in turn reduces the generalization capability of such models [4–6].

Most previous works on training robust classifiers have focused on analyzing relatively simple cases of synthetic noise [7, 8], either uniform (i.e. symmetric) or class-conditional (i.e. asymmetric). It is a common practice to evaluate these methods using popular datasets synthetically corrupted with label

*Equal contribution

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

noise, such as MNIST [9], ImageNet [10], CIFAR [11] or SVHN [12]. However, synthetic noise is not indicative of realistic label noise and thus deciding to use noisy label methods based on such benchmarks can lead to unsatisfactory results in real-world machine learning practice. Moreover, it has been shown that these benchmark datasets are already noisy themselves [13, 14], so the study of strictly synthetic noise in such a context is intrinsically flawed.

Realistic label noise is instance dependent, i.e. the labeling mistakes are caused not simply by label ambiguity, but by input uncertainty as well [15]. This is an inescapable fact when human annotators are responsible for the labeling process [16]. However, many existing approaches for mitigating instance-dependent noise have one drawback in common - they had to, in some capacity, artificially model the noise distribution due to the lack of existing benchmark datasets [17–22]. In addition, most of the focus in the field has been put on image classification, but with the ever-increasing importance of Transformer-based [23] architectures, the problem of label noise affecting the fine-tuning of natural language processing models needs to be addressed as well. There are many benchmark datasets for text data classification [24–27], but none of them are meant for the study of label noise. In most cases, the actual level of noise in these datasets is unknown, so using them for benchmarking label noise methods is unfeasible.

Moreover, the datasets used in this research area usually contain relatively few labels. The maximum reported number of labels is 1000 [28]. As such, there is a glaring lack of a benchmark dataset for studying label noise that provides realistic real-world noise, a high number of labels and text data at the same time.

We see a need for a textual benchmark dataset that would provide realistic instance-dependent noise distribution with a known level of label noise, as well as a relatively large number of target classes, with both clean and noisy labels. To this end, in this paper we provide the following main contributions:

- We introduce *AlleNoise* - a benchmark dataset for multi-class text classification with real-world label noise. The dataset consists of 502,310 short texts (e-commerce product titles) belonging to 5,692 categories (taken from a real product assortment tree). It includes a noise level of 15%, stemming from mislabeled data points. This amount of noise reflects the actual noise distribution in the data source (Allegro.com e-commerce platform). For each of the mislabeled data instances, the true category label was determined by human domain experts.
- We benchmark a comprehensive selection of well-established methods for classification with label noise against the real-world noise present in *AlleNoise* and compare the results to synthetic label noise generated for the same dataset. We provide evidence that the selected methods fail to mitigate real-world label noise, even though they are very effective in alleviating synthetic label noise.

2 Related work

Several classification benchmarks with real-world instance-dependent noise have been reported in the literature. ANIMAL-10N [29] is a human-labeled dataset of confusing images of animals, with 10 classes and an 8% noise level. CIFAR-10N and CIFAR-100N [30] are noisy versions of the CIFAR dataset, with labels assigned by crowd-sourced human annotators. CIFAR-10N is provided in three versions, with noise levels of 9%, 18% and 40%, while CIFAR-100N has a noise level of 40%. Clothing1M [31] is a large-scale dataset of fashion images crawled from several online shops. It contains 14 classes and the estimated noise rate is 38%. Similarly, WebVision [28] comprises of images crawled from the web, but it is more general - it has 1000 categories of diverse images. The estimated noise level is 20%. DCIC [32] is a benchmark that consists of 10 real-world image datasets, with several human annotations per image. This allows for testing algorithms that utilize soft labels to mitigate various kinds of annotation errors. The maximum number of classes in the included datasets is 10.

(a) **(b)**

Figure 1: Symmetric noise vs. *AlleNoise* in examples. Correct and noisy labels are marked in green and red, respectively. **(a)** Symmetric noise: an electric toothbrush incorrectly labeled as a winter tire is easy to spot, even for an untrained human. **(b)** *AlleNoise*: a ceiling dome is mislabeled as a pendant lamp. This error is semantically challenging and hard to detect. Note: *AlleNoise* dataset does not include images.

Dataset	Modality	Total examples	Classes	Noise level	Clean label
ANIMAL10N	Images	55k	10	8%	✓
CIFAR10N	Images	60k	10	9/18/40%	✓
CIFAR100N	Images	60k	10	40%	✓
WebVision	Images	2.4M	1000	∼20%	✗
Clothing1M	Images	1M	14	∼38%	✗
Hausa	Text	2,917	5	50.37%	✓
Yorùbá	Text	1,908	7	33.28%	✓
NoisyNER	Text	217k	4	unspecified	✓
AlleNoise	**Text**	**500k**	**5692**	**15%**	✓

Table 1: Comparison of *AlleNoise* to previously published datasets created for studying the problem of learning with noisy labels. All datasets contain real-world noise. *AlleNoise* is the biggest text classification dataset in this field, has a known level of label noise and provides clean labels in addition to the noisy ones.

With the focus in the label noise field being primarily on images, the issue of noisy text classification remains relatively unexplored. Previous works have either utilized existing classification datasets with synthetic noise [14, 17, 33] or introduced new datasets with real-world noise. NoisyNER [34] contains annotated named entity recognition data in the Estonian language, assigned to 4 categories. The authors do not mention the noise level, only that they provide 7 variants of real-world noise. NoisywikiHow [35] is a dataset of articles scraped from the wikiHow website, with accompanying 158 article categories. The data was manually cleaned by human annotators, which eliminated the real-world noise distribution. The authors performed experiments by injecting synthetic noise into their dataset. Thus, NoisywikiHow is not directly comparable to *AlleNoise*. Another two datasets are Hausa and Yorùbá [36], text classification datasets of low-resource African languages with 5 and 7 categories respectively. They both include real-world noise with the level of 50.37% for the former, and 33.28% for the latter.

While there is a number of text datasets containing e-commerce product data [17, 25, 27], none of them have verified clean labels and in most cases the noise level is unknown. Similarly, classification settings with large numbers (i.e. more than 1000) of classes were not addressed up to this point in the existing datasets (**Tab. 1**).

Offer title	Category label	True category label
Emporia PURE V25 BLACK	352	170
Metal Hanging Lid Rack Suspended	68710	321104
Miraculum Asta Plankton C Active Serum-Booster	5360	89000

Category label	Category name
352	Electronics > Phones and Accessories > GSM Accessories > Batteries
170	Electronics > Phones and Accessories > Smartphones and Cell Phones
68710	Home and Garden > Equipment > Kitchen Utensils > Pots and Pans > Lids
321104	Home and Garden > Equipment > Kitchen Utensils > Pots and Pans > Organizers
5360	Allegro > Beauty > Care > Face > Masks
89000	Allegro > Beauty > Care > Face > Serum

Figure 2: *AlleNoise* consists of two tables: the first table includes the true and noisy label for each product title, while the second table maps the labels to category names.

3 AlleNoise Dataset Construction

We introduce *AlleNoise* - a benchmark dataset for large-scale multi-class text classification with real-world label noise. The dataset consists of 502,310 e-commerce product titles listed on Allegro.com in 5,692 assortment categories, collected in January of 2022. 15% of the products were listed in wrong categories, hence for each entry the dataset includes: the product title, the category where the product was originally listed, and the category where it should be listed according to human experts.

Additionally, we release the taxonomy of product categories in the form of a mapping (category ID $\rightarrow$ path in the category tree), which allows for fine-grained exploration of noise semantics.

3.1 Real-world noise

We collected 75,348 mislabeled products from two sources: 1) customer complaints about a product being listed in the wrong category - such requests usually suggest the true category label, 2) assortment clean-up by internal domain experts, employed by Allegro - products listed in the wrong category were manually moved to the correct category.

The resulting distribution of label noise is not uniform over the entire product assortment - most of the noisy instances belong to a small number of categories. Such asymmetric distribution is an inherent feature of real-world label noise. It is frequently modeled with class-conditional synthetic noise in related literature. However, since the mistakes in *AlleNoise* were based not only on the category name, but also on the product features, our noise distribution is in fact instance-dependent.

3.2 Clean data sampling

The 75,348 mislabeled products were complemented with 426,962 products listed in correct categories. The clean instances were sampled from the most popular items listed in the same categories as the noisy instances, proportionally to the total number of products listed in each category. The high popularity of the sampled products guarantees their correct categorization, because items that generate a lot of traffic are curated by human domain experts. Thus, the sampled distribution was representative for a subset of the whole marketplace: 5,692 categories out of over 23,000, for which label noise is particularly well known and described.

3.3 Post-processing

We automatically translated all 500k product titles from Polish to English. Machine translation is a common part of e-commerce, many platforms incorporate it in multiple aspects of their operation [37, 38]. Moreover, it is an established practice to publish machine-translated text in product datasets [39].

Categories related to sexually explicit content were removed from the dataset altogether. Finally, categories with less than 5 products were removed from the dataset to allow for five-fold cross-validation in our experiments.

4 Methods

4.1 Problem statement

Let $\mathcal{X}$ denote the input feature space, and $\mathcal{Y}$ be a set of class labels. In a typical supervised setting, each instance x_i has a true class label y_i. However, in learning with noisy labels, $\tilde{y}_i$ is observed instead, which is with an unknown probability p (noise level) changed from the true y_i.

In this setting, we train a classifier $f : \mathcal{X} \to \mathcal{Y}$ that generalizes knowledge learnt from a dataset $\mathcal{D}$, consisting of training examples $(x_i, \tilde{y}_i)$. Because $\tilde{y}_i$ can be affected by label noise, the model's predictions $\hat{y}_i = f(x_i)$ might be corrupted by the distribution of noisy labels as well. Maximizing the robustness of such a classifier implies reducing the impact of noisy training samples on the generalization performance. In the *AlleNoise* dataset, x_i corresponds to the product title, $\tilde{y}_i$ is the original product category, and y_i is the correct category.

4.2 Synthetic noise generation

In order to compare the real-world noise directly with synthetic noise, we applied different kinds of synthetic noise to the clean version of *AlleNoise*: the synthetic noise was applied to each instance's true label y_i, yielding a new synthetic noisy label $\tilde{y}_i$. Overall, the labels were flipped for a controlled fraction $p = 15\%$ of all instances. We examined the following types of synthetic noise:

- Symmetric noise: each instance is given a noisy label different from the original label, with uniform probability p.
- Class-conditional pair-flip noise: each instance in class j is given a noisy label $j + 1$ with probability p.
- Class-conditional nested-flip noise: we only flip categories that are close to each other in the hierarchical taxonomy of categories. For example, for the parent category *Car Tires* we perform a cyclic flip between its children categories: *Summer* $\to$ *Winter* $\to$ *All-Season* $\to$ *Summer* with probability p. Thus, the noise transition matrix is a block matrix with a small number of off-diagonal elements equal to p.
- Class-conditional matrix-flip noise: the transition matrix between classes is approximated with the baseline classifier's confusion matrix. The confusion matrix is evaluated against the clean labels on 8% of the dataset (validation split) [8]. The resulting noise distribution is particularly tricky: we flip the labels between the classes that the model is most likely to confuse.

4.3 Model architecture

Next, we evaluated several algorithms for training classifiers under label noise. For a fair comparison, all experiments utilized the same classifier architecture as well as training and evaluation loops. We followed a fine-tuning routine that is typical for text classification tasks. In particular, we vectorized text inputs with XLMRoberta [40], a multilingual text encoder based on the Transformer architecture [23]. To provide the final class predictions, we used a single fully connected layer with a softmax activation and the number of neurons equal to the number of classes. The baseline model uses cross-entropy (CE) as a loss function.

Models were trained with the AdamW optimiser and linear LambdaLR scheduling (warmup steps $=$ 100). We have not used any additional regularization, i.e. weight decay or dropout. Key training parameters, such as batch size (bs $= 256$) and learning rate (lr $= 10^{-4}$) were tuned to maximize the validation accuracy on the clean dataset. All models have been trained for 10 epochs. Training of the baseline model, accelerated with a single NVIDIA A100 40GB GPU, lasted for about 1 hour.

We used five-fold stratified cross-validation to comprehensively evaluate the results of the models trained with label noise. For each fold, the full dataset was divided into three splits: $\mathcal{D}_{train}$, $\mathcal{D}_{val}$, $\mathcal{D}_{test}$, in proportion 72% : 8% : 20%. Following the literature on learning with noisy labels [2], both $\mathcal{D}_{train}$ and $\mathcal{D}_{val}$ were corrupted with label noise, while $\mathcal{D}_{test}$ remained clean.

All of the results presented in this study correspond to the last checkpoint of the model. We use the following format for presenting the experimental results: [m] $\pm$ [s], where m is an average over the five cross-validation folds, while s is the standard deviation. Experiments used a seeded random number generator to ensure the reproducibility of the results.

4.4 Evaluation metrics

Accuracy on the clean test set is the key metric in our study. We expect that methods that are robust to the label noise observed in the training phase, should be able to improve the test accuracy when compared to the baseline model.

Additionally, to better understand the difference between synthetic and real-world noise, we collected detailed validation metrics. The validation dataset $\mathcal{D}_{val}$ contained both instances for which the observed label $\tilde{y}_i$ was incorrect ($\mathcal{D}_{val}^{\texttt{noisy}}$) and correct ($\mathcal{D}_{val}^{\texttt{clean}}$). Noisy observations from $\mathcal{D}_{val}^{\texttt{noisy}}$ were used to measure the memorization metric $\texttt{memorized}_{val}$, defined as a ratio of predictions $\hat{y}_i$ that match the noisy label $\tilde{y}_i$. Notice that our memorization metric is computed on the validation set, contrary to the training set typically used in the literature [41]. Our metric increases when the model not only memorizes incorrect classes from the training observations, but also repeats these errors on unseen observations. Furthermore, we compute accuracy on $\mathcal{D}_{val}^{\texttt{noisy}}$ denoted as $\texttt{correct}_{val}^{\texttt{noisy}}$ and its counterpart on the clean fraction, $\texttt{correct}_{val}^{\texttt{clean}}$.

4.5 Benchmarked methods

We evaluated the following methods for learning with noisy labels: Self-Paced Learning (SPL) [42], Provably Robust Learning (PRL) [43], Early Learning Regularization (ELR) [41], Generalized Jensen-Shannon Divergence (GJSD) [44], Co-teaching (CT) [45], Co-teaching+ (CT+) [46] and Mixup (MU) [47]. Additionally, we implemented Clipped Cross-Entropy as a simple baseline (see Appendix **A**). These approaches represent a comprehensive selection of different method families: novel loss functions (GJSD), noise fltration (SPL, PRL, CCE, CT, CT+), robust regularization (ELR), data augmentation (MU) and training loop modifications (CT, CT+).

These methods are implemented with a range of technologies and software libraries. As such, in order to have a reliable and unbiased framework for comparing them, it is necessary to standardize the software implementation. To this end, we re-implemented these methods using PyTorch (version 1.13.1) and PyTorch Lightning (version 1.5.0) software libraries. We publish our re-implementations and the accompanying evaluation code on GitHub at `https://github.com/allegro/AlleNoise`.

To select the best hyperparameters (see Appendix **A**) for each of the benchmarked algorithms, we performed a tuning process on the *AlleNoise* dataset. We focused on maximizing the fraction of correct clean examples $\texttt{correct}_{val}^{\texttt{clean}}$ within the validation set for two noise types: 15% real-world noise and 15% symmetric noise. The tuning was performed on a single fold selected out of five cross-validation folds, yielding optimal hyperparameter values (**Tab. S1**). We then used these tuned values in all further experiments.

5 Results

The selected methods for learning with noisy labels were found to perform differently on AlleNoise than on several types of synthetic noise. Below we highlight those differences in performance and relate them to the dissimilarities between real-world and synthetic noise.

5.1 Synthetic noise vs *AlleNoise*

The selected methods were compared on the clean dataset, the four types of synthetic noise and on the real-world noise in *AlleNoise* (**Tab. 2**). The accuracy score on the clean dataset did not degrade for any of the evaluated algorithms when compared to the baseline CE. When it comes to the performance on the datasets with symmetric noise, the best method was GJSD, with CCE not too far behind. GJSD increased the accuracy by 1.31 percentage points (p.p.) over the baseline. For asymmetric noise types, the best method was consistently ELR. It significantly improved the test accuracy in comparison to CE, by 1.3 p.p. on average. Interestingly, some methods deteriorated the test accuracy. CT+ was worse than the baseline for all synthetic noise types (by 2.59 p.p., 2.12 p.p., 3.1 p.p., 2.02 p.p. for symmetric, pair-flip, nested-flip and matrix-flip noises, respectively), while SPL decreased the results for all types of asymmetric noise (by 3.63 p.p., 4.2 p.p., 5.17 p.p. for pair-flip, nested-flip and matrix-flip noises, respectively). CT+ seems to perform better for noise levels higher than 15% (see Appendix **B**). On *AlleNoise*, we observed nearly no improvement in accuracy for any of the evaluated algorithms, and CT+, PRL and SPL all deteriorated the metric (by 2.65 p.p., 2.05 p.p. and 4.61 p.p., respectively).

	Clean set	Symmetric	Pair-flip	Nested-flip	Matrix-flip	AlleNoise
CE	**74.85 ± 0.15**	71.97 ± 0.08	71.92 ± 0.08	71.77 ± 0.08	70.75 ± 0.17	63.71 ± 0.11
ELR	74.81 ± 0.11	72.15 ± 0.10	**73.21 ± 0.21**	**73.07 ± 0.11**	**72.02 ± 0.17**	63.72 ± 0.19
MU	74.73 ± 0.09	71.96 ± 0.08	71.95 ± 0.10	71.65 ± 0.14	70.73 ± 0.17	63.65 ± 0.12
CCE	74.80 ± 0.09	73.01 ± 0.10	71.86 ± 0.17	71.62 ± 0.10	70.61 ± 0.10	**63.73 ± 0.22**
CT	*74.85 ± 0.15	72.42 ± 0.13	71.99 ± 0.14	71.55 ± 0.08	70.57 ± 0.18	63.32 ± 0.25
CT+	*74.85 ± 0.15	↓69.38 ± 0.29	↓69.80 ± 0.24	↓68.67 ± 2.59	↓68.73 ± 0.27	↓61.06 ± 0.38
PRL	*74.85 ± 0.15	71.82 ± 0.17	71.95 ± 0.15	71.73 ± 0.16	71.12 ± 0.10	↓61.66 ± 0.17
SPL	*74.85 ± 0.15	72.56 ± 0.10	↓68.29 ± 0.15	↓67.57 ± 0.14	↓65.58 ± 0.15	↓59.10 ± 0.14
GJSD	74.63 ± 0.10	**73.28 ± 0.13**	71.67 ± 0.15	71.40 ± 0.10	70.55 ± 0.17	63.63 ± 0.19

Table 2: Accuracy of the evaluated methods on the clean dataset compared to various noisy datasets with 15% noise level. The noisy datasets include *AlleNoise*, symmetric synthetic noise, and asymmetric synthetic noises: pair-flip, nested-flip, and matrix-flip. * marks cases equivalent to the baseline CE. ↓ marks results significantly worse than the baseline CE. Best results for each noise type are bolded.

5.2 Noise type impacts memorization

To better understand the difference between synthetic noise types and *AlleNoise*, we analyze how the $\texttt{memorized}_{val}^{\texttt{noisy}}$, $\texttt{correct}_{val}^{\texttt{noisy}}$ and $\texttt{correct}_{val}^{\texttt{clean}}$ metrics (see 4.4) evolve over time. Memorization and correctness should be interpreted jointly with test accuracy (**Tab. 2**).

Synthetic noise types are memorized to a smaller extent than the real-world *AlleNoise* (**Fig. 3a**). For the two simplest synthetic noise types, symmetric and pair-flip, the value of $\texttt{memorized}_{val}$ is negligible (very close to zero). For the other two synthetic noise types, nested-flip and matrix-flip, memorization is still low (2-8%), but there are clearly visible differences between the benchmarked methods. While ELR, CT+ and PRL all keep the value of $\texttt{memorized}_{val}^{\texttt{noisy}}$ low for both nested-flip and matrix-flip noise types, it is only ELR that achieves test accuracy higher than the baseline.

However, for *AlleNoise*, the situation is completely different. All the training methods display increasing $\texttt{memorized}_{val}$ values throughout the training, up to 70% (**Fig. 3b**). PRL, SPL and CT+ give lower memorization than the other methods, but this is not reflected in higher accuracy. While these methods correct some of the errors on noisy examples, as measured by $\texttt{correct}_{val}^{\texttt{noisy}}$ (**Fig. 3d**), they display $\texttt{correct}_{val}^{\texttt{clean}}$ lower than other tested approaches (**Fig. 3c**), and thus overall they achieve low accuracy.

These results show that reducing memorization is necessary to create noise-robust classifiers. In this context, it is clear that *AlleNoise*, with its real-world instance-dependent noise distribution, is a challenge for the existing methods.

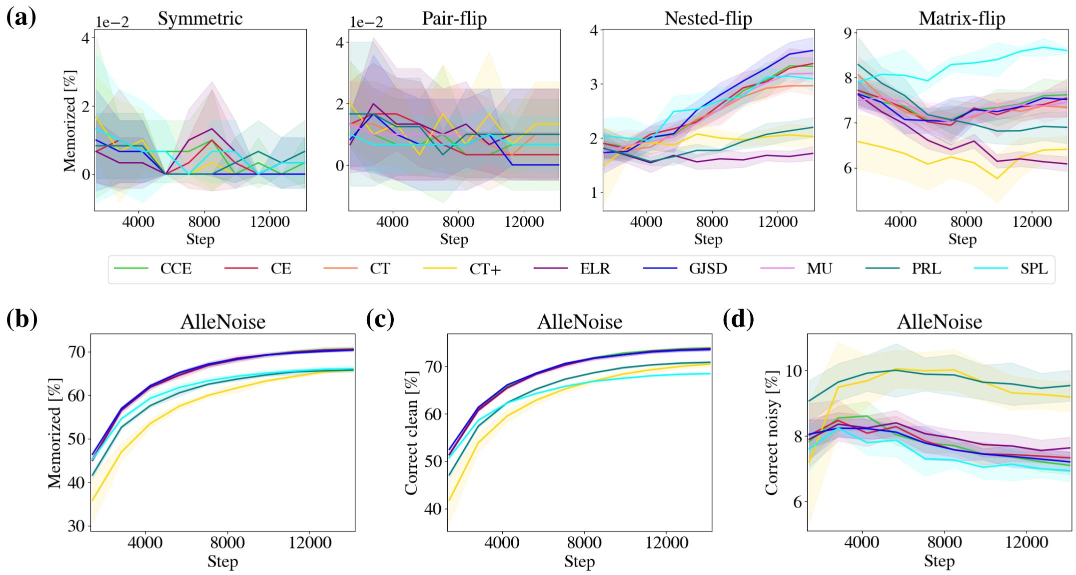

Figure 3: Memorization and correctness metrics as a function of the training step. **(a)** The value of $\texttt{memorized}_{val}$ for synthetic noise types. **(b)** The value of $\texttt{memorized}_{val}$ for *AlleNoise*. **(c)** The value of $\texttt{correct}^{\texttt{clean}}_{val}$ for *AlleNoise*. **(d)** The value of $\texttt{correct}^{\texttt{noisy}}_{val}$ for *AlleNoise*.

5.3 Noise distribution

To get even more insight into why the real-world noise in *AlleNoise* is more challenging than synthetic noise types, we analyzed the class distribution within our dataset. For synthetic noise types, there are very few highly-corrupted categories (**Fig. 4**). On the other hand, for *AlleNoise*, there is a significant number of such categories. The baseline model test accuracy is much lower for these classes than for other, less corrupted, categories. The set of these highly-corrupted classes is heavily populated by the following:

- *Specialized categories* that can be easily mistaken for a more generic category. For example, items belonging to the class *safety shoes* are frequently listed in categories *derby shoes*, *ankle boots* or *other*. In such cases, during the training, the model sees a large number of mislabeled instances of that class and very few correctly labeled ones, which is not enough to learn correct class associations.
- *Archetypal categories* that are considered the most representative examples of a broader parent category. For instance, car tires are most frequently listed in *Summer tires* even when they actually should belong to *All-season tires* or other specialized categories. In this case, the learnt representation of the class is distorted by a huge number of specialized items mislabeled as the archetypal class.

We hypothesize that these categories are the main culprits behind the poor performance of the model.

6 Discussion

Our experiments show that the real-world noise present in *AlleNoise* is a challenging task for existing methods for learning with noisy labels. We hypothesize that the main challenges for these methods stem from two major features of *AlleNoise*: 1) real-world, instance dependent noise distribution, 2) relatively large number of categories with class imbalance and long tail. While previous works have investigated challenges 1) [30] and 2) [35], this paper combines both in a single dataset and evaluation study, while also applying them to text data. We hope that making *AlleNoise* available

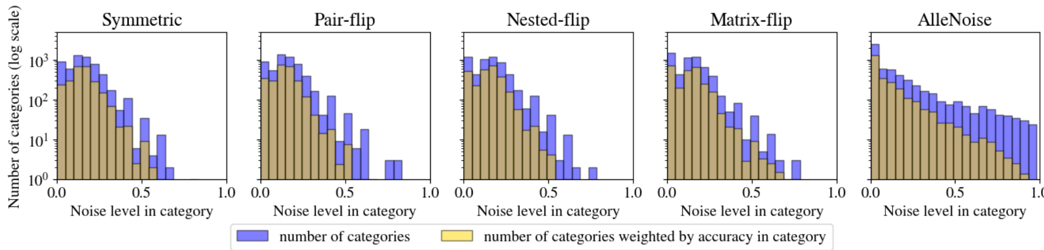

Figure 4: Noise level distribution over target categories (blue bars) shows that *AlleNoise* has a substantial fraction of classes with noise level over 0.5, contrary to synthetic noise. The same distribution multiplied by per-bin macro accuracy (yellow bars) shows that those specialized categories are particularly difficult to predict correctly.

publicly will spark new method development, especially in directions that would address the features of our dataset.

Based on our experiments, we make several interesting observations. The methods that rely on removing examples from within a batch perform noticeably worse than other approaches. We hypothesize that this is due to the large number of classes and the unbalanced distribution of their sizes (especially the long tail of underrepresented categories) in *AlleNoise* - by removing samples, we lose important information that is not recoverable. This is supported by the fact that such noise filtration methods excel on simple benchmarks like CIFAR-10, which all have a completely different class distribution. In order to mitigate the noise in *AlleNoise*, a more sophisticated approach is necessary. A promising direction seems to be the one presented by ELR. While for the real-world noise it did not increase the results above the baseline CE, it was the best algorithm for class-dependent noise types. The outstanding performance of ELR might be attributed to its target smoothing approach. The use of such soft labels may be particularly adequate to extreme classification scenarios where some of the classes are semantically close. Extending this idea to include an instance-dependent component may lead to an algorithm robust to the real-world noise in *AlleNoise*. Furthermore, based on the results of the memorization metric, it is evident that this realistic noise pattern needs to be tackled in a different way than synthetic noise. With the clean labels published as a part of *AlleNoise*, we enable researchers to further explore the issue of memorization in the presence of real-world instance-dependent noise.

7 Conclusions and future work

In this paper, we presented a new dataset for the evaluation of methods for learning with noisy labels. Our dataset, *AlleNoise*, contains a real-world instance-dependent noise distribution, with both clean and noisy labels, provides a large-scale classification problem, and unlike most previously available datasets in the field of learning from noisy labels, features textual data in the form of product names. We performed an evaluation of established noise-mitigation methods, which showed quantitatively that these approaches are not enough to alleviate the noise in our dataset. With *AlleNoise*, we hope to jump-start the development of new robust classifiers that would be able to handle demanding, real-world instance-dependent noise.

The scope of this paper is limited to BERT-based classifiers. As *AlleNoise* includes clean label names in addition to noisy labels, it could be used to benchmark Large Language Models in few-shot or in-context learning scenarios. We leave this as a future research direction.

Acknowledgments and Disclosure of Funding

Acknowledgements

We thank Karol Jędrzejewski for his help in pin-pointing the location of appropriate data records in Allegro data warehouse.

Funding

This work was funded fully by Allegro.com.

Competing interests

We declare no competing interests.

References

1. Frenay, B. & Verleysen, M. Classification in the Presence of Label Noise: A Survey. en. *IEEE Transactions on Neural Networks and Learning Systems* **25,** 845–869. ISSN: 2162-237X, 2162-2388. `http://ieeexplore.ieee.org/document/6685834/` (May 2014).
2. Song, H., Kim, M., Park, D., Shin, Y. & Lee, J.-G. *Learning from Noisy Labels with Deep Neural Networks: A Survey* arXiv:2007.08199 [cs, stat]. Mar. 2022. `http://arxiv.org/abs/2007.08199`.
3. Rolnick, D., Veit, A., Belongie, S. & Shavit, N. *Deep Learning is Robust to Massive Label Noise* en. arXiv:1705.10694 [cs]. Feb. 2018. `http://arxiv.org/abs/1705.10694`.
4. Arpit, D. *et al. A Closer Look at Memorization in Deep Networks* arXiv:1706.05394 [cs, stat]. July 2017. `http://arxiv.org/abs/1706.05394`.
5. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. *Understanding deep learning requires rethinking generalization* arXiv:1611.03530 [cs]. Feb. 2017. `http://arxiv.org/abs/1611.03530`.
6. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning (still) requires rethinking generalization. en. *Communications of the ACM* **64,** 107–115. ISSN: 0001-0782, 1557-7317. `https://dl.acm.org/doi/10.1145/3446776` (Mar. 2021).
7. Jindal, I., Nokleby, M. & Chen, X. *Learning Deep Networks from Noisy Labels with Dropout Regularization* 2017. arXiv: `1705.03419 [cs.CV]`.
8. Patrini, G., Rozza, A., Menon, A., Nock, R. & Qu, L. *Making Deep Neural Networks Robust to Label Noise: a Loss Correction Approach* arXiv:1609.03683 [cs, stat]. Mar. 2017. `http://arxiv.org/abs/1609.03683`.
9. Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine* **29,** 141–142 (2012).
10. Deng, J. *et al. Imagenet: A large-scale hierarchical image database* in *2009 IEEE conference on computer vision and pattern recognition* (2009), 248–255.
11. Krizhevsky, A. *Learning Multiple Layers of Features from Tiny Images* in (2009). `https://api.semanticscholar.org/CorpusID:18268744`.
12. Netzer, Y. *et al. Reading Digits in Natural Images with Unsupervised Feature Learning* in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011* (2011). `http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf`.
13. Northcutt, C. G., Athalye, A. & Mueller, J. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *arXiv:2103.14749 [cs, stat].* arXiv: 2103.14749. `http://arxiv.org/abs/2103.14749` (Nov. 2021).
14. Liu, B. *et al. Noise Learning for Text Classification: A Benchmark* in *Proceedings of the 29th International Conference on Computational Linguistics* (eds Calzolari, N. *et al.*) (International Committee on Computational Linguistics, Gyeongju, Republic of Korea, Oct. 2022), 4557–4567. `https://aclanthology.org/2022.coling-1.402`.

15. Goldberger, J. & Ben-Reuven, E. *Training deep neural-networks using a noise adaptation layer* in *International Conference on Learning Representations* (2017). https://openreview.net/forum?id=H12GRgcxg.

16. Krishna, R. A. *et al. Embracing Error to Enable Rapid Crowdsourcing* in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (ACM, May 2016). http://dx.doi.org/10.1145/2858036.2858115.

17. Nguyen, H. & Khatwani, D. *Robust Product Classification with Instance-Dependent Noise* in *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)* (Association for Computational Linguistics, Dublin, Ireland, May 2022), 171–180. https://aclanthology.org/2022.ecnlp-1.20.

18. Gu, K. *et al.* An Instance-Dependent Simulation Framework for Learning with Label Noise. *arXiv:2107.11413 [cs].* arXiv: 2107.11413. http://arxiv.org/abs/2107.11413 (Oct. 2021).

19. Chen, P., Ye, J., Chen, G., Zhao, J. & Heng, P.-A. *Beyond Class-Conditional Assumption: A Primary Attempt to Combat Instance-Dependent Label Noise* en. Number: arXiv:2012.05458 arXiv:2012.05458 [cs]. Dec. 2020. http://arxiv.org/abs/2012.05458.

20. Xia, X. *et al.* Part-dependent Label Noise: Towards Instance-dependent Label Noise. *arXiv:2006.07836 [cs, stat].* arXiv: 2006.07836. http://arxiv.org/abs/2006.07836 (Dec. 2020).

21. Algan, G. & Ulusoy, İ. Label Noise Types and Their Effects on Deep Learning. *arXiv:2003.10471 [cs].* arXiv: 2003.10471. http://arxiv.org/abs/2003.10471 (Mar. 2020).

22. Berthon, A., Han, B., Niu, G., Liu, T. & Sugiyama, M. Confidence Scores Make Instance-dependent Label-noise Learning Possible. *arXiv:2001.03772 [cs, stat].* arXiv: 2001.03772. http://arxiv.org/abs/2001.03772 (Feb. 2021).

23. Vaswani, A. *et al. Attention is All you Need* in *Advances in Neural Information Processing Systems* (eds Guyon, I. *et al.*) **30** (Curran Associates, Inc., 2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

24. Maas, A. L. *et al. Learning Word Vectors for Sentiment Analysis* in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, Portland, Oregon, USA, June 2011), 142–150. http://www.aclweb.org/anthology/P11-1015.

25. Lin, Y.-C., Das, P., Trotman, A. & Kallumadi, S. *A Dataset and Baselines for e-Commerce Product Categorization* en. in *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval* (ACM, Santa Clara CA USA, Sept. 2019), 213–216. ISBN: 978-1-4503-6881-0. https://dl.acm.org/doi/10.1145/3341981.3344237.

26. Wang, A. *et al. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding* 2019. arXiv: 1804.07461 [cs.CL].

27. Bhatia, K. *et al. The extreme classification repository: Multi-label datasets and code* 2016. http://manikvarma.org/downloads/XC/XMLRepository.html.

28. Li, W., Wang, L., Li, W., Agustsson, E. & Van Gool, L. WebVision Database: Visual Learning and Understanding from Web Data. *arXiv:1708.02862 [cs].* arXiv: 1708.02862. http://arxiv.org/abs/1708.02862 (Aug. 2017).

29. Song, H., Kim, M. & Lee, J.-G. *SELFIE: Refurbishing Unclean Samples for Robust Deep Learning* in *ICML* (2019).

30. Wei, J. *et al.* Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations. *arXiv:2110.12088 [cs, stat].* arXiv: 2110.12088. http://arxiv.org/abs/2110.12088 (Mar. 2022).

31. Xiao, T., Xia, T., Yang, Y., Huang, C. & Wang, X. *Learning from massive noisy labeled data for image classification* in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 2691–2699.

32. Schmarje, L. *et al. Is one annotation enough? A data-centric image classification benchmark for noisy and ambiguous label estimation* 2022. arXiv: 2207.06214 [cs.CV].

33. Jindal, I., Pressel, D., Lester, B. & Nokleby, M. *An Effective Label Noise Model for DNN Text Classification* 2019. arXiv: 1903.07507 [cs.LG].

34. Hedderich, M. A., Zhu, D. & Klakow, D. *Analysing the Noise Model Error for Realistic Noisy Label Data* 2021. arXiv: 2101.09763 [cs.LG].

35. Wu, T. *et al. NoisywikiHow: A Benchmark for Learning with Real-world Noisy Labels in Natural Language Processing* in *Findings of the Association for Computational Linguistics: ACL 2023* (eds Rogers, A., Boyd-Graber, J. & Okazaki, N.) (Association for Computational Linguistics, Toronto, Canada, July 2023), 4856–4873. https://aclanthology.org/2023.findings-acl.299.

36. Hedderich, M. A. *et al. 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)* (eds Webber, B., Cohn, T., He, Y. & Liu, Y.) (Association for Computational Linguistics, Online, Nov. 2020), 2580–2591. https://aclanthology.org/2020.emnlp-main.204.

37. Tan, L., Li, M. Y. & Kok, S. E-Commerce Product Categorization via Machine Translation. *ACM Trans. Manage. Inf. Syst.* **11.** ISSN: 2158-656X. https://doi.org/10.1145/3382189 (July 2020).

38. Zhang, B., Nakatani, T., Walter, S., Misra, A. & Milkovits, E. *Improve machine translation in e-commerce multilingual search with contextual signal from search sessions* in *SIGIR 2023 Workshop on eCommerce* (2023). https://www.amazon.science/publications/improve-machine-translation-in-e-commerce-multilingual-search-with-contextual-signal-from-search-sessions.

39. Ni, J., Li, J. & McAuley, J. *Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects* 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)* (eds Inui, K., Jiang, J., Ng, V. & Wan, X.) (Association for Computational Linguistics, Hong Kong, China, Nov. 2019), 188–197. https://aclanthology.org/D19-1018.

40. Conneau, A. *et al.* Unsupervised Cross-lingual Representation Learning at Scale. *CoRR* **abs/1911.02116.** arXiv: 1911.02116. http://arxiv.org/abs/1911.02116 (2019).

41. Liu, S., Niles-Weed, J., Razavian, N. & Fernandez-Granda, C. Early-Learning Regularization Prevents Memorization of Noisy Labels. *arXiv:2007.00151 [cs, stat].* arXiv: 2007.00151. http://arxiv.org/abs/2007.00151 (Oct. 2020).

42. Kumar, M., Packer, B. & Koller, D. *Self-Paced Learning for Latent Variable Models* in *Advances in Neural Information Processing Systems* **23** (Curran Associates, Inc., 2010). https://papers.nips.cc/paper/2010/hash/e57c6b956a6521b28495f2886ca0977a-Abstract.html.

43. Liu, B., Sun, M., Wang, D., Tan, P.-N. & Zhou, J. Learning Deep Neural Networks under Agnostic Corrupted Supervision. *arXiv:2102.06735 [cs, stat].* arXiv: 2102.06735. http://arxiv.org/abs/2102.06735 (Feb. 2021).

44. Englesson, E. & Azizpour, H. Generalized Jensen-Shannon Divergence Loss for Learning with Noisy Labels. *arXiv:2105.04522 [cs, stat].* arXiv: 2105.04522. http://arxiv.org/abs/2105.04522 (Oct. 2021).

45. Han, B. *et al.* Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. *arXiv:1804.06872 [cs, stat].* arXiv: 1804.06872. http://arxiv.org/abs/1804.06872 (Oct. 2018).

46. Yu, X. *et al.* How does Disagreement Help Generalization against Label Corruption? *arXiv:1901.04215 [cs, stat].* arXiv: 1901.04215. http://arxiv.org/abs/1901.04215 (May 2019).

47. Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. *arXiv:1710.09412 [cs, stat].* arXiv: 1710.09412. http://arxiv.org/abs/1710.09412 (Apr. 2018).

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We describe the dataset in Section 3 and show evaluation results in Sections 5.1, 5.2.
   (b) Did you describe the limitations of your work? [Yes] We discuss the limitations of our dataset in Section 3. However, to the best of our knowledge, the data in our dataset realistically reflects the actual distribution of products within an established e-commerce platform, used by over 20M daily active users.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Not applicable. Our dataset addresses an important problem in machine learning theory i.e. robustness to label noise, which is a significant research area in supervised learning. This does not have any societal impact per se.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We have carefully inspected the guidelines and made sure to conform to them.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] Not applicable
   (b) Did you include complete proofs of all theoretical results? [N/A] Not applicable
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We provide all the code and instructions in the supplementary GitHub repository at https://github.com/allegro/AlleNoise
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.3 and Section 4.5.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We have used 5 split cross-validation to estimate the variance in all experiments. See Section 4.3.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We used Google Cloud Platform virtual machines with NVIDIA A100 GPUs. See Section 4.3.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] The only existing asset used in the study is the XLMRoBERTa backbone, referenced in Section 4.3
   (b) Did you mention the license of the assets? [Yes] We specify the licence of our dataset in the supplementary data sheet.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include the data sheet of our dataset, with all relevant information and URLs as a supplementary material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We have not collected any data that would require user consent. We have required legal approval from Allegro to publish our data, which is stated in the supplementary data sheet.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] We described data post-processing in Section 3.3, i.e. filtering out potentially offensive product categories.
5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] Not applicable. The data in our dataset comes from pre-existing internal logs of Allegro.com.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] Not applicable. The data in our dataset comes from pre-existing internal logs of Allegro.com.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] While the data in our dataset comes from pre-existing internal logs of Allegro.com, we do state in the supplementary data sheet the guaranteed wage that human domain experts who originally verified the data were compensated with.