Datasheet for

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

1. Introduction

This datasheet describes *AlleNoise*, a benchmark dataset for large-scale multi-class text classification with real-world label noise. It consists of e-commerce product titles from Allegro.com with corresponding category labels. 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 and a meaningful, hierarchical taxonomy of categories.

2. Dataset Statistics

- **Number of data points:** 502,310
- **Number of classes:** 5,692 (unique product categories)
- **Data format:** Tab-separated CSV (two files; details in Section 4)

3. Task Description

The task is to predict the correct category label for a given product title, considering the presence of class imbalance and label noise in the original categories. 15% of the products were listed in incorrect categories, introducing noise into the dataset.

4. Dataset Files

- **full_dataset.csv:** The products dataset with four columns:

    - `offer_id`: ID of the offer on Allegro.com where the product was listed (string)
    - `text`: Product name (string, in English)
    - `clean_category_id`: True category label where the product should be listed according to domain experts (integer)
    - `noisy_category_id`: Noisy category label where the product was initially listed (integer, might be incorrect)
- **category_mapping.csv:** Category labels and their corresponding paths in the hierarchical taxonomy of assortment categories on Allegro.com (relevant for exploring noise semantics).:
    - `category_label`: Category label used in full_dataset.csv (integer)
    - `category_name`: Category path in the hierarchical taxonomy (string)

The dataset is available for download at https://github.com/allegro/AlleNoise

5. Metadata

We provide metadata in the ML Croissant format. It can be accessed at https://github.com/allegro/AlleNoise/blob/main/allenoise/metadata.json

6. Dataset Creation

- **Real-world noise:** We collected 75,348 mislabeled products from two sources:
    - customer complaints about a product being listed in a wrong category - such requests usually suggest the true category label

    - assortment clean-up by a domain expert - products listed in the wrong category were manually moved to the correct category.

- **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.

- **Post-processing:**
    - We automatically translated all 500k product titles from Polish to English.
    - Categories related to sexually explicit content were removed from the dataset altogether.
    - Categories with less than 5 products were removed from the dataset to allow for five-fold cross-validation in our experiments.

7. Usage Guidelines and Recommendations

- The dataset is intended for training and evaluating machine learning models for multi-class text classification with label noise.

- We recommend splitting the data into training, validation, and test sets while maintaining the class distribution and noise ratio.

- For model development, we recommend monitoring:

  - accuracy on the whole validation dataset
  - accuracy only on validation instances with noisy labels
  - the memorization metric, *i. e.* the ratio of predictions that match the noisy label to the number of noisy instances in the validation set
- Test accuracy should be used only for final model evaluation.

8. Baseline Performance

- Test accuracy of the baseline model (XLM-RoBERTa + linear layer) with cross-entropy loss is:
  - 74.85 ± 0.15 on clean labels
  - 63.71 ± 0.11 on noisy labels (with 15% real-world label noise)
- Test accuracies of selected methods for learning with noisy labels are reported in the accompanying paper [1].
- The code needed to reproduce the published results is available at: https://github.com/allegro/AlleNoise.

9. Additional Notes

- The 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.
- The dataset was submitted to the NeurIPS 2024 "Benchmarks and Datasets" Track.

- The DOI record for the dataset is https://zenodo.org/doi/10.5281/zenodo.11486108

10. License

The dataset is licensed under CC BY-NC-ND. The code is licensed under the MIT license.

11. Hosting and maintenance plan.

The dataset and the code are and will be available at: https://github.com/allegro/AlleNoise.

12. Author statement

We, the authors, bear all responsibility to withdraw our paper and data in case of violation of licensing or copyright of the data presented herein. Publication of the dataset has been approved by the Legal Department of Allegro.pl sp. z o. o.

13. Contact

For any questions or feedback regarding the AlleNoise dataset, please contact Alicja Rączkowska at alicja.raczkowska@allegro.com or Machine Learning Research at Allegro at mlr@allegro.com.

14. References

[1] Rączkowska, A., Osowska-Kurczab, A., Szczerbiński, J., Jasinska-Kobus, K., Nazarko, K., AlleNoise - large-scale text classification benchmark dataset with real-world label noise, 2024