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 internal domain experts - products listed in the wrong category were manually moved to the correct category. The domain experts were Allegro employees, employed in Poland. The minimum hourly wage in Poland is regulated by Polish law, see the ordinance Dz.U. 2021 poz. 1690 (legal basis: Dz.U. 2002 nr 200 poz. 1679).

- **Clean data sampling:** The 74,094 mislabeled products were complemented with 428,216 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. Ethical and Societal Impact of the Dataset

The realistic benchmark dataset for learning from noisy labels may jump-start jump-start the development of new robust classifiers that would be able to handle demanding, real-world instance-dependent noise, reducing errors in practical applications of text classifiers.

9. Limitations of the Dataset

- **Selection of Categories:** Allegro is a general marketplace that represents a wide spectrum of products from various categories and shopping intents. Our dataset comprises over 5,000 categories sampled from nearly 23,000 overall, following the distribution of the Allegro catalog. We undersampled the entire catalog to maintain a manageable dataset size and to control noise levels at around 15%.
- **Allegro as a Polish Marketplace**: Since the data originates from a Polish marketplace, the selection of products reflects items typical to this region. The diversity of products catering to minority groups might be limited due to the

popularity-based filtering used in the dataset. Additionally, due to EU regulations, the selection of products may not be representative of other marketplaces, such as those originating from the Americas, Asia, or Africa.

- **E-commerce Domain**: Our dataset, composed exclusively of e-commerce product names, may not be easily transferable to the broader NLP domain due to its specialized nature. Product titles often include domain-specific jargon, abbreviations, named entities, numbers, codes, and concise text that differs significantly from the more diverse and unstructured language found in general NLP tasks, such as web pages, articles, or conversations. This domain-specific focus can limit the generalizability of models trained on this data to other NLP applications.
- **Machine Translated Content**: Product titles have been translated using an in-house Neural Machine Translation (NMT) service, maintained by a team of over 30 machine learning specialists, software engineers, and language quality experts, following recent advancements in NMT. However, machine translation systems, often trained on general language corpora, may struggle with the domain-specific jargon, abbreviations, and structured product descriptions common in e-commerce, leading to inaccurate or misleading translations. Additionally, brand names, model numbers, and industry-specific terms may lack direct equivalents in other languages, resulting in translation errors that can compromise the clarity and reliability of the content. We mitigate these issues through model fine-tuning on in-house data, the use of translation glossaries, input data exceptions, and no-translate entity detection, but the model's accuracy is not perfect. More information on the quantitative impact of machine translations can be found in Appendix G of the accompanying research paper [1].
- **Malicious Content**: The product database is maintained daily by expert category managers to detect any malicious behavior on the platform, such as illegitimate,

disrespectful, or offensive products, personally identifiable information, derogatory language, etc. To the best of our knowledge, the dataset should be free from malicious content; however, we did not conduct extensive annotation in this regard.

- **Intended Use Case**: The intended use case of the dataset is to develop robust text classifiers for benchmarking algorithms that learn from noisy labels, which was our primary focus during its creation. We discourage any unintended usage of the AlleNoise dataset.
- **Competing Datasets**: To date, several similar benchmark datasets have been published for e-commerce applications of ML algorithms, such as Amazon [5], Rakuten [4], Skroutz [2,3], and Shopmania [2,3]. Our dataset competes in size (500,000 instances) and content (5,000 categories). A key difference is the known noise level available in the AlleNoise dataset.

10. 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.

11. 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

12. License

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

13. Hosting and maintenance plan.

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

14. 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.

15. 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.

16. 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

[2] Leonidas Akritidis, Athanasios Fevgas, and Panayiotis Bozanis. 2018. Effective Products Categorization with Importance Scores and Morphological Analysis of the Titles. In 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI), pages 213–220.

[3] Leonidas Akritidis, Athanasios Fevgas, Panayiotis Bozanis, and Christos Makris. 2020. A self-verifying clustering approach to unsupervised matching of product titles. Artificial Intelligence Review, pages 1–44.

[4] Yiu-Chang Lin, Pradipto Das, Andrew Trotman, and Surya Kallumadi. 2019. A Dataset and Baselines for e-Commerce Product Categorization. In Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '19). Association for Computing Machinery, New York, NY, USA, 213–216. https://doi.org/10.1145/3341981.3344237

[5] Bridging Language and Items for Retrieval and Recommendation, Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, Julian McAuley