ActiveAED: A Human in the Loop Improves Annotation Error Detection

Leon Weber[▲] and **Barbara Plank**[▲♢]
[▲]Center for Information and Language Processing (CIS), LMU Munich, Germany
[♢]Munich Center for Machine Learning (MCML), Munich, Germany
{leonweber, bplank}@cis.lmu.de

Abstract

Manually annotated datasets are crucial for training and evaluating Natural Language Processing models. However, recent work has discovered that even widely-used benchmark datasets contain a substantial number of erroneous annotations. This problem has been addressed with Annotation Error Detection (AED) models, which can flag such errors for human re-annotation. However, even though many of these AED methods assume a final curation step in which a human annotator decides whether the annotation is erroneous, they have been developed as static models without any human-in-the-loop component. In this work, we propose ActiveAED, an AED method that can detect errors more accurately by repeatedly querying a human for error corrections in its prediction loop. We evaluate ActiveAED on eight datasets spanning five different tasks and find that it leads to improvements over the state of the art on seven of them, with gains of up to six percentage points in average precision.

1 Introduction

Correct labels are crucial for model training and evaluation. Wrongly labelled instances in the training data hamper model performance (Larson et al., 2020; Vlachos, 2006), whereas errors in the test data can lead to wrong estimates of model performance (Alt et al., 2020; Larson et al., 2020; Reiss et al., 2020). This is a problem in practice, as even widely used benchmark datasets can contain a non-negligible number of erroneous annotations (Alt et al., 2020; Northcutt et al., 2021; Reiss et al., 2020). Researchers have developed a multitude of annotation error detection (AED) methods to detect such labelling errors as recently surveyed by Klie et al. (2022). After detection, there are multiple ways to deal with the found annotation errors. When it comes to training data, a reasonable strategy is to simply remove the instances flagged by an AED model (Huang et al., 2019). For evaluation data, however, this is not viable, because in many cases this would remove a significant fraction of hard but correctly labelled instances in addition to the errors (Swayamdipta et al., 2020), which would lead to an overestimation of model performance. Instead, researchers resorted to manual correction of the labels flagged by the AED method (Alt et al., 2020; Reiss et al., 2020; Northcutt et al., 2021; Larson et al., 2020). Strikingly, even though this manual correction requires human input, the typical workflow is to first apply the AED method once and afterwards correct the flagged errors, without using the human feedback in the AED step.

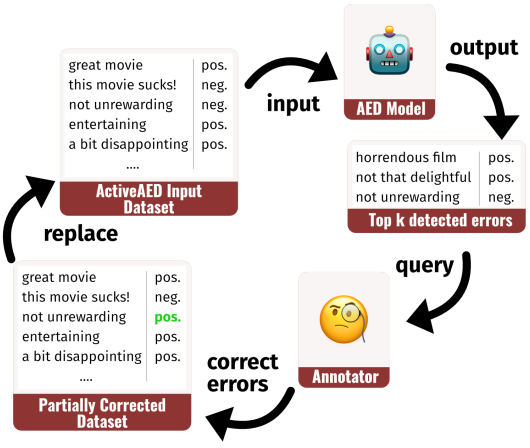

Figure 1: Prediction loop of ActiveAED

We hypothesize that connecting the human input and the AED prediction in a human-in-the-loop setup could increase the accuracy of the AED method without increasing the total amount of human intervention. To support this hypothesis, we propose ActiveAED, an AED method which includes human feedback in the annotation loop; see Figure 1 for an illustration. We base ActiveAED on the Area-under-the-Margin metric (AUM) (Pleiss et al., 2020), which was recently proposed to detect annotation errors in computer vision datasets. As an additional contribution, we

Findings of the Association for Computational Linguistics: ACL 2023, pages 8834–8845
July 9-14, 2023

propose a novel ensembling scheme to improve AUM's performance. In experiments on eight datasets spanning five different tasks, we show that ActiveAED improves over three baselines that performed well in a recent evaluation (Klie et al., 2022). On seven datasets, we observe improvements, with gains of up to six percentage points (pp) in average precision. Our ablation study shows that both the human-in-the-loop component and the ensembling scheme contribute to the improvements. We make code and data available under https://github.com/mainlp/ActiveAED.

2 Related Work

AED for Natural Language Processing (NLP) datasets has a long tradition which has recently been comprehensively evaluated and surveyed by the seminal work of Klie et al. (2022). We base our evaluation setup on theirs. Existing AED methods can be divided into six different categories (Klie et al., 2022): variation-based (Dickinson and Meurers, 2003; Larson et al., 2020), model-based (Amiri et al., 2018; Yaghoub-Zadeh-Fard et al., 2019; Chong et al., 2022), training-dynamics-based (Swayamdipta et al., 2020; Pleiss et al., 2020; Siddiqui et al., 2022), vector-space-proximity-based (Larson et al., 2019; Grivas et al., 2020), ensembling-based (Alt et al., 2020; Varshney et al., 2022) and rule-based (Květoň and Oliva, 2002). To the best of our knowledge, none of these AED methods has been developed or evaluated with a human in the loop, except for Vlachos (2006) who uses AED as part a larger framework for constructing a silver-standard dataset. Accordingly, they do not compare the performance of the AED component to competing approaches and they consider only a single dataset and task.

Additionally, one can distinguish between flaggers and scorers for AED (Klie et al., 2022). Flaggers output hard decisions of whether an instance contains an error, whereas scorers assign to each instance a score reflecting the likelihood of being an error. In this work, we focus on scoring methods, because ActiveAED requires error scores to rank the instances.

3 Active Annotation Error Detection

We propose ActiveAED, an AED method which uses the error corrections issued by an annotator in its prediction loop. The basic procedure of ActiveAED is this: In the first step, it uses a ranking-based AED method to find the k most likely annotation errors across the dataset. In the second step, the presumed annotation errors are forwarded to an annotator who checks them and corrects the labels if necessary. After this, the dataset is updated with the corrections issued by the annotator and the procedure continues with the first step. This loop continues until a stopping condition is met, e.g. that the fraction of errors in the batch drops to a user-defined threshold. See Figure 1 for an illustration of the process.

We consider a scenario where an annotator wants to correct annotation errors in a dataset with a given annotation budget of n instances. There are two options of how to apply an annotation error detection (AED) method to support this. The first is the state-of-the-art and the second one is our proposed approach: (1) Run the AED method once on the dataset to retrieve a list of instances ranked by their probability of containing an annotation error. Then, spend the annotation budget by correcting the top-n instances. (2) Run the AED method and spend some of the annotation budget by correcting the top-k instances with $k \ll n$. Then, run the AED method again on the now partially corrected dataset and repeat until the annotation budget is exhausted. Note, both approaches involve ranking instances based on their probability of containing annotation errors, and selection of a subset of instances for annotation based on this ranking. As a result, the outputs of both approaches can be fairly compared, because they use the same annotation budget and the same ranking-based score.

More formally, we assume a dataset with inputs X, (potentially erroneous) labels y, and true labels y^* which are initially unknown to us. After training the model for E epochs, we use (negative) AUM to assign error scores:

$$s_i = \frac{1}{E} \sum_{e=1}^{E} \max_{y' \neq y_i} p_{\theta_e}(y'|x_i) - p_{\theta_e}(y_i|x_i), \quad (1)$$

where $p_{\theta_e}(y_i|x_i)$ is the probability of the label assigned to x_i as estimated by θ_e and $\max_{y' \neq y_i} p_{\theta_e}(y'|x_i)$ the probability of the highest scoring label that is not the assigned one. Intuitively, correctly labelled instances on average obtain smaller (negative) AUM scores (Eq. 1) than incorrect ones, because the model will confidently predict their correct label earlier in the training. We chose AUM, because it performed well in preliminary experiments on SI-Flights (Larson et al., 2020)

and ATIS (Hemphill et al., 1990). Note, that this formulation differs from the original one in Pleiss et al. (2020) that uses raw logits instead of probabilities. We chose to use probabilities because this performed better in our experiments (see Table 1).

We extend AUM with a novel ensembling scheme based on training dynamics. For this, we train a model for E epochs in a C-fold cross-validation setup. For each fold $c \in \{1, ..., C\}$ and epoch $e \in \{1, ..., E\}$, we obtain a model $\theta_{c,e}$. We use the models of one fold c to assign an error score $s_{c,i}$ to each instance with AUM (Eq. 1). For each fold, we calculate the AUM score both on the train and on the test portion of the fold, which yields $C - 1$ training-based scores and one test-based score for each instance. For each instance, we first average the training-based scores and then compute the mean of this average and the test-based score, which results in the final score s_i:

$$s_i^{train} = \frac{1}{E-1} \sum_{c \in train_i} s_{c,i} \quad (2)$$

$$s_i = \frac{1}{2}(s_i^{train} + s_i^{test}), \quad (3)$$

where $train_i$ is the set of $C - 1$ folds in which instance i appears in the training portion. Then, we rank all uncorrected instances by s_i and route the k highest scoring ones to the annotator, who manually corrects their label by setting $y_i := y_i^*$. Finally, the procedure continues with the partially corrected dataset until a stopping condition is met. There are two kinds of motivation for the proposed ensembling scheme: s^{train} should improve the calibration of the model (Ovadia et al., 2019), which Klie et al. (2022) show to be helpful for AED. s^{test} derives from the observation that model-based AED methods benefit from computing statistics over unseen data (Klie et al., 2022).

4 Evaluation Protocol

4.1 Datasets & Evaluation Setting

We evaluate ActiveAED on eight datasets following the choice of datasets used by Klie et al. (2022):[1]

- The intent classification part of **ATIS** (Hemphill et al., 1990), for which we randomly perturb labels.
- The sentiment analysis dataset **IMDb** (Maas et al., 2011), for which Northcutt et al. (2021) provide semi-automatically detected annotation errors.
- The sentiment analysis dataset **SST** (Socher et al., 2013) with randomly perturbed labels.
- The UPOS annotations[2] from the Georgetown University Multilayer Corpus (**GUM**; Zeldes (2017)) with randomly perturbed labels.
- The **CoNLL-2003** Named Entity Recognition data (Tjong Kim Sang and De Meulder, 2003), for which Reiss et al. (2020) provide a version with corrected annotations.
- The slot three filling datasets **SI Companies**, **SI Flights**, and **SI Forex** (Larson et al., 2020) that contain manually corrected slot labels.

We provide Hugging Face datasets implementations and detailed statistics for all datasets; see Appendix A. Our evaluation setup for the sequence labelling datasets (GUM, CoNLL-2003, SI Companies, SI Flights, and SI Forex) differs from that proposed by Klie et al. (2022). We opt for a sequence-level setting because it is closer to our envisioned application scenario, as it makes more sense for an annotator to correct the entire sequence of annotations instead of a single one at a time. Specifically, we define errors on the sequence level, i.e. if at least one token annotation differs from the gold annotation, the sequence is treated as an error both during ActiveAED prediction and for evaluation. During prediction, ActiveAED aggregates token-level error scores by calculating the maximum over all tokens in the sequence. For the other parts of the evaluation setup we follow Klie et al. (2022).[3]

In all datasets in which we perturbed labels, we resample the label uniformly for 5% of all annotations. We use average precision (AP) as our evaluation metric, which we compute with scikit-learn v1.1.3 (Pedregosa et al., 2011). To be consistent with ActiveAED's application scenario, we cannot

[1] From this list, we exclude Plank et al. (2014) because it contains only annotation ambiguities and not corrected errors which are required for our evaluation setting.

[2] https://github.com/UniversalDependencies/UD_English-GUM

[3] Note, that our results are not comparable with the numbers for the state-of-the-art reported by Klie et al. (2022), because of the different treatment of sequence-labelling datasets. Additionally, for ATIS and SST the choice of randomly perturbed labels differs (but the fraction is the same) and for IMDb the dataset statistics reported by Klie et al. (2022) are different from those of the original dataset (Northcutt et al., 2021), which we use.

	ATIS	SI-Flights	IMDb	SST	GUM	CONLL-2003	SI-Companies	SI-Forex
CU	91.7±1.4	80.9±0.5	31.6±1.3	42.7±1.0	98.8±0.1	25.2±0.6	96.1±0.2	84.2±2.0
DM	97.2±0.2	79.2±2.4	30.1±3.0	47.1±1.0	99.3±0.1	30.2±0.7	97.5±0.2	80.6±0.9
AUM (p)	98.0±0.1	78.9±2.3	30.1±3.0	47.1±1.0	99.0±0.1	30.2±0.7	97.3±0.3	81.1±0.9
AUM (l)	97.3±0.4	72.6±0.3	27.5±2.5	39.6±1.3	**99.5±0.1**	29.3±0.2	97.2±0.2	66.6±1.5
ActiveAED	**98.6±0.1**	**86.6±0.5**	**36.6±0.1**	**53.0±0.2**	98.5±0.0	**33.3±0.2**	**99.3±0.0**	**89.7±0.6**
w/o active	98.7±0.1	80.3±0.6	36.0±0.4	52.9±0.4	98.4±0.0	31.7±0.4	97.9±0.1	85.5±0.6

Table 1: Evaluation results. All scores are mean and standard deviation of AP for AED in percent over three random seeds. The best score per dataset (without ablation) is in bold. We used ATIS and SI-Flights as development data. The last row is ActiveAED without the human-in-the-loop component. AUM (l) is the original version of AUM proposed by Pleiss et al. (2020), whereas AUM (p) is our variant in which we aggregate probabilities instead of raw logits.

use the standard train/dev/test split practice from supervised learning, because we will not have access to any known errors which we could use for development when we apply ActiveAED to a new dataset. Thus, we select the two datasets ATIS and SI-Flights as development datasets on which we devise our method, and reserve the remaining datasets for the final evaluation. We report the average and standard deviation across three random seeds. We follow the standard practice in active learning research and simulate the annotator by using gold-standard corrections (Settles, 2012; Zhang et al., 2022). Note, that here, we simulate a single annotator without accounting for inter- and intra-annotator variation (Jiang and de Marneffe, 2022; Plank, 2022). We set $k = 50$ (an ablation for k can be found in Section 5), because this is small enough so that an annotator can handle it in a single annotation session but large enough that gains can be observed after a single iteration on SI Flights. We stop the prediction loop after 40 iterations or when the whole dataset was annotated. We perform 10-fold cross validation in all experiments. We describe the remaining hyperparameters in Appendix B.

4.2 Baselines

As baselines, we choose the top-performing scorer methods recommended by Klie et al. (2022):

- (Negative) Area-under-the-margin (**AUM**) (Pleiss et al., 2020): $s_i^{AUM} = \frac{1}{E}\sum_{e=1}^{E} \max_{y' \neq y_i} p_{\theta_e}(y'|x_i) - p_{\theta_{ce}}(y_i|x_i)$
- (Negative) Data Map Confidence (**DM**) (Swayamdipta et al., 2020): $s_i^{DM} = -\frac{1}{E}\sum_{e=1}^{E} p_{\theta^{(e)}}(y_i|x_i)$
- Classification Uncertainty (**CU**) (Klie et al., 2022): $s_i^{CU} = -p_{\theta^*}(y_i|x_i)$,

where AUM and DM are both computed over a single training run and CU is computed with cross-validation over the test portions using the model θ^* achieving the lowest test loss for the given fold.

5 Results

The results of our evaluation can be found in Table 1. ActiveAED outperforms the three baselines on seven of the eight datasets, with gains ranging from 0.6 to 6 pp AP. We observe a large variance of the AP scores across different datasets, which is in concordance with the findings of Klie et al. (2022). We suspect that the relatively low scores on IMDb and CoNLL-2003 are because the errors were manually annotated after automatic filtering and thus are limited by the recall of the filtering method. We disentangle the contribution of our proposed ensembling strategy from that of the human-in-the-loop component by ablating the human-in-the-loop (last row in Table 1). We find that on four of the eight datasets, the ensembling alone improves results, whereas on SI Companies, SI Flights, and SI Forex, the main driver for improvements is the human-in-the-loop component. Generally, the human-in-the-loop component improves over the non-active variant on seven out of eight datasets.

A natural question that arises is whether the human-in-the-loop procedure of ActiveAED can also improve AED methods other than our modified version of AUM. To investigate this, we evaluate unmodified versions of (negative) AUM and DM on SI Flights and ATIS with our human-in-the-loop setup. We find that, for SI Flights, AUM/DM improves by 7.4/6.9 pp AP, whereas for ATIS, DM improves by 0.8 pp and AUM's result diminishes by 0.2 pp. This suggests that a human in the loop might not be helpful for all combinations of datasets and methods, but that it has the potential

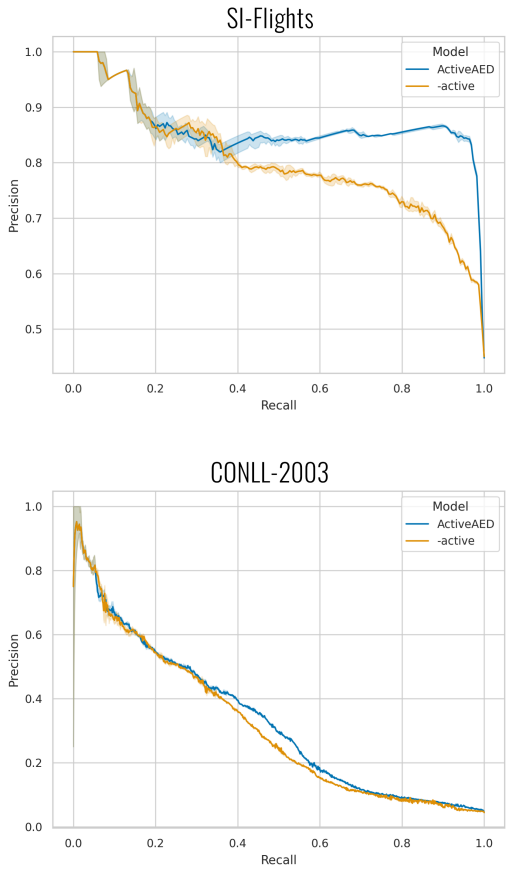

Figure 2: Comparison of the precision-recall curves of ActiveAED and its non-active ablation. The gains of ActiveAED are made in the mid-to-high recall regime for both datasets. Curves are mean and error bars are standard deviation across three random seeds.

to significantly improve results for other methods than ActiveAED.

It is instructive to compare the precision-recall curves of ActiveAED to that of its non-active variant. The graphs for datasets SI Flights and CoNLL-2003 can be found in Figure 2. On both datasets, the precision gains are present in the mid-to-high recall regime (> 0.4), which intuitively makes sense, because ActiveAED requires a few rounds of human annotation to produce different outputs than its non-active variant. This suggests that one could increase the efficiency of ActiveAED by starting with a more lightweight AED method, e.g. one that does not require cross validation or ensembling and only later switch to the more compute-intensive ensembling of ActiveAED. We leave the investigation of this option for future work. We describe the ablation study of our proposed ensembling scheme and for different choices of k in Appendix C. Here, we find that test ensembling is crucial, that train ensembling sometimes improves results and that increasing k for the small SI-Flights dataset harms results. We provide example outputs of ActiveAED in Appendix E.

6 Conclusion

We have proposed ActiveAED, an AED method that includes human feedback in its prediction loop. While the proposed approach could be used with every ranking-based AED method, we base ActiveAED on the recently proposed AUM score, which we augment with a novel ensembling scheme based on training dynamics. We evaluate ActiveAED on eight datasets spanning five different tasks and find that it improves results on seven of them, with gains of up to six pp AP. In future work, we plan on extending ActiveAED to generative models and structured prediction tasks. Additionally, we want to use ActiveAED to clean benchmark datasets. We also plan to investigate the reasons for the observed performance gains of ActiveAED, for instance by exploring the role of model capacity and dataset characteristics (Ethayarajh et al., 2022). Finally, we would like to study the interplay between ActiveAED and human label variation (Jiang and de Marneffe, 2022; Plank, 2022).

Limitations

A major limitation of ActiveAED is that it is significantly more compute-intensive than other scoring-based AED methods such as AUM or DM. This is inherent to the proposed method because the ensemble requires training of multiple models and, after receiving human feedback, the full ensemble has to be re-trained. Also, the ensembling of ActiveAED requires more training runs than training-dynamics-based AED methods. However, most model-based methods require a cross-validation scheme (Klie et al., 2022). The ensembling component of ActiveAED is more data-efficient than these approaches, because it makes use of the training dynamics captured during cross-validation instead of discarding them. A second limitation of this work is that while we chose baselines that performed strongly in Klie et al. (2022), they represent only a fraction of the scoring-based AED methods described in the literature. Finally, our evaluation is limited to a single language model and it would be interesting to investigate how ActiveAED interacts with larger language models than DistilRoBERTa.

Ethics Statement

Datasets with fewer annotation errors can improve model training and evaluation. While this generally seems desirable, it is subject to the same dual-use concerns as the NLP models that are improved with AED methods. Additionally, using ActiveAED instead of AUM or DM can make the AED results more accurate, but that comes at the expense of a higher runtime. This, in turn, leads to increased energy consumption and, depending on the source of the energy, more CO_2 released (Strubell et al., 2019), which is highly problematic in the face of the climate crisis.

Acknowledgements

We thank the reviewers for their constructive feedback which helped to improve the paper. Many thanks to the members of MaiNLP and NLPNorth for their comments on the paper. This research is in parts supported by European Research Council (ERC) grant agreement No. 101043235.

References

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED Revisited: A Thorough Evaluation of the TACRED Relation Extraction Task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, Online. Association for Computational Linguistics.

Hadi Amiri, Timothy Miller, and Guergana Savova. 2018. Spotting Spurious Data with Neural Networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2006–2016, New Orleans, Louisiana. 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Markus Dickinson and W. Detmar Meurers. 2003. Detecting Errors in Part-of-Speech Annotation. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary. Association for Computational Linguistics.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding Dataset Difficulty with $\mathcal{V}$-Usable Information.

Andreas Grivas, Beatrice Alex, Claire Grover, Richard Tobin, and William Whiteley. 2020. Not a cute stroke: Analysis of Rule- and Neural Network-based Information Extraction Systems for Brain Radiology Reports. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 24–37, Online. Association for Computational Linguistics.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS Spoken Language Systems Pilot Corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. 2019. O2U-Net: A Simple Noisy Label Detection Approach for Deep Neural Networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3325–3333.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating Reasons for Disagreement in Natural Language Inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2022. Annotation Error Detection: Analyzing the Past and Present for a More Coherent Future. *Computational Linguistics*, pages 1–42.

Pavel Květoň and Karel Oliva. 2002. (Semi-)Automatic Detection of Errors in PoS-Tagged Corpora. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Stefan Larson, Adrian Cheung, Anish Mahendran, Kevin Leach, and Jonathan K. Kummerfeld. 2020. Inconsistencies in Crowdsourced Slot-Filling Annotations: A Typology and Identification Methods. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5035–5046, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Stefan Larson, Anish Mahendran, Andrew Lee, Jonathan K. Kummerfeld, Parker Hill, Michael A. Laurenzano, Johann Hauswald, Lingjia Tang, and Jason Mars. 2019. Outlier Detection for Improved Data Quality and Diversity in Dialog Systems. 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 517–527, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Alberto Moro and Laura Lonza. 2018. Electricity carbon intensity in European Member States: Impacts on GHG emissions of electric vehicles. *Transportation Research Part D: Transport and Environment*, 64:5–14.

Curtis Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model' s uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. 2020. Identifying Mislabeled Data using the Area Under the Margin Ranking. In *Advances in Neural Information Processing Systems*, volume 33, pages 17044–17056. Curran Associates, Inc.

Frederick Reiss, Hong Xu, Bryan Cutler, Karthik Muthuraman, and Zachary Eichenberger. 2020. Identifying Incorrect Labels in the CoNLL-2003 Corpus. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 215–226, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter.

Burr Settles. 2012. Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114.

Shoaib Ahmed Siddiqui, Nitarshan Rajkumar, Tegan Maharaj, David Krueger, and Sara Hooker. 2022. Metadata Archaeology: Unearthing Data Subsets by Leveraging Training Dynamics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: A web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022. ILDAE: Instance-Level Difficulty Analysis of Evaluation Data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3412–3425, Dublin, Ireland. Association for Computational Linguistics.

Andreas Vlachos. 2006. Active Annotation. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)*.

Mohammad-Ali Yaghoub-Zadeh-Fard, Boualem Benatallah, Moshe Chai Barukh, and Shayan Zamanirad. 2019. A Study of Incorrect Paraphrases in Crowd-sourced User Utterances. 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 295–306, Minneapolis, Minnesota. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Shujian Zhang, Chengyue Gong, Xingchao Liu, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. 2022. ALLSH: Active Learning Guided by Local Sensitivity and Hardness.

A Datasets

Table 2 lists statistics for all datasets that we used in this work, together with links to the HuggingFace datasets implementations we provide.

B Hyperparameters

As base model, we choose the 82M parameter model DistilRoBERTa-base[4] (Sanh et al., 2020), which is licensed under apache-2.0. In all experiments, we perform 10-fold cross validation. We manually optimize the hyperparameters of ActiveAED on ATIS and SI Flights, resulting in a learning rate of 5e-5 and a batch size of 64. We adapt the number of epochs to the size of the dataset: for the SI datasets, we set it to 40, for ATIS to 20, for GUM, CoNLL and SST to 10, and for IMDb to 5 and use Adam (Kingma and Ba, 2015). We set the number of instances that the annotator corrects in a single pass k to 50 for all datasets because this is small enough so that an annotator can handle it in a single annotation session but large enough that gains could be observed after a single iteration on SI Flights.

C Further Ablation Studies

Table 3 gives results for our full ablation study. We find that for ATIS, where ensembling was the main driver of improved results, ablating both train and test ensembling leads to worse results. For SI-Flights, the variant without test ensembling leads to worse results, whereas omitting train ensembling improves results. We hypothesized that, for small datasets, increasing k would lead to worse results. Our results confirm this. Setting $k = 100$ leaves results almost unchanged, whereas $k = 200$ leads to a dramatic drop of 3.9 pp AP on SI-Flights, without affecting performance on the much larger ATIS.

D Compute Resources for Experiments

We estimate the total computational cost of our experiments including development of the method to be around 1000 GPU hours on an 80GB A100. As per the ML CO_2 Impact tool (Lacoste et al., 2019)[5] and an average carbon intensity of electricity for Germany of $0.485 \frac{\text{kg}}{\text{kWh}} CO_2$ (Moro and Lonza, 2018) this amounts to roughly 121 kg CO_2 emitted.

E Example Outputs

Example outputs for IMDb and CONLL-2003 can be found in Figure Appendix 3. We show the five instances with the highest error scores assigned by ActiveAED. All instances contain an annotation error.

[4] https://huggingface.co/distilroberta-base

[5] https://mlco2.github.io/impact/

Review	Original Label
SPOILERS AHEAD It is really unfortunate that a movie so well produced turns out to be such a disappointment. [...]	Positive
Lois Weber's film "Hypocrites" was and still kind of is a very bold and daring film. I enjoyed it and was very impressed by the filming and story of it. [...]	Negative
I really liked this quirky movie. The characters are not the bland beautiful people that show up in so many movies and on TV. It has a realistic edge, with a captivating story line. The main title sequence alone makes this movie fun to watch.	Negative
I went to see this 3 nights ago here in Cork, Ireland. It was the world premiere of it, in the tiny cinema in the Triskel Arts Centre as part of the Cork Film Festival. I found "Strange Fruit" to be an excellent movie. [...]	Negative
This movie was pure genius. John Waters is brilliant. It is hilarious and I am not sick of it even after seeing it about 20 times since I bought it a few months ago. The acting is great, although Ricki Lake could have been better. And Johnny Depp is magnificent. He is such a beautiful man and a very talented actor. And seeing most of Johnny's movies, this is probably my favorite. I give it 9.5/10. Rent it today!	Negative

Original Annotation

1 Person Regula Susana Siegfried , 50 , and Misc Nicola Person Fleuchaus , 25 , were released after 71 days after a $ 200,000 ransom was paid.

2 Person Laurence Courtois (Location Belgium) beat Location Flora Person Perfetti (Location Italy) 6-4 3-6 6-2

3 Organization Hapoel Haifa 3 Maccabi Org Tel Aviv 1

4 Org Sporting Gijon 15 4 4 7 15 22 16

5 Org St. Gallen 9 4 4 1 6 5 16

Corrected Annotation

1 Person Regula Susana Siegfried , 50 , and Person Nicola Fleuchaus , 25 , were released after 71 days after a $ 200,000 ransom was paid.

2 Person Laurence Courtois (Location Belgium) beat Person Flora Perfetti (Location Italy) 6-4 3-6 6-2

3 Organization Hapoel Haifa 3 Organization Maccabi Tel Aviv 1

4 Organization Sporting Gijon 15 4 4 7 15 22 16

5 Organization St. Gallen 9 4 4 1 6 5 16

Figure 3: Five instances with highest error scores assigned by ActiveAED for IMDb (top) and CONLL-2003 (bottom; visualized with brat (Stenetorp et al., 2012)). All original annotations are erroneous.

	$\lvert\mathcal{I}\rvert$	$\lvert\mathcal{I}_\epsilon\rvert$	$\frac{\lvert\mathcal{I}_\epsilon\rvert}{\lvert\mathcal{I}\rvert}\%$	$\lvert\mathcal{A}\rvert$	$\lvert\mathcal{A}_\epsilon\rvert$	$\frac{\lvert\mathcal{A}_\epsilon\rvert}{\lvert\mathcal{A}\rvert}\%$	Datasets URL	License
ATIS	4978	238	4.8	4978	238	4.8	mainlp/aed_atis	LDC
IMDb	25,000	725	2.9	25000	725	2.9	mainlp/pervasive_imdb	GPL3
SST	8544	427	5.0	8544	427	5.0	mainlp/aed_sst	unknown
GUM	1117	552	49.4	13480	929	6.9	mainlp/aed_gum	Online
CoNLL-2003	18,463	761	4.1	13870	1133	8.2	mainlp/aed_conll	Online
SI-Companies	500	454	90.8	7310	1650	22.6	mainlp/inconsistencies_companies	CC-BY4
SI-Flights	500	224	44.8	2571	420	16.3	mainlp/aed_atis	CC-BY4
SI-Forex	520	143	27.5	1632	326	20.0	mainlp/inconsistencies_forex	CC-BY4

Table 2: Statistics for all datasets that we used in this study. $\lvert\mathcal{I}\rvert$ is the number of total instances, whereas $\lvert\mathcal{I}_\epsilon\rvert$ is the number of instances with at least one wrong annotation. $\lvert\mathcal{A}\rvert$ is the number of total annotations and $\lvert\mathcal{A}_\epsilon\rvert$ the number of erroneous annotations. Datasets URL is the name of our implementation of the dataset in the huggingface datasets hub, which allows to deterministically reproduce all datasets.

	ATIS	SI-Flights
ActiveAED	98.6±0.1	86.6±0.5
w/o active	98.7±0.1	**80.3±0.6**
w/o test ens.	**98.3±0.1**	**84.3±0.5**
w/o train ens.	**97.4±0.3**	89.2±1.2
k = 100	**98.5±0.1**	**86.4±0.5**
k = 200	98.7±0.0	**82.7±0.7**

Table 3: Results of the ablation study. All modifications denote independent changes to ActiveAED. I.e. 'w/o test ens.' is ActiveAED without the test ensembling but with train ensembling and the human-in-the-loop component. Scores with a lower average than that of ActiveAED are in bold.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Left blank.

☑ A2. Did you discuss any potential risks of your work?
Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4

☑ B1. Did you cite the creators of artifacts you used?
Section 4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix A

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix A

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix A

C ☑ Did you run computational experiments?

Sections 4 and 5, Appendices C, D, E, and F

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix E

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.