Revisiting Sample Size Determination in Natural Language Understanding

Ernie Chang[†*], **Muhammad Hassan Rashid**[‡*], **Pin-Jie Lin**[‡*],
Changsheng Zhao[†], **Vera Demberg**[‡], **Yangyang Shi**[†] and **Vikas Chandra**[†]
†Reality Labs, Meta Inc.
‡Saarland Informatics Campus, Saarland University, Germany
{erniecyc, cszhao, yyshi, vchandra}@meta.com
hassanrashid725@gmail.com
pinjie@lst.uni-saarland.de
vera@coli.uni-saarland.de

Abstract

Knowing exactly how many data points need to be labeled to achieve a certain model performance is a hugely beneficial step towards reducing the overall budgets for annotation. It pertains to both active learning and traditional data annotation, and is particularly beneficial for low resource scenarios. Nevertheless, it remains a largely under-explored area of research in NLP. We therefore explored various techniques for estimating the training sample size necessary to achieve a targeted performance value. We derived a simple yet effective approach to predict the maximum achievable model performance based on small amount of training samples – which serves as an early indicator during data annotation for data quality and sample size determination. We performed ablation studies on four language understanding tasks, and showed that the proposed approach allows us to forecast model performance within a small margin of mean absolute error ($\backsim$ 0.9%) with only 10% data[1].

1 Introduction

Labeled data play an important role in creating performant machine learning models, which makes data annotation a fundamental process for any natural language application pipeline (Lewis and Catlett, 1994). Recent work has sought to reduce the annotation costs through the use of active learning (Ducoffe and Precioso, 2018; Margatina et al., 2021) and data sampling (Sener and Savarese, 2018; Coleman et al., 2019; Killamsetty et al., 2021a,b). Indeed, these approaches are shown to be effective in identifying or constructing data subsets needed to achieve a competitive model performance. For instance, the active learning paradigm adds new data iteratively to the existing set before model retraining (Agarwal et al., 2020; Margatina et al., 2021), improving upon the traditional human annotation pipeline that obtains the entire labeled set all at once.

Nevertheless, the data labeling process typically annotates as much data as the annotation budget permits, or by clearly defined stopping criteria to terminate the labeling process. Unfortunately, this is usually challenging as annotators do not have the knowledge of the effect of added labels to model performance nor how much more data is needed to arrive at the desired model generalizability (Killamsetty et al., 2020). The stopping condition is in fact tied to the quality of data samples w.r.t. model parameters (Hu et al., 2021), which influences the effective sample size[2], and it is then beneficial to obtain an approximation of the expected performance (Vlachos, 2008; Olsson and Tomanek, 2009a; Zhu et al., 2010; Ishibashi and Hino, 2020). Therefore, knowing the approximate amount of training data needed for this particular performance would serve as an useful knowledge not only for deciding when to stop adding labeled data, but also as an early indication for the data quality. For instance, by having early label quality signals, we can decide between two different types of annotation, or even between two pools of annotators with different expertise.

To this end, we explored the relationship between *data sample size* and *model performance* in the context of language understanding via learning curve modeling, which defines model performance as a function of dataset sizes. By modeling this relationship in low resource settings, we obtain useful early signals with approximated accuracies for any given the labeled set, which can provide an idea for the sample size and data quality (Olsson and Tomanek, 2009b; Figueroa et al., 2012). Previous studies have shown that nonlinear weighted

∗ These authors contributed equally to this work.

[1]Our code is available at: https://github.com/pjlintw/sample-size.

[2]It is the size of datasets which could have been achieved by an effective unweighted random sample (Guo et al., 2022).

Findings of the Association for Computational Linguistics: ACL 2023, pages 6716–6724
July 9-14, 2023

curve fitting methods such as inverse power laws or exponential functions can provide decent approximations of the empirical predictive performances (Frey and Fisher, 1999; Figueroa et al., 2012). We thus put forward an ensemble of these functions which we showed to display a consistently highly correlated behavior across four language understanding benchmarks and with as little as 10% of the entire training set. This work makes the following contributions:

1. We revisit the task of sample size determination in four natural language understanding benchmarks and empirically explore the correlation strengths of several successful techniques.

2. Based on our findings, we propose an ENSEMBLE function and demonstrated across several benchmarks and low resource settings that the ensemble function is consistently providing a high correlation with the empirical learning curve plots.

2 Background

Our method is a sample size determination technique that helps to design annotation projects by determining the necessary sample size. Previous methods have focused on identifying the sample size required to reach a specific target performance, such as a high correlation coefficient (Beal, 1989; Stalbovskaya et al., 2007; Beal, 1989), which often involves predicting the sample size necessary for a classifier to attain a specific accuracy level (Fukunaga and Hayes, 1989). There are two main approaches for predicting the sample size needed to achieve a particular classifier performance: (1) Dobbin et al. (2008) present a model-based method for predicting the number of samples required for classifying microarray data. (2) A more general approach involves fitting a classifier's learning curve to inverse power law models (Figueroa et al., 2012). Examples of this approach include algorithms proposed by Mukherjee et al. (2003); Boonyanunta and Zeephongsekul (2004); Last (2007).

3 The Approach

Learning Curve Modeling. A learning curve is a graphical representation of how a classifier's performance changes as the size of the training set increases. The curve typically has three sections: an initial section where performance improves rapidly with increasing training set size, a middle section where the rate of improvement slows down, and a final section where the classifier reaches its maximum performance and further increases in training set size do not lead to significant improvements. This relationship can be quantified using a set of data points, each of which represents the expected performance of the classifier E_{acc} on a particular training set size D_k. These data points can be plotted to create the learning curve, which can help to understand the behavior of the classifier and inform decision-making about how much training data is needed to achieve a desired performance level.

Task Description. Given a downstream classification task with N_{total} data points, a learning curve model F predicts the expected performance E_{acc} when a classifier trained on the an observed range of training set size (D_k; $k >= N$). The empirical learning curve is assessed by the parametric models for the learning algorithm performance extrapolation. In our settings, we set $k << N_{total}$ to simulate practical settings, where few data points consisting of (E_{acc}, D_K) are to be obtained.

Types of Extrapolations. Here, we study different forms of learning curve models with few learnable parameters that have been proven as simple yet effective. The simplest type of learning curve model *exponential function* (EXP) only introduces two parameters a and b to fit the exponent behavior of learning curve (Frey and Fisher, 1999). The second form, *Inverse Power Law function* (INVERSE), fits the inverse power law (Figueroa et al., 2012) and has three parameters. The third form uses a function from the power law family – Power4 function (POW4) (Kolachina et al., 2012) with four parameters. Lastly, we propose to combine all functions into one (ENSEMBLE) so that it has all their characteristics in order to make it more robust across benchmarks. Table 1 shows the formulae of our investigated extrapolating functions.

EXTRAPOLATING FUNCTIONS	FORMULA
EXP (A)	$a \cdot N^b$
INVERSE (B)	$(1-a) - b \cdot N^c$
POW4 (C)	$a - (b \cdot N + c)^{-d}$
ENSEMBLE (A+B+C)	–

Table 1: Overview of extrapolating functions

4 Experimental Settings

We study four NLU tasks: (1) IMDB (Maas et al., 2011) is a binary classification dataset (25K/–/25K)[3] where model predicts the sentiment (positive/negative) for movie reviews from IMDB; (2) SST2 (Socher et al., 2013) is also a sentiment classification datatset (67K/0.8K/1.8K) containing reviews of different movies and since the model predicts if the review is positive or negative, it also falls in the category of binary classification; (3) AG NEWS is a multi-class classification dataset (120K/–/7.6K) containing texts from different news where the model predicts whether the news text is about sports, science/technology, world or business from the four different classes. We also consider one other multi-class classification task, (4) DBPEDIA dataset (560K/–/70K) , since it could help us in testing the robustness of the methods used in our experiments.

Configs. To investigate how changes in data size affect the predictiveness of the learning curves, under the assumption that the model structure and settings remain unchanged, we perform all experiments using a transformer model (Vaswani et al., 2017) and average the results over 3 initialization runs. The embedding and hidden layer dimensions are 1000 and 1820; and we use a 6-layer encoder with 4 multi-heads, and the dropout is 0.2. To find the parameters of learning curve models, we consider unweighted and for the gradient descent and non-linear least squares optimizers. The Adam algorithm (Kingma and Ba, 2014) was used as the optimizer with learning rate of 1e-5 and ReLU was used as the activation function. The cross-entropy objective was used for all classification benchmarks, and we select the models using loss values. Finally, we chose a batch size of 8 with 200 number of epochs.

Evaluation. We use the aforementioned functions: EXP, INVERSE, POW4 and ENSEMBLE for fitting the empirical learning curve. For each dataset, we select training set sizes ranging from 1% to 10% data sizes at an interval of 1%. The learning curve testsets were created with the data splits in the range $[55, 100]$ at 5% interval by training the classifier, and obtaining the testset[4] performance for each corresponding data split. Therefore, we collect the accuracies against different sample sizes and report the mean absolute error (MAE) as the evaluation metric for learning curve modeling.

[3]Expressed in the order (train/dev/test).

[4]Here, we make the distinction between testset for learning curve and the original testset split.

5 Results and Analysis

We present results of ensemble method for learning curve modeling on the NLU benchmarks.

5.1 Main Results

Figure 1 demonstrates that by using only 10% of the data for learning curve modeling, ENSEMBLE is able to effectively predict model performance within a 0.9% margin of the actual model performance. Moreover, we observe the same trend across all four benchmarks consisting of different training set sizes (i.e. ranging from 25K to 250K) and varying number of classification classes (i.e. ranging from 2 to 14), see the appendix A for remaining figures. Our result shows that the proposed approach is not confined by the classification types and sample sizes.

Table 2 shows the saturated points of the learning curve when the performance improvement is less than a threshold $\alpha = 0.2$ – we found that the predicted performance with only 19% data is within 2.44 accuracy points from the trained model performance for IMDB. Another key observation is that the size (%) needed to predict a low L1 distance increases as the number of classification classes goes up, which indicates that task difficulty does influence the ease of extrapolation. An example is that AG NEWS requires up to 51% to predict a low L1 distance. Next, we perform further ablation studies to investigate the effect of sample size, types of non-linear functions used, or the effect of data weighting.

BENCHMARK	CLS (#N)	SIZE (%)	SIZE (#N)	L1↓	100%
		$\alpha = 0.2$			
IMDB	2	36%	6,300	2.44	17K
SST2	2	19%	8,958	5.57	47K
AG NEWS	4	51%	42,840	2.6	84K
DBPEDIA	14	51%	199,920	2.39	392K

Table 2: CLS (#N) stands for the number of classes, SIZE (%) for the percentages of the data size for the learning curve modeling. SIZE (#N) is the number of the corresponding data size, L1 is the L1 distance between the accuracy of models using all the training data and the estimated accuracy based on the saturated point. 100% specifies all training samples for learning curve.

5.2 Ablation Study

Effect of sample size. In Figure 1, we study the correlation between sample sizes and the absolute

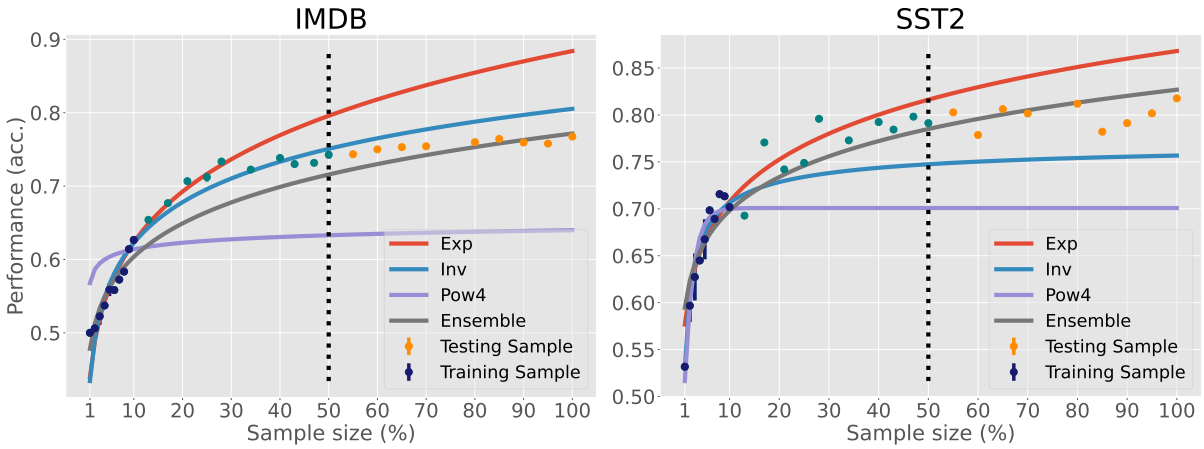

Figure 1: Learning curves on 10% sample size of IMDB and SST2 datasets. We plot learning curves using the exponential (Exp), inverse power law (Inv), power4 (Pow4) function, and the combination of the aforementioned forms (Ensemble). The learning curves only fit on 10% training sample (blue) and generalize on the unseen data sizes. We evaluate the learning curves on the testing sample (yellow). Data sizes from 10% to 50% (teal) are neither used in training nor testing.

mean error between the learning curve model and empirical model performance trend. Surprisingly, we discovered by having more samples does not necessarily help with modeling a better learning curve[5], and that with only 10% data to build the (D_k, E_{acc}) data points is sufficient to obtain rather small errors across all four benchmarks.

Types of learning curve functions. We are also interested in seeing how each of the non-linear learning curve function fare against each other in simpler settings. To this end, we used up to 10% data to model the learning curves and obtained their respective mean absolute error values. In Figure 1, we present this comparison where we showed that on IMDB and SST2, the ENSEMBLE function consistently fit best against the empirical data. We observed a similar trend across other benchmark DBPEDIA with the exception of AG NEWS. We placed the plot for AG NEWS in appendix A.3.

Influence of data weighting. Previous work (Paul et al., 2021; Guo et al., 2022) has found that not all data points are equally important in terms of curve fitting. In fact, data points at a later phase corresponding to more samples are to be given more weight compared to earlier points. We thus investigate this phenomenon in the context of our benchmark, and we observed this to be true anecdotally. The detailed result can be found in Appendix A.2. The reason for this is that the more data samples there are, the more closely they resemble the entire training set, and this makes their signals a better estimation of a point on the actual learning curve. Another perspective is that the more data samples are used, the less the effect of random sampling on the performance, which affects model performance in extremely low resource scenarios.

FUNCTION TYPE	NON-LINEAR LEAST SQUARES	
	UNWEIGHTED	WEIGHTED
EXP	0.0417	**0.0244**
INV	0.00777	**0.00442**
POW4	0.00795	0.00795

Table 3: Better curve fitting when weighting data points at latter phase. We examine the effectiveness of weighting data size on the exponential (EXP), inverse power law (INV), power4 (POW4) function using non-linear least squares method. The learning curves fit on 5%, 10%, 25% and 50% data sizes of IMDB and is evaluated on testing sample with mean absolute error (MAE).

6 Conclusions and Future Works

In this work, we investigated techniques for estimating the amount of training data needed to achieve a target performance in four natural language understanding benchmarks. We demonstrated that our approach allows for accurate prediction of model performance using only a small portion of the data, which can be useful in scenarios with limited resources. Nevertheless, we also recognize the limitation in our current study. For instance, we did not explore sampling techniques other than random sampling; while recent works (Yuan et al., 2020; Paul et al., 2021; Guo et al., 2022) have shown promising directions in data sampling that outperforms random sampling. Another interesting

[5]We showed this result in the Appendix A.5.

direction is to explore the model architecture's influence on generalizability, and thus the learning curve, which we left for future works.

Limitations

While the effectiveness of the expressive learning curve in settings with limited data has been demonstrated, it is uncertain if this success can be replicated in more complex natural language understanding tasks, such as question answering or tasks that involve a large amount of data. Furthermore, it is assumed that all data samples have the same impact on the model's performance. However, the actual performance of the model may vary based on the method used to select the data or the specific set of tasks being performed, e.g., coreset selection. Similarly, the quality of the labels used for the data can also play a significant role in predicting the performance of the model. Overall, we plan to further investigate these questions and explore them in future studies.

Ethics Statement

We address the efficiency of data annotation by investigating learning curves to estimate the necessary training sample size to reach a desired model performance. However, it is imperative to take into consideration the potential biases that may exist in the model predictions when utilizing a reduced amount of labeled data in the system construction process. Furthermore, when addressing complex tasks such as machine translation and text summarization, it is essential to guarantee the factuality of output generated by the system trained with the suggested data sample size.

References

Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. 2020. Contextual diversity for active learning. In *ECCV*, pages 137–153. Springer.

S L Beal. 1989. Sample size determination for confidence intervals on the population mean and on the difference between two population means. *Biometrics*, 45(3):969–977.

Natthaphan Boonyanunta and Panlop Zeephongsekul. 2004. Predicting the relationship between the size of training sample and the predictive power of classifiers. *Knowledge-Based Intelligent Information and Engineering Systems*, 3215:529–535.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. 2019. Selection via proxy: Efficient data selection for deep learning. *arXiv preprint arXiv:1906.11829*.

Kevin K Dobbin, Yingdong Zhao, and Richard M Simon. 2008. How large a training set is needed to develop a classifier for microarray data? *Clin. Cancer Res.*, 14(1):108–114.

Melanie Ducoffe and Frederic Precioso. 2018. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*.

Rosa L Figueroa, Qing Zeng-Treitler, Sasikiran Kandula, and Long H Ngo. 2012. Predicting sample size required for classification performance. *BMC Med Inform Decis Mak*, 12.

Lewis J. Frey and Douglas H. Fisher. 1999. Modeling decision tree performance with the power law. In *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*, volume R2 of *Proceedings of Machine Learning Research*. PMLR. Reissued by PMLR on 20 August 2020.

K Fukunaga and R R Hayes. 1989. Effects of sample size in classifier design. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(8):873–885.

Chengcheng Guo, Bo Zhao, and Yanbing Bai. 2022. Deepcore: A comprehensive library for coreset selection in deep learning.

Xia Hu, Lingyang Chu, Jian Pei, Weiqing Liu, and Jiang Bian. 2021. Model complexity of deep learning: A survey. *Knowledge and Information Systems*, 63(10):2585–2619.

Hideaki Ishibashi and Hideitsu Hino. 2020. Stopping criterion for active learning based on deterministic generalization bounds. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 386–397. PMLR.

Krishnateja Killamsetty, S Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. 2021a. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *ICML*, pages 5464–5474.

KrishnaTeja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh K. Iyer. 2020. GLISTER: generalization based data subset selection for efficient and robust learning. *CoRR*, abs/2012.10630.

Krishnateja Killamsetty, Xujiang Zhao, Feng Chen, and Rishabh Iyer. 2021b. Retrieve: Coreset selection for efficient and robust semi-supervised learning. *arXiv preprint arXiv:2106.07760*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Prasanth Kolachina, Nicola Cancedda, Marc Dymetman, and Sriram Venkatapathy. 2012. Prediction of learning curves in machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22–30, Jeju Island, Korea. Association for Computational Linguistics.

Mark Last. 2007. Predicting and optimizing classifier utility with the power law. *Seventh IEEE International Conference on Data Mining Workshops*, pages 219–224.

David D Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier.

Andrew 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.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. *arXiv preprint arXiv:2109.03764*.

Sayan Mukherjee, Pablo Tamayo, Simon Rogers, Ryan Rifkin, Anna Engle, Colin Campbell, Todd R Golub, and Jill P Mesirov. 2003. Estimating dataset size requirements for classifying dna microarray data. *Comput Biol*, 10:119–142.

Fredrik Olsson and Katrin Tomanek. 2009a. An intrinsic stopping criterion for committee-based active learning. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 138–146, Boulder, Colorado. Association for Computational Linguistics.

Fredrik Olsson and Katrin Tomanek. 2009b. An intrinsic stopping criterion for committee-based active learning. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 138–146.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34:20596–20607.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *ICLR*.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. 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.

Viktoriya Stalbovskaya, Brahim Hamadicharef, and Emmanuel C Ifeachor. 2007. Sample size determination using ROC analysis. In *3rd International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Andreas Vlachos. 2008. A stopping criterion for active learning. *Computer Speech and Language*, 22(3):295–312.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online. Association for Computational Linguistics.

Jingbo Zhu, Huizhen Wang, Eduard Hovy, and Matthew Ma. 2010. Confidence-based stopping criteria for active learning for data annotation. *ACM Trans. Speech Lang. Process.*, 6(3).

A Detailed Results

A.1 Predicting the Required Data Size

Table 4 presents the results of required data size prediction using threshold $\alpha = 0.1$ and $\alpha = 0.3$.

BENCHMARK	CLS (#)	SIZE (%)	SIZE (#N)	L1↓	100%
		$\alpha = 0.1$			
IMDB	2	19%	16,800	6.56	17K
SST2	2	8%	25,458	8.27	47K
AG NEWS	4	28%	82,320,	2.96	84K
DBPEDIA	14	27%	384,160	3.44	392K
		$\alpha = 0.3$			
IMDB	2	96%	3,325	5.84	17K
SST2	2	54%	3,772	0.704	47K
AG NEWS	4	98%	23,521	9.9	84K
DBPEDIA	14	98%	105,840	9.68	392K

Table 4: We show the results with the thresholds $\alpha = 0.1$ and $\alpha = 0.3$. SIZE (%) stands for the percentages of the data size for the learning curve modeling, SIZE (#N) is the number of the corresponding data size, L1 is the L1 distance between the accuracy of models using all the training data and the estimated accuracy based on the saturated point. 100% specifies all training samples.

A.2 Data Weighting

We apply data weighting on three extrapolating functions using gradient decent methods in 5.

EXTRAPOLATING	GRADIENT DESCENT	
	UNWEIGHTED	WEIGHTED
EXP	0.0417	**0.0342**
INV	0.0706	**0.0519**
POW4	0.0979	**0.0652**

Table 5: Better curve fitting when weighting data points at latter phase. We examine the effectiveness of weighting data size on the exponential (EXP), inverse power law (INV), power4 (POW4) function using gradient decent method. The learning curves fit on 5%, 10%, 25% and 50% data sizes of IMDB and is evaluated on testing sample with mean absolute error (MAE).

A.3 Learning curve on 10% data sizes of AG NEWS

Figure 2 shows the learning curves fitting on 10% data sizes of AG NEWS dataset.

A.4 Learning curve on 10% data sizes of DBpedia

Figure 3 shows the learning curves fitting on 10% data sizes of DBPEDIA dataset.

A.5 Effect of Sample Sizes for Learning Curve Fitting

We examined the relationship between sample sizes and the difference in mean absolute error (MAE) between the predicted and actual performance trends across four benchmarks. Table 6 showed MAEs when ENSEMBLE fitting on 50% and 10% of data respectively. We observed that having more samples does not necessarily lead to a better model and that using only 10% resulted in smaller MAEs on all four benchmarks. Therefore, we select 10% of data points for learning curve modeling.

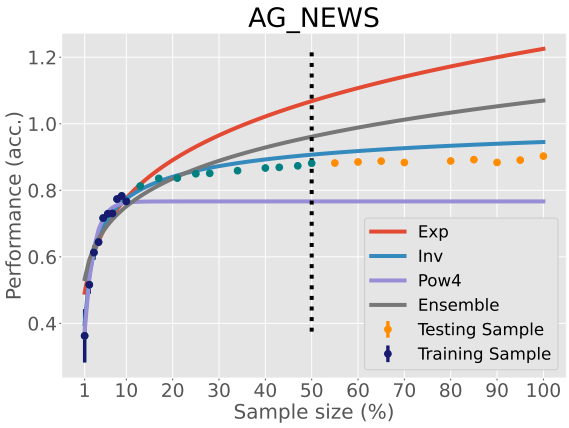

Figure 2: Learning curves on 10% data size using AG NEWS.

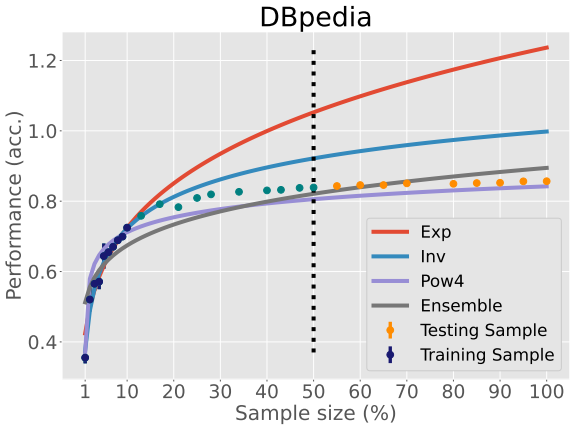

Figure 3: Learning curves on 10% data size using DBPEDIA.

BENCHMARK	SAMPLE SIZES	
	50%	10%
IMDB	0.0458	**0.00961**
SST2	0.0299	**0.0132**
AG NEWS	0.0704	**0.0209**
DBPEDIA	0.0734	**0.0158**

Table 6: Learning Curve Fitting on 50% and 10% data size respectively.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7 (after conclusions)

☑ A2. Did you discuss any potential risks of your work?
8 (ethics statement)

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

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

B ☑ Did you use or create scientific artifacts?

4

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

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We adopted widely-used datasets for our investigation.

☑ 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)?
4

☐ 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. We do not collect data and we adopted widely-used datasets for our investigation.

☐ 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.
described in section 3

C ☑ Did you run computational experiments?

4 and 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

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?
4

☑ 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?
4 and 5

☑ 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.)?
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.?
Not applicable. Left blank.

☐ 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)?
Not applicable. Left blank.

☐ 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?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.