Data Augmentation for Text Classification with EASE

A M Muntasir Rahman[1] **Wenpeng Yin**[2] **Guiling "Grace" Wang**[1]
[1]Department of Computer Science, New Jersey Institute of Technology
[2]Department of Computer Science & Engineering, Penn State University
{ar238,gwang}@njit.edu
wenpeng@psu.edu

Abstract

In this work, we present **EASE**, a simple but dependable Data Augmentation (DA) technique for Text Classification (TC) that has four easy steps: **E**xtract Units, **A**cquire Labels, **S**ift and **E**mploy. We extract meaningful units as augmented samples from original data samples and use powerful tools to acquire labels for them before they are sifted and merged. Previous DA techniques, like EDA-Easy DA (Wei and Zou, 2019) and AEDA-An Easier DA (Karimi et al., 2021), excel with sequential, RNN-based models but struggle with BERT (Devlin et al., 2019) and other transformer-based models that heavily rely on token order. EASE, in contrast, performs well with these models, demonstrating stability, speed, and minimal adverse effects. We tested our intuitive method on multiple challenging datasets sensitive to augmentation, and experimental results have indicated the efficacy of DA with EASE.

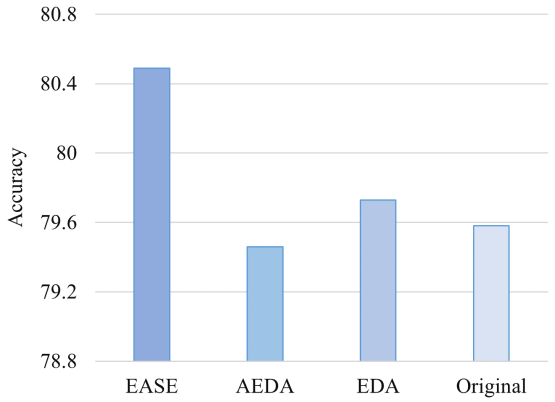

Figure 1: Averaged accuracy accross all datasets and models used in the low-resource experiments.

.

1 Introduction

DA is a fairly common technique in Machine Learning, especially in Computer Vision and Speech Recognition, and there are many standard ways of doing it. For example, simply flipping or rotating an image and labeling it the same as the original sample is quite logical. While these techniques do involve elements of randomness, they can still be regarded as logically labeled samples, distinct from random noise. This distinction is essential for enhancing the interpretability of complex deep learning models, a challenge often encountered in several notable NLP DA techniques, including EDA (Wei and Zou, 2019) and AEDA (Karimi et al., 2021), among others.

In EDA (Wei and Zou, 2019), four random operations—Random Synonym Replacement, Random Insertion, Random Swap, and Random Deletion—are employed. These operations, when applied even moderately, can significantly alter the original text's meaning in text classification. A single token replacement, for instance, can reverse the sentiment of a sentence. Similarly, In AEDA (Karimi et al., 2021), random punctuation marks like question marks and periods are inserted into samples, radically altering sentence structure and causing confusion in the training model. Despite their proven effectiveness in ideal situations, these techniques often hinder performance. Particularly in the era of Transformers (Vaswani et al., 2017), where positional encodings are crucial and depend on token order, random rearrangement disrupts the models' contextual understanding. Hence, the demand for a DA method that accounts for this critical aspect became apparent.

The rise of large language models, such as BERT-base (110M parameters), necessitates a DA (DA) technique that avoids substantial expansion of the training set and the associated increase in training time. Notably, EDA and AEDA suggest a substantial 9-10 times dataset size augmentation, significantly impacting fine-tuning duration. Furthermore, transformer-based models have eclipsed RNN-based models, rendering experiments with EDA or AEDA on the latter obsolete. These models' potent bidirectional contextual representations

demand robust DA methods and more challenging datasets. Given the substantial resources needed for fine-tuning, a reliable DA approach that minimizes hyper-parameter search and ensures favorable outcomes is essential. Additionally, previous complexities attributed to resource constraints, like GPUs and user-friendly frameworks, are no longer valid arguments, enabling the seamless application of intricate processes to crucial tasks such as DA.

We developed a 4-step technique for text classification data augmentation that is time-efficient, stable, intuitive and outperforms existing DA methods. Our experiments with five transformer-based models and four datasets validate our approach, showcasing its superior performance and reliability (Figure 1 and 2).

2 Relevant Studies

In NLP, DA can be challenging due to the contextual nature of the data. Preserving relative word positions is crucial for contextual text embedding, but many existing DA techniques disrupt coherence by introducing random synonyms, punctuations, or altering token order. Regarding ground truth, research falls into two categories: one conserves the original ground truth, while the other generates ground truth based on the augmented sample, with subsequent studies aligning with one of these approaches.

Fadaee et al. (2017) introduced Translation DA for Neural Machine Translation (NMT) by replacing common words with unique words in both source and target sentences. Sennrich et al. (2016) used automatic translation of additional monolingual data for NMT augmentation. Back-translation techniques, as employed by Silfverberg et al. (2017) and Yu et al. (2018), aimed to capture paraphrases for various NLP settings. In addition to EDA (Wei and Zou, 2019), other studies focused on synonym replacements (e.g., Wang and Yang, 2015; Kolomiyets et al., 2011; Zhang et al., 2015). Kobayashi (2018) replaced words with predicted words from BERT, while Andreas (2020) replaced sentence segments with similar contextual counterparts. Sun et al. (2020) used transformers to interpolate input sequences for generating new samples and labels. Additionally, Karimi et al. (2021) compared their work with Xie et al. (2017), viewing it as a data-noising approach to enhance training architectures in NLP.

Many of these approaches, such as AEDA (Karimi et al., 2021) and the work by Xie et al. (2017), often resemble data-noising methods rather than true DA, lacking coherent sentence structures in augmented samples. This falls short of achieving the clarity and human interpretability found in computer vision's approach. To address this, our method extracts coherent, meaningful units from samples, leading to logical samples that surpass existing techniques that disrupt token orders. While most of our experiments focus on text classification due to space constraints, our approach is adaptable to various NLP tasks and holds the potential to become an industry standard.

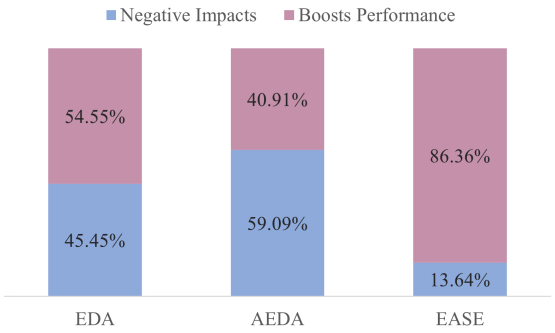

Figure 2: EASE has significantly fewer negative-impacts on performance with different hyper-parameters compared to EDA & AEDA

.

3 EASE

DA with EASE has 4 easy steps that are intuitive and effective.

Extracting Units: In EASE, the most critical step involves extracting meaningful units as augmented samples. The choice of unit depends on the sample structure. For paragraphs, we recommend extracting sentences using the NLTK library (Bird et al., 2009). When dealing with sentences, we suggest extracting "Facts" as introduced by Yuan et al. (2020). These Facts represent coherent sentence units containing logical information. They also preserve token sequences crucial for attention mechanisms in Transformer-based models. For detailed information on extracting facts from sentences, we refer readers to Yuan et al. (2020).

Label Acquisition: In the subsequent EASE step, labels are obtained using pretrained models. The extracted meaningful token sequences make it straightforward for pretrained models to generate high-quality labels without the need for additional training. For our experiments, we employed the

Dataset	Method	Accuracy
Large Movie Review Dataset	Original	86.94%
	EDA	86.72%
	AEDA	86.5%
	EASE	**87.64%**
Sentiment 140	Original	58.67%
	EDA	57.93%
	AEDA	58.23%
	EASE	**60.20%**
Financial Phrase Bank	Original	84.50%
	EDA	**85.20%**
	AEDA	84.72%
	EASE	85.10%
Customer Review	Original	88.55%
	EDA	88.54%
	AEDA	88.32%
	EASE	**89.10%**

Table 1: Comparing EASE, EDA, AEDA for four different datasets in low-resource scenarios by varying the number of augmented samples from small to full size. For Customer Review, the numbers represent average accuracy over three different training subsets with one model, for the other datasets the average is taken over 5 different models and augmentation size variations (Complete detail available in Appendix D). **Bold** suggests the best performance across each column for each dataset.

Dataset	Method	Accuracy
Large Movie Review Dataset	Original	53.37%
	EDA	56.02%
	AEDA	53.08%
	EASE	**67.82%**
Sentiment 140	Original	44.15%
	EDA	44.91%
	AEDA	45.69%
	EASE	**46.00%**
Financial Phrase Bank	Original	55.12%
	EDA	55.35%
	AEDA	55.89%
	EASE	**56.22%**

Table 2: Comparison among EASE, EDA, AEDA for three different datasets in **extremely** low-resource scenarios (only 10 training samples). The performances represent the average over 5 different models (Complete detail available in Appendix D). **Bold** suggests the best performance across each column for each dataset, and parentheses suggest a negative impact on performance)

default pretrained DistilBERT model (fine-tuned on the SST-2 dataset (Socher et al., 2013)) from the HuggingFace library (Wolf et al., 2020) for label generation. In the results section, we present ablation studies to highlight the significance of this step. Nevertheless, it is worth noting that our method can yield promising results even without the label acquisition process.

Sift & Employ : In the "Sift" step, we recommend filtering out smaller-length samples. In our experiments, we retained 10%, 25%, 50%, or 100% of the augmented samples, but it rarely adversely affects performance. This optional step underscores the stability of our method, which is not a random noise injector but a DA technique that complements original training samples. Subsequently, in the "Employ" step, the augmented samples are seamlessly integrated with the original ones completing the final training set.

4 Experimental Setup

We view EDA & AEDA to be the most relevant to our study and showcase performance comparisons for these two methods. Fine-tuning for transformers is usually performed for 5-15 epochs, and from all our experiments, we observe that max validation accuracy is reached before the 30th epoch for these models, but we still performed all the fine-tuning for up to 50 epochs for completeness (More detail on performance saturation in Appendix B). The compared methods differ in augmentation processes: they generate a fixed number of augmented samples per original sample (recommended from 1 to 16), while our approach adapts to sample structure. On average, Fact extraction increases the training dataset by 2.3 times, and sentence extraction by 5.92 times.

4.1 Datasets and Models

For our experiments we used four different sentiment classification datasets. Large Movie Review Dataset (IMDB50K or IMDB) (Maas et al., 2011), Financial Phrasebank (Malo et al., 2014), Customer Review (Hu and Liu, 2004), and Sentiment 140 (Go et al., 2009). We used five different models for our experiments. These are, Bert-base-cased, Bert-base-uncased (Devlin et al., 2019), Distilbert-base-cased, Distilbert-base-uncased (Sanh et al., 2019) and Albert-base-v1 (Lan et al., 2020). We used Huggingface's (Wolf et al., 2020) implementation of these models, a popular Transformer library.

5 Results

5.1 Low-Resource Setting

The original datasets, comprising high number of samples (e.g., 25,000 for IMDB50K, 1.6 Million for Sentiment 140), is adequate for high-performing models like Transformers. To simulate a low-resource scenario, we use only a small subset (e.g., 1000 for Sentiment 140) of the original training sets for data augmentation and generate significantly lower amount of augmented samples compared to EDA & AEDA. (DA) (Details in Appendix D).

In the Sentiment 140 dataset, we have observed that EASE derives benefits from generating augmented samples for the Neutral class, a class that is absent in the original training set but exists in the test set. This stands in contrast to EDA and AEDA. Additionally, EASE demonstrates superior stability and performance.

We observe higher accuracy gain and fewer negative impacts with EASE on average across the board (table 1 & figure 2). Even though, on average, the accuracy gain seems to be higher for EDA, we see the highest accuracy gain of 3.2% in bert-base-cased and fewer negative-impacts with our method for Financial Phrase Bank (Complete table in Appendix D).

Although this study focuses on low-resource scenarios, we still show that our method has promise in high-resource scenarios. Tests on the CR dataset using different portions of the original dataset (500, 2000, and Full) shows that even with the complete dataset, our method outperforms the two other methods, with approximately 10-16x fewer number of augmented samples required (table 1, see Appendix fig. 5 for details).

On an average, we see the best accuracy improvement in 3 out of the 4 datasets with EASE (figure 3). While the other two methods fail to achieve performance boost on an average on 3 out of the 4 datasets, EASE steadily increases performance across all the four different datasets, speaking to the robust nature of our method.

5.2 Extremely Low-Resource Setting

We test the robustness of our method by simulating extremely low-resource scenarios where only 10 training samples are available for fine-tuning and therefore, augmentation. Table 2 demonstrates that even in extremely low-resource setting our method outperforms the other two methods.

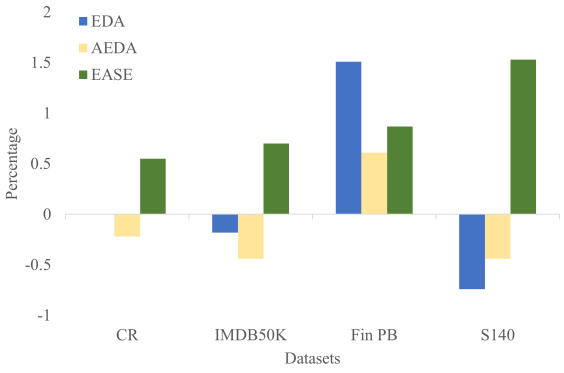

Figure 3: Average accuracy increase over different datasets. EASE showing greater number of and more stable accuracy improvement compared to EDA & AEDA

.

6 Ablation Study

	EASE	EASE-A
Avg. Acc. Gain	1.12%	-0.50%
Neg. Impact	16%	60%
Pos. Impact	84%	40%

Table 3: Average Performance of EASE vs EASE without Acquiring labels on IMDB50K & S140

As an ablation study, we try to measure how important acquiring new labels for the augmented samples is. We use IMDB50K & S140 dataset and test our method by preserving labels. We use the same augmented and original dataset partitions used in typical experiments. The details are summarized in table 3. See Appendix table 8 for details.

7 Conclusion

We introduced an efficient DA technique for TC that improves accuracy without significantly extending training time. Our method outperforms AEDA & EDA in performance, stability, and efficiency. While currently tailored for TC, we envision its adaptation to various NLP tasks with minimal modifications. For instance, equivalent units can be derived from larger samples for Machine Translation using the same technique as EASE to feed the model augmented samples that provide a more nuanced and granular understanding of the training text. Future work will explore additional extraction units and label acquisition methods.

References

Jacob Andreas. 2020. Good-enough compositional data augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7556–7566, Online. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

William Falcon. 2019. `https://www.pytorchlightning.ai`. Accessed: 2022-08-13.

Alec Go, Richa Bhayani, and Lei Huang. 2009. `http://help.sentiment140.com/home`. Accessed: 2022-08-13.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021. AEDA: An easier data augmentation technique for text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2748–2754, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2011. Model-portability experiments for textual temporal analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 271–276, Portland, Oregon, USA. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *ICLR*. OpenReview.net.

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.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Miikka Silfverberg, Adam Wiemerslage, Ling Liu, and Lingshuang Jack Mao. 2017. Data augmentation for morphological reinflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 90–99, Vancouver. Association for Computational Linguistics.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–465, Sofia, Bulgaria. Association for Computational Linguistics.

Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip Yu, and Lifang He. 2020. Mixup-transformer: Dynamic data augmentation for NLP tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3436–3440, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, Lisbon, Portugal. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ziang Xie, Sida I. Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y. Ng. 2017. Data noising as smoothing in neural network language models.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension.

Ruifeng Yuan, Zili Wang, and Wenjie Li. 2020. Fact-level extractive summarization with hierarchical graph mask on BERT. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5629–5639, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

A Example Augmentations

Table 4 shows two different kinds of augmentated samples with EASE.

Data	Text	Label
	Fact Extraction	
Orig.	The monitor is simply amazing, however, it does not support HDMI input	pos
Aug. 1	The monitor is simply amazing, however,	pos
Aug. 2	it does not support HDMI input	neg
	Sentence Extraction	
Orig.	Actually I'm surprised there were so many comments about this movie. I saw it as part of a Slavic film festival at a major American University. But nobody in USA has heard of it, which is a real shame!	pos
Aug. 1	Actually I'm surprised there were so many comments about this movie.	pos
Aug. 2	I saw it as part of a Slavic film festival at a major American University.	pos
Aug. 3	But nobody in USA has heard of it, which is a real shame!	neg

Table 4: Original sentence and the augmented samples generated and labelled through EASE using Fact or Sentence Extraction.

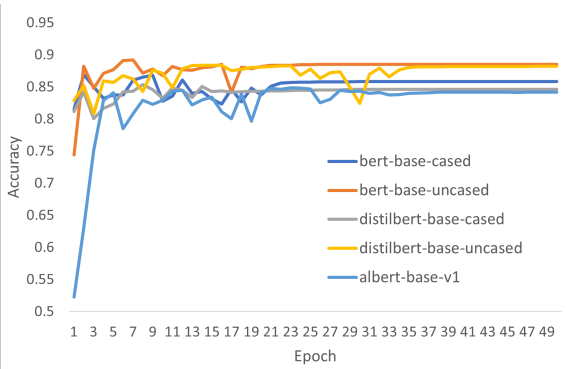

Figure 4: Performance Saturation after 30 epochs for the unaugmented IMDB50K dataset with 500 samples over different models

.

B Performance Saturation

Since the transformer models are already pretrained on unlabelled data, very little amount of fine-tuning is required to gain good task-oriented performance from them. It also must be noted that because of

	Original		EASE			EDA			AEDA		
	Train	Test	Small	Med	Full	Small	Med	Full	Small	Med	Full
CR	500	451	56	282	564	500	2500	4500	500	4000	8000
CR	2000	451	216	1083	2167	2000	10000	18000	2000	8000	32000
CR	4067	451	443	2217	4434	4067	20335	36603	4067	32536	65072
IMDB	500	25000	1000	-	5962	500	-	4500	500	-	4000
FinPB	1000	485	123	-	1235	1000	-	8000	1000	-	9000
S140	1000	497	111	559	1118	1000	5000	9000	1000	4000	8000

Table 5: Number of augmentations used in each experiments for each dataset and each method

the large size of the Transformer based models, even fine-tuning for 50 epochs on multiple GPUs using distributed strategies requires a long time. We discuss more about this in the subsequent section. In all our experiments, we have observed that the validation accuracy in most scenarios saturates after the 30th epoch. In figure 4 we show how fine-tuning for more than 30 epochs is not required. Nevertheless, we still performed all our experiments for 50 epochs for completeness.

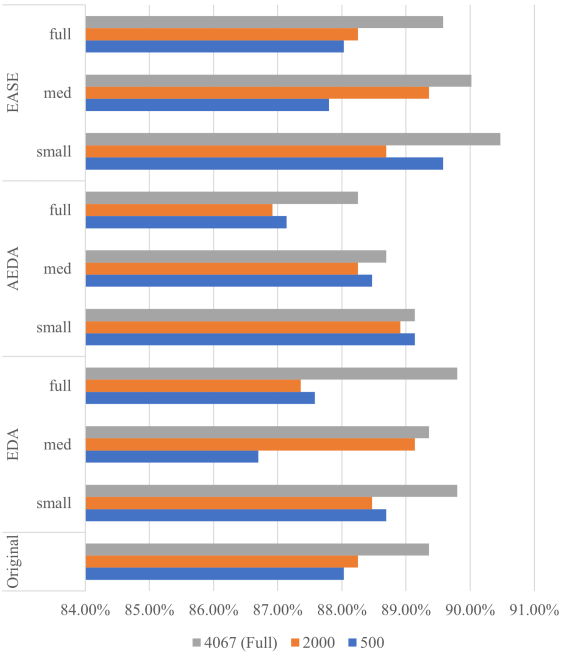

Figure 5: Performance comparison on CR dataset on different training set size using bert-base-cased

.

C Discussion on Training Time

While GPUs are more accessible and distributed training with tools like PyTorch Lightning (Falcon, 2019) has simplified, neural network models are growing larger to balance it out. Transformer models are notorious for taking a massive amount of time for training. To put things into perspective, fine-tuning the Bert-base-cased model for 50 epochs with AEDA-full-augmented IMDB50K dataset (4500 training samples & 25000 testing samples) with 2 Nvidia Tesla P100 GPUs (Each with 16GB Memory) required 12.6 Hours and AEDA-full-augmented Customer Review dataset (65,072 training samples and 451 testing samples) required 26.3 hours. Naturally, searching hyperparameter (number of augmentation) to figure out the optimal augmented dataset that boosts performance is a non-trivial factor to consider while choosing the data augmentation method. For the Customer Review dataset, it took more than 2 days of training to get the results for the different number of augmentation samples, while our method took only 3.4 hours of training. After exploring this vast search space, our method boosted performance 8 out of 9 times, whereas AEDA boosted performance 3 out of 9 times (average performance gain is also in the negative for AEDA). In a low-training-resource scenario, the amount of DA is essential, so a dependable method is required. For these reasons, although our method outperforms EDA & AEDA, we also want to focus on the time-efficient and stable nature of our method.

D Training Set Size and Performance Details

To simulate low-resource settings, small subsets of original training sets were used. Table 5 presents these numbers for each dataset. Model-wise performances are laid out in table 6 for low-resource experiments, in table 7 for extremely-low resource experiments, and in table 8 for the ablation study of label preservation. Customer Review dataset were partitioned into 3 different sets and the accuracy comparisons are showcased in figure 5.

	bert-base-cased	bert-base-uncased	distilbert-base-cased	distilbert-base-uncased	albert-base-v1
Large Movie Review Dataset					
Original	86.92%	89.20%	85.36%	**88.38%**	84.86%
+EDA-small	87.04%	(88.58%)	(85.03%)	(86.84%)	85.17%
+EDA-full	87.26%	(88.22%)	(85.26%)	(87.63%)	86.59%
+AEDA-small	87.31%	(88.98%)	(84.62%)	(87.37%)	85.59%
+AEDA-full	(85.83%)	(88.26%)	(84.98%)	(86.50%)	85.52%
+EASE-small	87.74%	89.40%	85.47%	(87.72%)	86.46%
+EASE-full	**88.01%**	**89.60%**	**86.80%**	(87.72%)	**87.50%**
Sentiment 140					
Original	60.36%	60.56%	59.15%	57.75%	55.53%
+EDA-small	60.56%	61.77%	59.15%	57.95%	56.14%
+EDA-medium	(58.35)%	(58.95%)	(57.34%)	58.15%	(54.12%)
+EDA-full	(57.75)%	(58.55%)	(57.95%)	(57.55%)	(54.73%)
+AEDA-small	(58.95)%	(60.36%)	(58.55%)	58.55%	56.14%
+AEDA-medium	(59.96)%	(60.36%)	(58.35%)	(56.94%)	56.74%
+AEDA-full	(58.15)%	(59.96%)	(56.34%)	57.75%	56.34%
+EASE-small	(59.76)%	**63.18%**	(58.75%)	58.55%	57.75%
+EASE-medium	60.97%	62.17%	**59.96%**	60.36%	**57.95%**
+EASE-full	**62.98%**	62.37%	**59.96%**	**61.97%**	56.34%
Financial Phrase Bank					
Original	84.12%	87.01%	83.71%	84.33%	83.30%
+EDA-small	85.36%	(86.80%)	84.33%	**85.77%**	**84.12%**
+EDA-full	84.95%	87.01%	**85.98%**	85.77%	(81.86%)
+AEDA-small	86.80%	(84.95%)	(83.30%)	84.33%	(83.09%)
+AEDA-full	(84.74%)	**87.84%**	(83.09%)	85.36%	83.71%
+EASE-small	**87.22%**	(84.74%)	83.92%	84.54%	83.92%
+EASE-full	86.19%	87.63%	84.54%	84.95%	83.30%

Table 6: Comparing EASE, EDA, AEDA for the IMDB50K, S140 & FinPB datasets in low-resource scenarios by varying the number of augmented samples from small to full size. **Bold** suggests best performance across each column for each dataset, and parentheses suggest negative-impact on performance

	bert-base-cased	bert-base-uncased	distilbert-base-cased	distilbert-base-uncased	albert-base-v1
Large Movie Review Dataset					
Original	51.85%	54.19%	52.36%	55.97%	52.49%
EDA	54.75%	58.67%	54.41%	61.92%	50.36%
AEDA	(51.52%)	(53.18%)	52.72%	56.11%	(51.91%)
EASE	**64.18%**	**74.85%**	**69.27%**	**72.16%**	**58.66%**
Sentiment 140					
Original	40.24%	60.00%	40.44%	40.04%	40.04%
EDA	(37.22%)	60.00%	45.67%	(37.83%)	43.86%
AEDA	(37.63%)	60.62%	45.88%	(39.64%)	**44.67%**
EASE	**40.44%**	(59.79%)	**46.88%**	**41.05%**	41.85%
Financial Phrase Bank					
Original	59.38%	36.61%	59.38%	60.83%	59.38%
EDA	59.38%	38.83%	59.38%	61.03%	(58.14)%
AEDA	59.38%	**39.64%**	59.38%	61.65%	(59.18)%
EASE	59.38%	38.63%	**61.86%**	**61.86%**	**59.38%**

Table 7: Comparing EASE, EDA, AEDA for the IMDB50K, S140 & FinPB datasets in **extremely** low-resource scenarios (10 Samples) over 5 different models. **Bold** suggests best performance across each column for each dataset, and parentheses suggest negative-impact on performance

	bert-base-cased	bert-base-uncased	distilbert-base-cased	distilbert-base-uncased	albert-base-v1
Large Movie Review Dataset					
Original	86.92%	89.20%	85.36%	**88.38%**	84.86%
+EASE-small	87.74%	89.40%	85.47%	(87.72%)	86.46%
-A	(86.48%)	(88.93%)	(84.30%)	(87.16%)	86.02%
+EASE-full	**88.01%**	**89.60%**	**86.80%**	(87.72%)	**87.50%**
-A	87.13%	89.21%	85.41%	(87.36%)	86.01%
Sentiment 140					
Original	60.36%	60.56%	59.15%	57.75%	55.53%
+EASE-small	(59.76)%	**63.18%**	(58.75%)	58.55%	57.75%
-A	(58.75%)	60.56%	59.36%	59.36%	(54.53%)
+EASE-medium	60.97%	62.17%	**59.96%**	60.36%	**57.95%**
-A	(58.35)%	60.76%	(58.15%)	58.15%	(53.52%)
+EASE-full	**62.98%**	62.37%	**59.96%**	**61.97%**	56.34%
-A	(58.95)%	(59.15%)	(55.73%)	(56.94%)	(54.93%)

Table 8: EASE's performance after preserving labels (EASE vs EASE-A). **Bold** suggests the best performance across each column of each dataset, and parentheses suggest a negative impact on performance.