# Improving Imbalanced Learning by Pre-finetuning with Data Augmentation

**Yiwen Shi** YIWEN.SHI@DREXEL.EDU
*College of Computing and Informatics, Drexel University, Philadelphia, USA*
**Taha ValizadehAslani** TAHA.VALIZADEHASLANI@DREXEL.EDU
*College of Engineering, Drexel University, Philadelphia, USA*
**Jing Wang** JING.WANG1@FDA.HHS.GOV
**Ping Ren** PING.REN@FDA.HHS.GOV
**Yi Zhang** YI.ZHANG@FDA.HHS.GOV
**Meng Hu** MENG.HU@FDA.HHS.GOV
**Liang Zhao** LIANG.ZHAO@FDA.HHS.GOV
*Office of Research and Standards, Office of Generic Drugs, Center for Drug Evaluation and Research Food and Drug Administration, Silver Spring, USA*
**Hualou Liang** HUALOU.LIANG@DREXEL.EDU
*School of Biomedical Engineering, Science & Health Systems, Drexel University, Philadelphia, USA*

**Editor:** Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michał Woźniak and Shuo Wang.

## Abstract

Imbalanced data is ubiquitous in the real world, where there is an uneven distribution of classes in the datasets. Such class imbalance poses a major challenge for modern deep learning, even with the typical class-balanced approaches such as re-sampling and re-weighting. In this work, we introduced a simple training strategy, namely pre-finetuning, as a new intermediate training stage in between the pretrained model and finetuning. We leveraged the idea of data augmentation to learn an initial representation that better fits the imbalanced distribution of the domain task during the pre-finetuning stage. We tested our method on manually contrived imbalanced datasets (both two-class and multi-class) and the FDA drug labeling dataset for ADME (i.e., absorption, distribution, metabolism, and excretion) classification. We found that, compared with standard single-stage training (i.e., vanilla finetuning), our method consistently attains improved model performance by large margins. Our work demonstrated that pre-finetuning is a simple, yet effective, learning strategy for imbalanced data.

**Keywords:** Finetuning, Data Augmentation, BERT, Natural Language Processing

## 1. Introduction

Real-world data often exhibit long-tailed distributions with heavy class imbalance (e.g., Buda et al., 2018; Liu et al., 2019; Van Horn and Perona, 2017). When training machine learning models on imbalanced datasets, where certain classes contain many more samples than others, the models tend to learn better on the samples of majority classes but generalize poorly on minority classes (Branco et al., 2016; Buda et al., 2018; He and Garcia, 2009; Van Horn and Perona, 2017). Learning on such classes is of crucial importance in high-stakes settings such as the diagnosis of rare disease or unfairness for minority groups.

Data imbalance is particularly challenging for pretrained language models (LMs) in natural language processing (NLP, Brown et al., 2020; Devlin et al., 2019; Raffel et al., 2020, *inter alia*). Finetuning is the prevalent paradigm for using large pretrained LMs to perform downstream tasks. In this paradigm, a large LM such as BERT, which stands for Bidirectional Encoder Representations from Transformers, is trained on vast amounts of text, then finetuned on a specific downstream task (Figure 1A). Despite widely used, the finetuning may predispose the pretrained model to overfitting and poor generalization due to the large model and relatively small data samples in the downstream task. Owing to the paucity of samples, learning on the minority classes presents a persisting hindrance to improving task performance, even with the specialized class-balanced techniques such as re-sampling (Buda et al., 2018; He and Garcia, 2009; Van Horn and Perona, 2017) and re-weighting (Cao et al., 2019; Huang et al., 2016).

To tackle the above challenges, we proposed a simple training strategy, *pre-finetuning* (Figure 1B), as an additional training stage in between the pretrained model and finetuning. In our approach, we leveraged the idea of Data Augmentation (DA, Dhole et al., 2021; Feng et al., 2021; Wei and Zou, 2019) to produce a vast amount of augmented data that preserve a similar distribution as the original data for handling data scarcity and enhancing data diversity. Our pre-finetuning strategy (or pre-finetuning with DA) encourages the pretrained model to be better adapted to the target data, thereby leading to a good initialization for next stage of standard finetuning.

We first validated our proposed method on two manually created imbalanced benchmark datasets (both two-class and multi-class). We then applied our approach to a real-world FDA drug labeling dataset for enhancing product-specific guidance (PSGs) assessment (Shi et al., 2022). PSGs, recommended by the United States Food and Drug Administration (FDA), are instrumental to promote and guide generic drug product development. The FDA assessor needs to take extensive time and effort to manually retrieve supportive drug information of absorption, distribution, metabolism, and excretion (ADME) from the reference listed drug labeling for the PSG assessment. As a result, it is highly desirable to automate this process by developing a text classification model to automatically label ADME paragraphs with their semantic meaning. The dataset is by nature heavily class-imbalanced.

The contributions of our work are as follows:

- We introduced a novel method, pre-finetuning with data augmentation, to improve imbalanced learning before the vanilla finetuning takes place.
- We validated our approach on two benchmark datasets of text classification (both two-class and multi-class) and both achieved increased performance by improving the generalization of the minority classes.
- We presented a real-world application for ADME semantic labeling task, which gained superior performance when our approach was applied.

## 2. Related Work

**Pretraining and Finetuning Framework** Finetuning is the prevalent paradigm for using large LMs (Devlin et al., 2019; Radford et al., 2019) to perform downstream tasks. In this paradigm, a large LM such as BERT, is trained on vast amounts of text, then finetuned on a specific downstream task. Among different finetuning approaches, vanilla finetuning is perhaps the most popular approach, which finetunes some or all the layers of the LM and then adds one or two simple task-specific output layers (known as the classifier or the

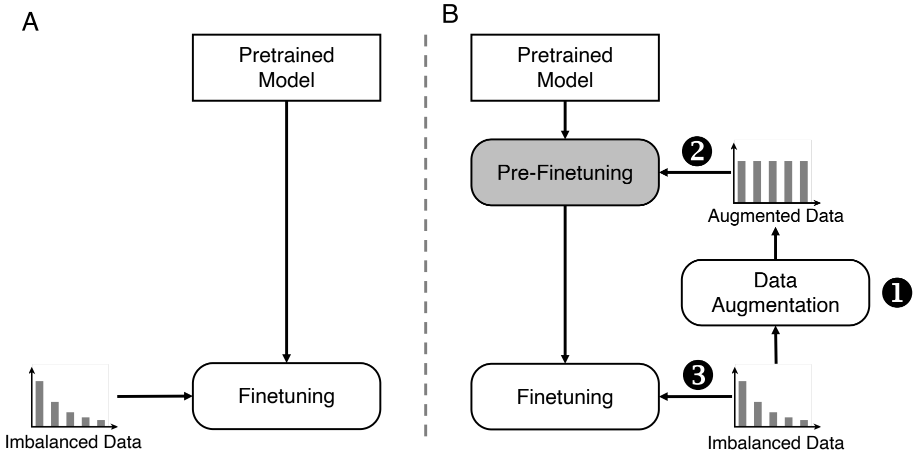

**Figure 1:** Schematics showing the vanilla finetuning pretrained model (A) and the pre-finetuning with data augmentation we proposed (B) for imbalanced data. A: The vanilla finetuning initializes with the general-purpose model pretrained with a large unlabeled corpus, and then performs a small amount of task-specific parameter updates. B: Our method introduces an additional *pre-finetuning* stage to adapt the pretrained model to have an initial representation of the target data before the standard finetuning takes place. It consists of three steps: (1) data augmentation, (2) finetuning with augmented data, and (3) standard finetuning.

head, Wolf et al., 2020). In this work, we built upon the vanilla finetuning by introducing the pre-finetuning to deal with imbalanced data.

**Imbalanced Learning** There is rich literature on learning imbalanced data. Re-sampling and re-weighting are two popular approaches (e.g., Geifman and El-Yaniv, 2017; Huang et al., 2016; Japkowicz and Stephen, 2002; Krawczyk, 2016; Van Horn and Perona, 2017; Yin et al., 2019). Re-sampling involves either over-sampling the minority classes or under-sampling the majority classes, or both. Such re-sampling incurs the cost of overfitting or losing the important information respectively. In addition, new samples can also be generated by interpolating neighboring samples or synthesizing for minority classes (Chawla et al., 2002; He et al., 2008). Re-weighting, on the other hand, is to modify the loss function to compensate for class imbalance by assigning weights to different samples according to the class distribution (Cao et al., 2019; Cui et al., 2019; Huang et al., 2016). There are different importance weighting schemes. A simple way is to assign sample weights proportionally to the inverse of the class frequency (Huang et al., 2016; Wang et al., 2017). Such a scheme tends to perform poorly when training on large-scale, imbalanced datasets (Huang et al., 2016). Instead of using the total number of samples present in each class, re-weighting loss by the inverse effective number of samples is introduced for better class-balance (Cui et al., 2019). Recently, a label-distribution-aware margin loss function is proposed to encourage larger margins for minority classes, leading to significantly improved performance on a variety of benchmark vision tasks (Cao et al., 2019).

Ensemble-based approach is also widely used for imbalanced learning. It is known to effectively deal with imbalanced data by merging the outputs of multiple classifiers (Chawla et al., 2003; Liu et al., 2020; Wang and Yao, 2009) or by combining individual classifiers in a multi-expert framework (Wang et al., 2022). These methods achieve the state-of-the-art performance mainly by reducing the model variance to obtain robust predictions (Krawczyk, 2016). However, many of them are direct combinations of a resampling/reweighting scheme and an ensemble learning framework (Chawla et al., 2003), which hence inherit the similar shortcomings of existing class-balancing strategies, such as the redundancy for easy samples by uniformly assigning experts to all classes (Wang et al., 2022). In most cases, it is difficult to obtain an appropriate cost matrix given by domain experts (Krawczyk et al., 2014).

**Data Augmentation (DA)** DA is a set of techniques for increasing training data diversity without directly collecting new data. It has proven widely effective in computer vision, albeit relatively challenging in NLP, due to the discrete nature of language data (Dhole et al., 2021; Feng et al., 2021). Broadly, there are three types of text data augmentation: rule-based, interpolation-based, and model-based. Rule-based methods manipulate words and phrases in a sentence to generate augmented text while ideally retaining the semantic meaning and labels of the original text. Easy Data Augmentation (EDA) is the representative in this category by employing a set of text editing techniques such as random insertion, deletion, replacement, and swap (Wei and Zou, 2019). They are easy to implement but usually offer unstable improvements due to the possibility that random perturbations can completely change the meanings of sentences (Niu and Bansal, 2018).

Interpolation-based methods generate new examples through interpolating operations over the original text directly (Chawla et al., 2002; He et al., 2008) or their latent states representations (Chen et al., 2020). Notably, SMOTE - Synthetic Minority Over-sampling Technique - generates synthetic samples for minority classes by linearly interpolating samples in the same class (Chawla et al., 2002). The model remains popular but is error-prone due to noise in the synthetic samples.

Model-based methods create augmented examples by leveraging either generative adversarial networks (Goodfellow et al., 2014) to add adversarial perturbations to the original data or the trained language models such as BERT to encode the class category along with its associated text to generate new samples with some modifications. The backtranslation is perhaps the most popular model-based method (Edunov et al., 2018; Sennrich et al., 2016) that translates sentences into certain intermediate languages and then back into the original language. This model-based approach requires significant training effort, but once the pretrained models are built, they are readily used to create novel and diverse data that might be unseen in the original dataset, leading to better performance.

## 3. Our Approach: Pre-finetuning with Data Augmentation

Our approach was both empirically and theoretically motivated. Recent empirical work showed that weighting has a significant effect *early in training*, and the impact of importance weighting diminishes over successive epochs of training (Byrd and Lipton, 2019). Theoretical analysis (Fang et al., 2021) predicted the emergence of *Minority Collapse* in imbalanced learning, i.e., the minority classes collapsed to a single vector *in the topmost layer*, which placed a fundamental limit on the model performance for the minority classes. As such, we proposed a novel early training stage between the pretrained model and finetuning, pre-finetuning, to adapt the pretrained model to have a rough-ready representation of the target data before the vanilla finetuning takes place.

Central to our approach is the innovative use of Data Augmentation (DA, Dhole et al., 2021; Feng et al., 2021; Wei and Zou, 2019) technique that allows us to produce a vast amount of augmented data that preserve a similar distribution as the original data to handle data scarcity and enhance data diversity. Our approach was henceforth referred to as *pre-finetuning with DA*. In this work, we used the back translation, a model-based approach for data augmentation due to its excellent performance (Edunov et al., 2018; Sennrich et al., 2016). In contrast to the vanilla finetuning (Figure 1A), our proposed approach, as schematically shown in Figure 1B, consists of three stages, as described below.

First, we used the back translation (Edunov et al., 2018; Sennrich et al., 2016) to generate a perturbed version of the training data while preserving the semantics of the original sentences. We leveraged two translation models, one translating the source text into a certain intermediate language and the other translating it back to the original language. For example, we can translate original sentences from English to German and then translate them back to get the paraphrases. Back-translated texts should maintain the semantics and basic syntactic structure of original texts. For data with a given label, back translation can generate a potentially infinite amount of new augmented data samples, thus can drastically avoid overfitting. In this work, we used German and Russian as the intermediate languages to enhance linguistic variety.

Second, we used the BERT as the pretrained model in this study. During the pre-finetuning stage, we froze every layer of BERT except the topmost layer and the classifier, which indicated we only tuned the topmost layer of BERT and the classifier with a larger learning rate on the augmented data from the data augmentation stage.

Third, we unfroze every layer and tuned the entire BERT model with the original imbalanced data as we do with the vanilla finetuning.

## 4. Experimental Setup

### 4.1. Datasets

We evaluated our pre-finetuning strategy on artificially created versions of two benchmark datasets: IMDB (an abbreviation of Internet Movie Database, Maas et al., 2011) and the 20 Newsgroups (Lang, 1995), and a real-world application of ADME semantic labeling (Shi et al., 2022).

**IMDB** This dataset consists of 50,000 movie reviews for binary sentiment classification. The number of positive and negative reviews is evenly distributed in the original dataset. We manually created an imbalanced training dataset by removing 90% of negative reviews. The testing dataset remained unchanged.

**The 20 Newsgroups** This dataset is a collection of approximately 20,000 newsgroup documents, partitioned almost evenly across 20 different newsgroups. We manually created the imbalanced version of the training set by reducing the number of training examples per class until a given imbalance ratio was reached and kept the test set unchanged. We defined the imbalance ratio $\rho$ as the ratio between sample sizes ($N_i$) of the least frequent class and the most frequent class, i.e., $\rho = min(N_i)/max(N_i)$. Two types of imbalance were considered to ensure that our method applicable to different settings. One is the step imbalance (Buda et al., 2018), where we artificially created a class-imbalanced training set with the imbalance ratio of 0.1. This is done by selecting ten classes, which sizes were between 591 to 600, then randomly sampled about 10% of records (60 records per class) for the remaining classes. Another is the long-tailed imbalance (Cui et al., 2019), where the data was also created by following the exponential decay distribution. Specifically, for each class, we randomly draw $max(N_i) \times \rho^{i/(n-1)}$ samples, where $max(N_i)$ is the maximum sample size for class $i$, $n$ is the number of the classes, $\rho$ is the imbalance ratio, which was set to 0.1. The distributions of the training datasets were showed in Figure **??** of the Appendix **??**.

**ADME Semantic Labeling** We applied our methods on the FDA drug labeling dataset for ADME classification. The FDA drug labeling dataset was retrieved from the DailyMed[1], which is a free drug information resource provided by the U.S. National Library of Medicine. The electronic drug labeling in DailyMed follows the Structured Product Labeling standard, which specifies various drug label sections by Logical Observation Identifiers Names and Codes (LOINC). ADME is a part of the pharmacokinetics section (LOINC code: 43682-4) of drug labeling. The rule-based method was used to extract 5,687 ADME paragraphs with explicit ADME titles and 5,367 paragraphs under other topics (e.g., "specific populations", "drug interaction studies", etc.) and hence labeled them as "Other" from the pharmacokinetics section in drug labeling. For details about data preparation, please refer to Shi et al. (2021). We randomly split 85% of the dataset for training and the rest 15% for testing, so both training and testing datasets remained class imbalanced. In addition to the hold-out method, we also performed 5-fold cross-validation (CV) on this dataset for additional check.

### 4.2. Implementation Details

We used PyTorch (Paszke et al., 2019) for all experiments. To generate the augmented data for pre-finetuning, we employed the back-translation method. Specifically, we first randomly sampled from the training set to get equal numbers of the labeled data for each class, then used *nlpaug*[2] to generate the corresponding augmented data by selecting German and Russian as intermediate languages for back translation. Hence, the input sentence was altered by back translation, while the class label was maintained. For example, for a sentence from IMDB: *"You'd better choose Paul Verhoeven's even if you have watched it."*, the augmented texts through German and Russian were, respectively, *"You* ***should*** *choose Paul Verhoeven's, even if you* ***saw*** *it."* and *"You'd better* ***pick*** *Paul Verhoeven, even if you* ***were watching*** *him."*

In our implementation, we used the BERT base model, *bert-base-uncased*[3], as the pretrained model due to its use in NLP predominantly[4]. The batch size and maximum sequence length remained the same as the vanilla finetuning. With the augmented data, we used 1 epoch with a larger learning rate of 1e-4 to only tune the topmost layer and the head in the pre-finetuning stage, which enables the model to learn quickly and preserve the pretrained features. The hyperparameters used for finetuning in the second stage kept the same as the vanilla finetuning for all the datasets. We provided further details of the hyperparameters

[1] https://dailymed.nlm.nih.gov/dailymed

[2] https://github.com/makcedward/nlpaug

[3] https://huggingface.co/bert-base-uncased

[4] We note that our method does not depend on the BERT base per se; other pretrained models such as the BERT large can also be used. However, it is generally observed that larger models have higher accuracy.

used for each dataset in Table **??** of Appendix **??**. All the experiments were run on either a single Nvidia Tesla P100-PCIE-16GB or Nvidia Tesla V100-SXM2-32GB.

### 4.3. Baselines

We benchmarked our proposed method against the vanilla finetuning as the our primary baseline. We used the *bert-base-uncased* model which was pretrained and initialized with the parameters released by (Devlin et al., 2019), which can be accessed from Huggingface (Wolf et al., 2020). The model configuration we used was consistent with the recommendations in the original release. We grid searched a batch size of {8, 16, 32}, and a learning rate of {5e-6, 1e-5, 3e-5, 5e-5}, with the optimal hyperparameters for each dataset shown in Table **??** of the Appendix **??**.

For completeness, we also included results for additional baselines: (1) Finetuning via oversampling, which is perhaps the most popular sampling scheme to oversample training examples from the minority classes (Buda et al., 2018; Cui et al., 2019; Cao et al., 2019), and (2) LP-FT, a two-step strategy of linear probing (LP or head tuning) followed by full finetuning (FT) which has been shown to achieve competitive results (Levine et al., 2016; Kanavati and Tsuneki, 2021; Kumar et al., 2022).

### 4.4. Evaluation Metrics

We used the F1-score as our primary metric to assess the model performance as it is sensitive to data distribution. Since we dealt with both two-class and multi-class problems, we reported both the overall F1 score (micro-F1) and per-class F1 score to quantify the generalization performance of both majority and minority classes in the data. When we reported the F1 score for multi-class classification, we computed the micro-F1 *on the non-majority classes* to remove the dominance of the majority class, hence it is not biased toward the majority class. Note that we mainly used the hold-out method for model evaluation. While the hold-out method is particularly attractive in deep learning where model training is expensive, the results can depend on a particular random choice of the data set. To reduce the potential sampling bias, we reported all the results that were based on the average of the five independent runs, each with different seed. Additionally, we performed the stratified 5-fold CV on the ADME dataset to check the robustness of our results since the CV is a common, albeit costly, practice to obtain better estimates. The uncertainty in estimates was represented by error bar based on the standard error of the mean (SEM).

## 5. Results

### 5.1. Imbalanced Benchmark Datasets: IMDB and the 20 Newsgroups

**IMDB** Before we showed our main results, we performed a quick sanity check on the quality of the data augmentation technique we used. To assess to what extent the augmented data preserved a similar distribution as the original, we compared the distributions of the text embedding for the augmented data and for the original data. To do so, we first generated the augmented data for 2,500 samples from the training dataset (1,250 for each class), and similarly set aside 2,500 hold-out original data samples for validation. We then obtained the embeddings for both the augmented data and the hold-out data through the finetuned BERT models. Finally, we compared the distributions of their embeddings using t-SNE visualization (Van der Maaten and Hinton, 2008).

The results were displayed in Figure 2. We can observe that the distribution of the augmented data very well matched with that of the original hold-out dataset in each class. The comparison indicated that the augmented data indeed possess a similar distribution as the original data. Note that the original data used for data augmentation and the hold-out data were from the same data distribution, but independent of each other to avoid potential confounding.

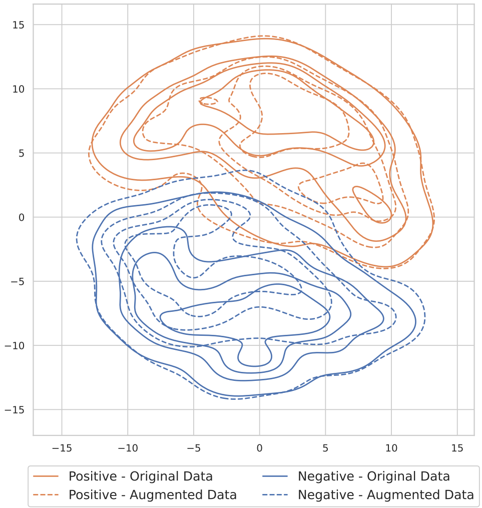

**Figure 2:** The t-SNE visualization of IMDB original data and augmented data via back translation from German.

We reported the F1 scores and per-class F1 scores for the IMDB dataset in Table 1 for both our method and the baselines: the vanilla finetuning, finetuning via oversampling and linear probing-then-full finetuning (LP-FT). Our method showed the best performance of both the F1 and per-class F1 when compared to all three baselines. The negative class was the minority class, which achieved more improvement than the majority class. This suggested that, with the pre-finetuning, our method was successful in regularizing minority classes more strongly.

**Table 1:** F1 and per-class F1 (SEM) comparisons on the IMDB dataset between our method (pre-finetuning with DA) and three baselines (the vanilla finetuning, finetuning via oversampling and LP-FT). The Negative class is the minority class, which has a larger improvement than the majority class (Positive class). Note that the testing dataset is *balanced*.

| | F1 | Per-class F1 | |
|---|---|---|---|
| | | Negative | Positive |
| Vanilla Finetuning | 0.8637 (0.0031) | 0.8447 (0.0037) | 0.8786 (0.0021) |
| Finetuning via Oversampling | 0.8583 (0.0019) | 0.8373 (0.0023) | 0.8745 (0.0013) |
| LP-FT | 0.8639 (0.0028) | 0.8446 (0.0035) | 0.8789 (0.0019) |
| Pre-finetuning with DA | 0.8651 (0.0026) | 0.8463 (0.0032) | 0.8799 (0.0018) |

**The 20 Newsgroups** We reported the per-class F1 scores in Figure 3. We observed that both our pre-finetuning with DA and the baseline method (the vanilla finetuning) showed the minority classes had much lower F1 than the majority classes, but our method exhibited better generalization on minority classes with a significant improvement of F1 than the majority classes. Overall, the F1 score increased from 0.6433 for the vanilla finetuning to 0.6781 for our method.

To ensure our method applicable to different settings, we tested our pre-finetuning strategy with a long-tailed imbalance training dataset on the 20 newsgroups benchmark (The class distribution of the training dataset was showed in Figure **??**C in the Appendix **??**).

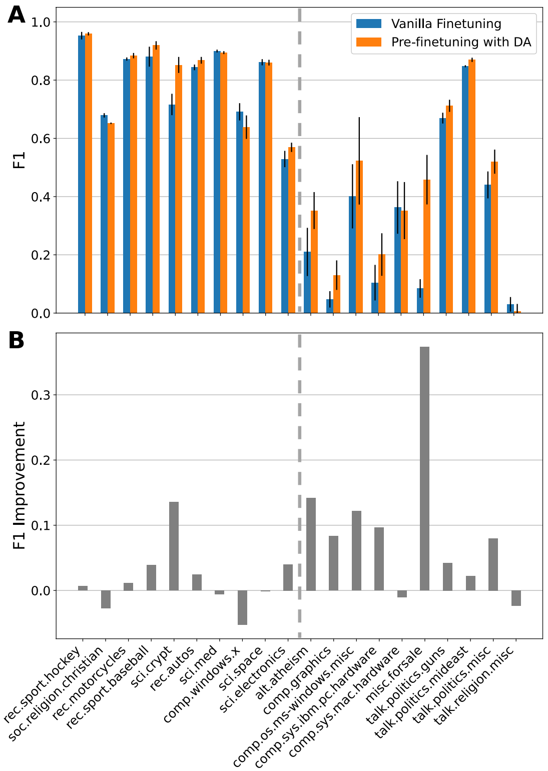

**Figure 3:** Comparison of per-class F1 scores between the vanilla finetuning and our pre-finetuning with DA trained on the 20 Newsgroups data with *step-imbalance*. The classes on the left of the dashed line are the majority classes, and the ones on the right are the minority classes. A: Comparison of the F1 scores between the vanilla fine-tuning and our method. The error bar: SEM. B: The absolute improvements as the difference between the vanilla method and our pre-finetuning with DA, where we see the substantial improvements in most of minority classes.

Our method with the F1 score of 0.7571 outperformed the baseline method with the F1 of 0.7137. The per-class F1 scores of minority classes gained more improvements than the majority classes, which had a similar trend as those with the step-imbalanced dataset (Figure **??** in the Appendix **??**). The results show our method can be used to adapt to different imbalanced types of the training dataset.

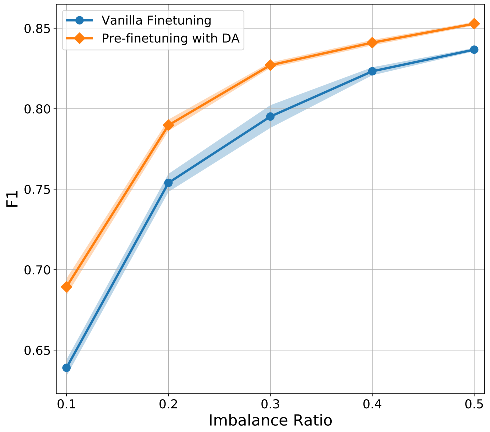

**Figure 4:** Comparison between two methods (the vanilla finetuning and our pre-finetuning with DA) on the 20 Newsgroups benchmarks by showing that the F1 score changes as a function of the imbalance ratio. The shaded represents the SEM.

In our method, we have used the back translation (Edunov et al., 2018; Sennrich et al., 2016) in the data augmentation. To examine if other data augmentation techniques can also be used, we instead experimented with EDA (Wei and Zou, 2019) method in our pre-finetuning strategy. The results showed that with EDA, our method worked equally well by improving the overall F1 and per-class F1 (Figure **??** in the Appendix **??**).

**Impact of imbalance ratio** To assess the impact of imbalance ratio on the model performance, we systematically vary the imbalance ratio from 0.1 to 0.5, with increments of 0.1. Figure 4 showed that the overall F1 performance improved when the imbalance ratio increased, regardless of what method was used. However, our pre-finetuning with DA outperformed the baseline method across all the different imbalance ratios. In addition, we observed that the performance gain of F1 score decreased when the imbalance ratio increased.

### 5.2. Real-world Application: ADME Semantic Labeling

We applied our method to the FDA drug labeling dataset for ADME classification, which was used as an example to demonstrate our method could be applied to a real-world problem. We first reported the results based on the hold-out method. Figure 5A showed the class distribution of the ADME training data, which was inherently imbalanced. Compared to the vanilla finetuning, our pre-finetuning with DA improved the F1 score from 0.8936 to 0.9054. The per-class F1 scores were provided in Figure 5B with the absolute improvements shown in Figure 5C. We observed from these data that most of the F1 improvements in minority classes (e.g. ADME) were much more substantial than the majority class (e.g. Other). Moreover, the per-class F1 improvement increased while the sample size per class decreased. We further performed the 5-fold CV on this dataset, and found that our method still improved the overall F1 score from 0.8993 to 0.9070 when compared to the vanilla finetuning. The per-class F1 scores obtained by the 5-fold CV were provided in Appendix **??**.

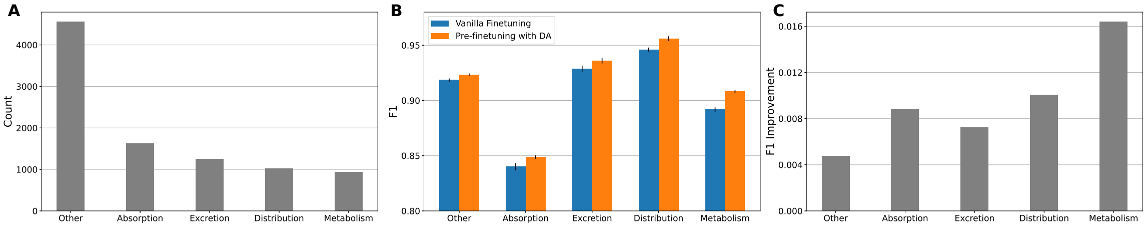

**Figure 5:** Per-class F1 scores and improvements on ADME dataset. A: The class distribution of the training set. B: The per-class F1-score comparison between the two methods (The vanilla method and our pre-finetuning with DA). C: Improvement in per-class F1 of our method over the vanilla method. The F1-score increase is observed in all classes with our method. The improvement becomes larger in the less frequent classes and smaller in the more frequent classes. Error bars: SEM.

## 6. Discussion and Conclusion

Imbalanced data is inherent in many real-world problems. It occurs when there are one or more classes (majority classes) that are more frequently occurring than the other classes (minority classes). This problem plagues most machine learning algorithms that assume the classes are roughly equal in size. When training a model on an imbalanced dataset, the model becomes biased toward the majority classes, and hence generalizes poorly on minority classes (Buda et al., 2018; He and Garcia, 2009; Van Horn and Perona, 2017).

In this work, we showed that pre-finetuning is a simple, yet effective, learning strategy for imbalanced data. The core idea was to add the pre-finetuning as a new intermediate training stage in between the pretrained model and finetuning. Specifically, we leveraged the augmented data to learn an initial representation of the imbalanced data. With this additional training stage, we can further increase the similarity between the general domain and the target domain, which enabled the model to fit the imbalanced distribution of the downstream task potentially better. We showed that standard pretrained representations, when further refined with pre-finetuning, consistently improved performance on downstream tasks, as evaluated on two manually created imbalanced datasets and an FDA drug labeling dataset for ADME semantic labeling.

With data augmentation, we can generate potentially an unlimited size of synthetic data in the vicinity of the original data space. Ideally, the augmented data should possess a similar distribution as the original data. The distribution of the augmented data should neither be too similar nor too different from the original, which may respectively lead to model overfitting or poor performance through training on examples not representative of the given domain. Therefore, we noted that augmentation does not always improve performance. It is still challenging to determine under what conditions DA approaches are effective as there remains a lack of theoretical understanding as to why DA works (Dao et al., 2019).

We note that our pre-finetuning strategy is a rather general approach to improving downstream performance. While the method we proposed in this work is designed for learning imbalanced data, the idea can also be applied to balanced data as well as other domains of machine learning such as computer vision. Although our work focuses on the imbalanced classification, extending our investigation to imbalanced regression (Steininger et al., 2021; Torgo et al., 2013; Yang et al., 2021) is also of interest. In our approach, we have used the augmented data for pre-finetuning. We note other strategies such as re-weighting can be also exploited from the model perspective (ValizadehAslani et al., 2022). Prior work (Levine et al., 2016; Kanavati and Tsuneki, 2021; Kumar et al., 2022) has shown that LP-FT, a two-step strategy of linear probing then full fine-tuning, provides better results than either does alone. Our work, albeit conceptually similar, has several major differences: (1) we aim at the class imbalance problem; (2) we use the data augmentation for pre-finetuning; (3) we additionally finetune the final layer of the model, as motivated by theoretical analysis (Fang et al., 2021); and (4) we focus on the NLP tasks.

In the current setting, we have limited the model pre-finetuned to only the topmost layer, which was mainly determined by the theoretical analysis (Fang et al., 2021). In practice, it is plausible that the last few layers, rather than only the topmost layer, could be pre-finetuned for improved performance when it comes to different datasets. When using our method to learn imbalanced data, two important hyperparameters need to be determined, that is, the number of epochs and the learning rate for each stage. We observe empirically overall good performance when we use only 1 epoch and a relatively large learning rate (1e-4) for the pre-finetuning in the first stage. This enables the model to roughly adapt to the distribution of the target domain to learn a good initial representation. For the finetuning stage, we followed the standard BERT finetuning parameters and trained the model using a 1e-5 learning rate for 3 to 5 epochs. The learning rate in the finetuning stage was relatively small compared to the pre-finetuning stage so the second stage did not move the weights very far. It is conceivable that the optimal parameters are application dependent. Further research is needed to determine the optimal number of epochs for pre-finetuning.

When reporting the model performance, it has been an on-going debate about how to choose an appropriate metric for imbalanced data (Japkowicz and Shah, 2011), particularly for multi-class classification problems. In this work, we have mainly used the F1-score to assess the model performance, though other measures can be more appropriate in different domains such as medical diagnosis. As the F1-score is more sensitive to data distribution, it is a suitable measure for classification problems on imbalanced datasets. As we work with both two-class and multi-class problems, we report both the overall F1 score (micro-F1) and per-class F1 score to quantify the generalization performance of both majority and minority classes for the benchmark datasets. We also test the per-class top-1 error, which obtains similarly consistent performance. When reporting the model performance for imbalanced data, we deem that it is important to explicitly document what evaluation metrics are exactly used.

Although the testing data in the two benchmark datasets in our experiments followed the relative uniform distribution, the testing dataset could be imbalanced naturally, such as the ADME semantic labeling task. When test distribution is imbalanced, we observed that pre-finetuning works equally well. Specifically, we artificially generated the imbalanced testing datasets that have different class distributions than the training datasets for both IMDB and the 20 Newsgroups. Table **??** in the Appendix **??** showed results for IMDB test data of step-imbalanced distribution, and Figure **??** in the Appendix **??** showed the results for the 20 Newsgroups test data of different class imbalance distributions (both the step imbalance and long-tailed imbalance). For both benchmarks, we found similar improvements in micro-F1 and per-class F1 as the balanced testing datasets. These results suggested that our method can also be used broadly where the test label distribution was not necessarily uniform.

## Acknowledgments

This work was partly supported by the FDA Contract #: 75F40119C10106. Work reported here partly was run on hardware supported by Drexel's University Research Computing Facility.

## Disclaimer

The opinions expressed in this article are the author's own and do not reflect the view of the Food and Drug Administration, the Department of Health and Human Services, or the United States government.

## References

P. Branco, L. Torgo, and R. P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2):1–50, 2016.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

M. Buda, A. Maki, and M. A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

J. Byrd and Z. Lipton. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pages 872–881. PMLR, 2019.

K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery*, pages 107–119. Springer, 2003.

J. Chen, Z. Yang, and D. Yang. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *arXiv preprint arXiv:2004.12239*, 2020.

Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

T. Dao, A. Gu, A. Ratner, V. Smith, C. De Sa, and C. Ré. A kernel theory of modern data augmentation. In *International Conference on Machine Learning*, pages 1528–1537. PMLR, 2019.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

K. D. Dhole, V. Gangal, S. Gehrmann, A. Gupta, Z. Li, S. Mahamood, A. Mahendiran, S. Mille, A. Srivastava, S. Tan, et al. Nl-augmenter: A framework for task-sensitive natural language augmentation. *arXiv preprint arXiv:2112.02721*, 2021.

S. Edunov, M. Ott, M. Auli, and D. Grangier. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*, 2018.

C. Fang, H. He, Q. Long, and W. J. Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43), 2021.

S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*, 2021.

Y. Geifman and R. El-Yaniv. Deep active learning over the long tail. *arXiv preprint arXiv:1711.00941*, 2017.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

H. He, Y. Bai, E. A. Garcia, and S. Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE, 2008.

C. Huang, Y. Li, C. C. Loy, and X. Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016.

N. Japkowicz and M. Shah. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.

N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

F. Kanavati and M. Tsuneki. Partial transfusion: on the expressive influence of trainable batch norm parameters for transfer learning. In *Medical Imaging with Deep Learning*, pages 338–353. PMLR, 2021.

B. Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

B. Krawczyk, M. Woźniak, and G. Schaefer. Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing*, 14:554–562, 2014.

A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang. Fine-tuning can distort pre-trained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

K. Lang. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pages 331–339. Elsevier, 1995.

S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019.

Z. Liu, W. Cao, Z. Gao, J. Bian, H. Chen, Y. Chang, and T.-Y. Liu. Self-paced ensemble for highly imbalanced massive data classification. In *2020 IEEE 36th international conference on data engineering (ICDE)*, pages 841–852. IEEE, 2020.

A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. 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, 2011.

T. Niu and M. Bansal. Adversarial over-sensitivity and over-stability strategies for dialogue models. *arXiv preprint arXiv:1809.02079*, 2018.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2020.

R. Sennrich, B. Haddow, and A. Birch. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*, 2016.

Y. Shi, P. Ren, Y. Zhang, X. Gong, M. Hu, and H. Liang. Information extraction from FDA drug labeling to enhance Product-Specific Guidance assessment using natural language processing. *Frontiers in Research Metrics and Analytics*, 6, 2021.

Y. Shi, J. Wang, P. Ren, T. ValizadehAslani, Y. Zhang, M. Hu, and H. Liang. Fine-tuning BERT for automatic ADME semantic labeling in FDA drug labeling to enhance Product-Specific Guidance assessment. *arXiv preprint arXiv:2207.12376*, 2022.

M. Steininger, K. Kobs, P. Davidson, A. Krause, and A. Hotho. Density-based weighting for imbalanced regression. *Machine Learning*, 110(8):2187–2211, 2021.

L. Torgo, R. P. Ribeiro, B. Pfahringer, and P. Branco. Smote for regression. In *Portuguese conference on artificial intelligence*, pages 378–389. Springer, 2013.

T. ValizadehAslani, Y. Shi, J. Wang, P. Ren, Y. Zhang, M. Hu, L. Zhao, and H. Liang. Two-stage fine-tuning: A novel strategy for learning class-imbalanced data. *arXiv preprint arXiv:2207.10858*, 2022.

L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

G. Van Horn and P. Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.

S. Wang and X. Yao. Diversity analysis on imbalanced data sets by using ensemble models. In *2009 IEEE symposium on computational intelligence and data mining*, pages 324–331. IEEE, 2009.

X. Wang, L. Lian, Z. Miao, Z. Liu, and S. X. Yu. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2022.

Y.-X. Wang, D. Ramanan, and M. Hebert. Learning to model the tail. *Advances in Neural Information Processing Systems*, 30, 2017.

J. Wei and K. Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2020.

Y. Yang, K. Zha, Y. Chen, H. Wang, and D. Katabi. Delving into deep imbalanced regression. In *International Conference on Machine Learning*, pages 11842–11851. PMLR, 2021.

X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Feature transfer learning for deep face recognition with under-represented data. *arXiv preprint arXiv:1803.09014*, 2019.