WeLT: Improving Biomedical Fine-tuned Pre-trained Language Models with Cost-sensitive Learning

Ghadeer Mobasher [1, 2], **Wolfgang Müller**[1], **Olga Krebs** [1], **and Michael Gertz**[2]

[1]Scientific Databases and Visualization, Heidelberg Institute for Theoretical Studies (HITS gGmbH), Germany
[2]Institute of Computer Science, Heidelberg University, Germany
{ghadeer.mobasher,wolfgang.mueller,olga.krebs}@h-its.org
gertz@informatik.uni-heidelberg.de

Abstract

Fine-tuning biomedical pre-trained language models (BioPLMs) such as BioBERT has become a common practice dominating leaderboards across various natural language processing tasks. Despite their success and wide adoption, prevailing fine-tuning approaches for named entity recognition (NER) naively train BioPLMs on targeted datasets without considering class distributions. This is problematic especially when dealing with imbalanced biomedical gold-standard datasets for NER in which most biomedical entities are underrepresented. In this paper, we address the class imbalance problem and propose **WeLT**, a cost-sensitive fine-tuning approach based on new rescaled class weights for the task of biomedical NER. We evaluate **WeLT**'s fine-tuning performance on mixed-domain and domain-specific BioPLMs using eight biomedical gold-standard datasets. We compare our approach against vanilla fine-tuning and three other existing reweighting schemes. Our results show the positive impact of handling the class imbalance problem. WeLT outperforms all the vanilla fine-tuned models. Furthermore, our method demonstrates advantages over other existing weighting schemes in most experiments.

1 Introduction

Fine-tuning biomedical pre-trained language models has become a mainstream practice dominating leaderboards across various biomedical NLP tasks. Despite major advancements and wide usage, recent work reported fine-tuning instabilities for general-domain NLP tasks (Devlin et al., 2019; Lee et al., 2019). Lee and colleagues reported that small training datasets (i.e., less than 10,000 examples) is one of the potential reasons for fine-tuning instabilities. Moreover, a few studies point out the positive impact of handling the class imbalance before fine-tuning (ValizadehAslani et al., 2022). Most of the real-world biomedical datasets are highly imbalanced (Akkasi et al., 2018). Nevertheless the impact of class imbalance before fine-tuning biomedical datasets is often explored, especially not for named entity recognition.

The commonly used tagging scheme for biomedical named entity recognition (BioNER) is the Inside–outside–beginning (IOB) format (Shen and Sarkar, 2005). It consists of three classes: (a) the **B** tag represents the beginning or first token of a biomedical entity, (b) the **I** tag denotes the continuation of the first token as an inside biomedical entity, and (c) the **O** tag represents a token that is not part of a biomedical entity. Thus **B** and **I** classes are the positive samples, while the **O** class is the negative sample. Table 1 shows the imbalanced nature of BioNER corpora for multiple entity types, including chemical, disease, gene, and species entities. It is clear to conclude that biomedical ground-truth training datasets are highly skewed. Such high data imbalance and scarcity make many BioNER models biased towards the **O** class, thus, they often misclassify entities (B and I classes).

Dataset	O	B	I
NCBI (Doğan et al., 2014)	74.44	12.67	12.89
BC5CDR-Disease (Li et al., 2016)	93.99	3.54	2.47
BC5CDR-Chemical (Li et al., 2016)	93.99	4.40	1.61
BC4CHEMD (Krallinger et al., 2015)	92.69	3.30	4.01
BC2GM (Smith et al., 2008)	89.50	4.28	6.22
BioRED-Chem (Luo et al., 2022)	96.72	2.34	0.94
BioRED-Disease (Luo et al., 2022)	94.78	3.00	2.22
Linnaeus (Gerner et al., 2010)	98.84	0.75	0.41

Table 1: Class distribution percentage for biomedical ground-truth training datasets.

Since these gold-standard biomedical datasets are curated by domain experts, we avoided using traditional resampling approaches, as there might be a possibility of information loss, and duplicating training examples leads to poor performance of language models (Lee et al., 2021). On the other hand, fine-tuning BioPLMs (Alrowili and Shanker, 2021) naively on highly skewed datasets

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 427–438

is problematic, since it can lead to a bias in the trained models, which can negatively affect the performance. Therefore, we investigate the impact of handling the class imbalances while fine-tuning. We develop a weighted loss trainer (WeLT) that addresses class imbalances by introducing coefficients that penalize majority classes and give more weight to the rare ones before fine-tuning.

To evaluate our approach, we conduct extensive BioNER experiments on eight biomedical gold-standard datasets. We fine-tune different variants of biomedical BERT models (mixed/continual training and domain-specific) and BioELECTRA. We do not only compare WeLT against vanilla fine-tuning approach (i.e., it assumes that the misclassification errors are equal for all class labels), but also with three other existing comparable weighting schemes. WeLT and other weighting schemes outperform vanilla fine-tuned models. This proves the hypothesis that class imbalance has a negative impact on fine-tuning. We suggest that class imbalance should be considered as an additional factor for fine-tuning.

In summary, our main contributions are as follows:

1. We propose WeLT, a class-balanced re-weighting scheme added to a trainer's loss function before fine-tuning models.

2. We compare WeLT to vanilla fine-tuning approach, and existing cost-sensitive class weighting methods, including Inverse Number of Samples, Inverse of Square Root of Number of Samples, and Effective Number of Samples (Suri, 2022; Cui et al., 2019). We conduct our experiments on several transformers such as BERT and ELECTRA.

3. We publicly release the code and hyper-parameters to reproduce our research results[1]. We also release all the fine-tuned models on the Hugging Face Hub[2] (Wolf et al., 2020).

2 Related Work

Large pre-trained language models have been the state-of-the-art across a variety of tasks in natural language processing (Devlin et al., 2019; Clark et al., 2020). However, direct application of these models to biomedical text mining applications yields unsatisfactory results, since the word distributions of general domain and biomedical corpora are very diverse (Lee et al., 2020). Therefore, Lee and colleagues adapted BERT for biomedical applications and pre-trained BERT on PubMed abstracts and PubMed Central full-text articles. However, BioBERT is a mixed-domain pre-training approach, as authors used the original vocabulary of BERT (i.e., Wikipedia and BookCorpus). BlueBERT (Peng et al., 2019) is pre-trained on PubMed abstracts and clinical notes. SciBERT (Beltagy et al., 2019) generates the vocabulary and pretrains from scratch, using biomedical and computer science literature. Subsequent work has been devoted to train the model from scratch on a biomedical vocabulary only, such as PubMedBERT (Gu et al., 2021). Their results show that domain-specific pretraining from scratch can substantially outperform the traditional mixed-domain approach. BioELECTRA is pre-trained on PubMed and PubMed Central (Kanakarajan et al., 2021). Biomedical BERT variants and BioELECTRA require minimal modification of the architecture. Leaman and colleagues developed various ensemble methods for the chemical identification task (Leaman et al., 2023) mostly focusing on fine-tuning Biomedical BERT variants. The fine-tuning instability of BERT has been pointed out in various studies. The observed instabilities are identified due to the following potential reasons: (a) catastrophic forgetting, (b) small size of the fine-tuning datasets (Devlin et al., 2019; Lee et al., 2019; Dodge et al., 2020), and (c) optimization difficulties early in training (vanishing gradients) (Mosbach et al., 2020). One of the possible reasons for poor fine-tuning results is using an imbalanced dataset even if the training dataset is large. Most of the biomedical datasets are skewed (Zhao et al., 2018).

Most existing algorithms for learning imbalanced datasets can be divided in to two categories: (a) re-sampling, and (b) re-weighting. In the biomedical domain, Akkasi and colleagues directly investigated the class imbalance problem for biomedical datasets. They propose a balanced undersampling approach for sequence data and enhance the classification performance by systematically removing negative samples from training data (Akkasi et al., 2018). Re-weighting approaches allow minority samples to contribute more to the loss function. One of the earliest approaches introduced a factor that can be multiplied by a certain thresh-

[1] https://github.com/mobashgr/WeLT
[2] https://huggingface.co/mobashgr

old, resulting in higher weights for misclassification of minority classes (Elkan, 2001). However, this factor assumes that one deals with a binary classification problem. Recent work addresses class imbalance when applying BERT for sentence classification. For example, Madabushi and colleagues (Madabushi et al., 2020) applied cost-weighting for a binary classification problem on which the exact weight is related to the dissimilarity of training, development, and test datasets. They report that the weights are obtained experimentally through a hyper-parameter search. Weighting by inverse class frequency or a smoothed version of inverse square root of class frequency are often adopted (Suri, 2022). Cui and colleagues proposed a re-weighting loss by using the inverse effective number of samples for addressing class imbalance (Cui et al., 2019).

Handling class imbalance for different NLP downstream tasks is challenging, due to the following reasons:

- Most of the training data are part of gold standard datasets that are annotated and curated by domain experts. The human annotation and curation of biomedical data are tedious tasks that require manpower and time, therefore, there are only limited biomedical gold-standard datasets. Consequently, undersampling is not the appropriate solution to address class imbalance.
- Even if the training data is not a gold standard, transformers are powerful enough to memorize duplicate instances. Moreover, recent research proves that de-duplicating training data makes the language models better and allows faster training (Lee et al., 2021).
- Removing some instances from the majority classes can result in information loss.
- Running multiple experiments to acquire new class weights by applying cost-learning approaches is a tedious task that does not provide a generic method that can be used to acquire new weights for both majority and minority classes.

Based on the extensive literature survey related to the existing class imbalance research, we conclude that the cost-sensitive learning approaches for biomedical BERT-based models have been neglected for BioNER.

3 WeLT - Weighted Loss Trainer

Vanilla fine-tuned models assume that misclassification errors and cost, respectively, contribute equally to all class labels. However, in many real-world applications, such as BioNER, the differences between different misclassification errors can be quite large and thus critical. In other words, the error costs for rare classes in a trainer's loss function should be higher. This can be achieved by re-weighting the instances of each class according to their misclassification costs. For this, we proposed WeLT, a cost-sensitive fine-tuning approach for BioNER as depicted in Figure 1.

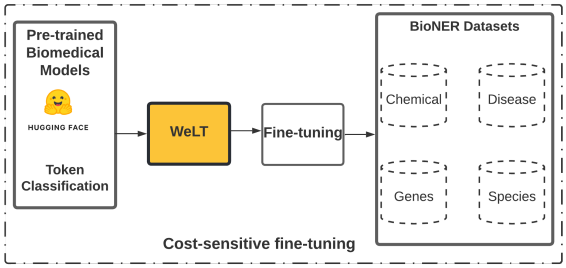

Figure 1: WeLT: A class-balanced re-weighting loss function applied before fine-tuning on BioNER targeted datasets.

WeLT sets a higher class weight for the minority classes and a lower weight for the majority classes. Based on the context of our work with the IOB tagging scheme, **O** is the majority class. While **B** and **I** are the minority classes as shown in Figure 2. We present a generic way of calculating the re-scaled class weights for the IOB tags without having hyper-parameter search factors. The new class weights are calculated using the inverse ratio of each corresponding class distribution over the total class distributions of the training datasets, as shown in Equation (1).

$$w_x = 1 - \frac{I_x}{I_t}, \text{ with } x \in \{o, b, i\} \quad (1)$$

We illustrate the computation based on the NCBI dataset. Let TD_{ncbi} be the training data of NCBI, consisting of I_t = **5,868,591** instances of the three classes o, b, and i. The number of instances per class in TD_{ncbi} is denoted I_o, I_b and I_i for the **O**, **B**, and **I** classes, respectively.

The three class instances are as follows:

- I_o = **4,262,718**
- I_b = **389,892**
- I_i = **1,215,981**

The new weight for each of the three classes is computed using Equation (1), leading to the following values :

- $w_o \approx$ **0.27**
- $w_b \approx$ **0.79**
- $w_i \approx$ **0.93**

Let $\vec{WV}$ be the vector that has three elements, i.e. $WV = \left(\begin{bmatrix} w_o \\ w_b \\ w_i \end{bmatrix} \right)$. Then, $\vec{WV}$ is normalized using Softmax, as shown in Equation (2).

$$\sigma(\vec{WV})i = \frac{e^{WV_i}}{\sum_{c=1}^{t} e^{WV_c}} \quad (2)$$

where $\vec{WV}$ is the input vector to Softmax. WV_i are the elements of $\vec{WV}$, so WV_i of TD_{ncbi} contains three elements.

The standard exponential function, denoted e^{WV_i}, is applied to each element of the input vector $\vec{WV}$.

The normalization term, denoted $\sum_{c=1}^{t} e^{WV_c}$, ensures that all the output values of the function will sum to 1, where $t = 3$.

Thus, the values for normalized vector of TD_{ncbi} denoted, $\sigma(\vec{WV})$, are $\approx \left(\begin{bmatrix} \mathbf{0.2167} \\ \mathbf{0.4192} \\ \mathbf{0.3641} \end{bmatrix} \right)$

As previously mentioned vanilla fine-tuned trainers use cross-entropy loss. Cross-entropy loss as shown in Equation (3) minimizes the training error by assuming that individual samples and classes are equally important as if the class frequencies are sufficiently balanced.

$$L_{ce} = -\sum_{i=1}^{c} y_i \; log \; \hat{y}_i \quad (3)$$

where $c = 3$ represents the different classes to be predicted, $y_i \in [0, 1]$ denotes the ground truth class, and $\hat{y}_i \in [0, 1]$ is the model's estimated probability for each class with ground truth. Since the biomedical gold standard training datasets are highly imbalanced, we use the weighted cross-entropy loss.

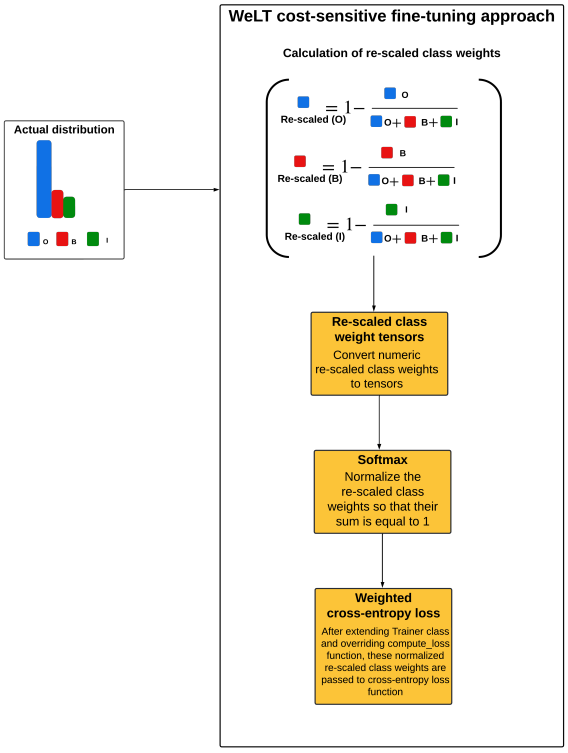

Figure 2: Calculation of normalized re-scaled class weights in WeLT.

The normalized vector $\sigma(\vec{WV})i$ is passed to the weighted cross-entropy loss function, as illustrated in Equation (4), after extending the class trainer and overriding the *compute_loss* function (Paszke et al., 2017).

$$L_{wce} = \sigma(\vec{WV})i \; L_{ce} \quad (4)$$

Subsequently, the models are fine-tuned using WeLT with the exact training cost as the vanilla fine-tuned approach.

4 Experiments

We empirically demonstrate the performance of our method through various experiments on eight biomedical gold-standard datasets focusing on the BioNER task. We evaluate WeLT on both mixed-domain and domain-specific pre-trained models. We compared WeLT to their corresponding vanilla fine-tuning approach and three existing weighting schemes. We assess the behaviour of WeLT when being fine-tuned while dealing with different dataset sizes and a variety of class distributions. In addition to the experimental analysis, we further share the implementation details and evaluation metrics.

4.1 Evaluation Datasets

NCBI Disease The American National Institutes of Health released the NCBI disease corpus to promote disease NER research. The public release of the NCBI disease corpus contains 6,892 disease mentions, which are mapped to 790 unique disease concepts.

BC5CDR-Disease and BC5CDR-Chemical The BioCreative V Chemical Disease Relation (CDR) corpus was created for the Chemical Disease Relation (CDR) Task. It consists of human annotations of all chemicals, diseases, and their interactions in 1,500 PubMed articles.

BC4CHEMD The BioCreative IV Chemical and Drug (BC4CHEMD) named entity recognition task corpus. It contains 10,000 abstracts annotated for mentions of chemical and drug names.

BC2GM The BioCreative II Gene Mention task corpus. It consists of 20,000 sentences from biomedical publication abstracts, annotated genes, and proteins.

Linnaeus Linnaeus corpus has 100 full-text documents for species annotations.

BioRED-Disease and BioRED-Chemical The BioRED corpus was created for multiple biomedical relations. It consists of human annotations of all different biomedical entities and their interactions in 600 PubMed abstracts. In this work, we only focus on chemical and disease instances.

4.2 Evaluation Metrics

Regarding the evaluation metrics, we report the entity-level precision, recall, and F1 scores. However, we assess the performance based on the entity-level F1 score. Due to the nature of our work that investigates the impact of addressing the class imbalance before fine-tuning, we do not compete with the state-of-the-art BioNER baselines. However, we compare the vanilla fine-tuning approach to WeLT and three existing weighting schemes using the same hyper-parameters.

Besides the overall assessment of the aforementioned fine-tuned models, we evaluate the annotation quality for species entities on the Linnaeus dataset. Thus, we used two sequence labeling metrics: (a) seqeval (Nakayama, 2018), and (b) FairEval (Ortmann, 2022). FairEval is one of the latest metrics on which Ortmann argues that the traditional evaluation metric causes double penalties for close-to-correct annotations. Therefore, Ortmann developed FairEval, which ensures that every error is counted only once.

4.3 Implementation

We adapt the BioBERT (Lee et al., 2020) PyTorch named entity recognition code to develop WeLT. In our experiments, we used the five following pre-trained model variants (a) BioBERT, (b) BlueBERT, (c) PubMedBERT, (d) SciBERT, and (e) BioELECTRA. For more details on the hyper-parameters, see Appendix A. All the experiments were carried out using two Tesla P40 GPUs with 24GB memory.

4.4 Compared Methods

As previously mentioned, we avoid using traditional resampling approaches, as there might be a possibility of information loss, and duplicating training examples leads to poor performance of language models (Lee et al., 2021). Therefore, in this paper, we focus on existing weighting schemes.

We compare WeLT to a vanilla fine-tuned approach (i.e., without handling class imbalance). We also evaluate WeLT with other weighting schemes: (a) Inverse of Number of Samples (INS), (b) Inverse of Square Root of Number of Samples (ISNS) , and (c) Effective Number of Samples (ENS) (Suri, 2022; Cui et al., 2019). Regarding the ENS approach, we used different values for β as follows: (a) 0.3, (b) 0.5, and (c) 0.9. For more details about the aforementioned weighting schemes, see Appendix B.

5 Results and Discussion

We present the results of seven fine-tuning experiments that include (a) INS, (b) ISNS, (c) ENS with $\beta = 0.3$, (d) ENS with $\beta = 0.5$, (e) ENS with $\beta = 0.9$, (f) vanilla, and (g) WeLT. For a fair comparison of all eight biomedical datasets, we applied the same experimental settings for all the fine-tuning approaches. We fine-tuned five Bio-PLMs using six weighting schemes and a vanilla approach on eight targeted datasets. Thus, we have 280 experimental results as presented in Tables (2)-(9).

At first glance, addressing the class imbalance enhances the performance. WeLT, different variants of ENS, and ISNS have the highest F1-score compared to their corresponding vanilla fine-tuning approach.

We extensively report the results based on the three following criteria: (a) size of training datasets,

Model	Metrics	INS	ISNS	ENS (0.3)	ENS (0.5)	ENS(0.9)	Vanilla	WeLT
BioBERT	P	**92.52**	90.10	89.32	91.81	92.17	91.81	91.21
	R	78.57	83.87	**86.39**	85.34	86.25	85.34	86.25
	F	84.98	86.88	87.83	88.46	**89.11**	88.46	88.66
PubMedBERT	P	**88.49**	85.47	85.31	85.63	85.15	85.63	86.98
	R	79.41	77.59	76.62	79.48	**80.87**	79.48	80.66
	F	**83.70**	81.34	80.73	82.44	82.96	82.44	**83.70**
BlueBERT	P	91.23	91.10	90.73	91.15	90.24	90.97	**91.35**
	R	50.87	64.34	65.59	64.41	64.54	85.83	**86.25**
	F	65.32	75.41	76.14	75.47	75.26	88.33	**88.72**
SciBERT	P	88.43	91.41	91.02	90.51	90.65	90.51	**92.44**
	R	46.96	62.38	65.80	64.61	66.36	64.61	**66.57**
	F	61.34	74.16	76.38	75.40	76.63	75.40	**77.40**
BioELECTRA	P	79.07	82.38	80.56	82.82	82.39	82.82	**84.15**
	R	70.41	75.71	79.83	81.08	81.99	81.08	**82.62**
	F	74.49	78.90	80.19	81.94	82.19	81.94	**83.38**

Table 2: Linnaeus fine-tuning scores. Precision (P), Recall (R), and F1-score (F) evaluation metrics. The best scores are shown in bold and the second best ones are underlined.

Model	Metrics	INS	ISNS	ENS (0.3)	ENS (0.5)	ENS(0.9)	Vanilla	WeLT
BioBERT	P	84.38	**84.46**	83.91	81.52	84.03	81.52	83.79
	R	84.57	85.66	86.21	84.46	84.68	84.46	**86.54**
	F	84.48	85.06	85.05	82.96	84.35	82.96	**85.14**
PubMedBERT	P	58.57	66.77	**71.77**	68.92	69.79	68.92	67.95
	R	43.32	64.87	**68.70**	65.75	67.50	65.75	67.50
	F	49.81	65.81	**70.20**	67.30	68.63	67.30	67.72
BlueBERT	P	65.56	68.88	67.16	65.10	66.56	65.10	**69.32**
	R	56.67	64.44	68.27	66.95	**68.38**	66.95	66.52
	F	60.79	66.59	67.71	66.01	67.45	66.01	**67.89**
SciBERT	P	68.98	**73.75**	71.91	71.68	72.33	69.29	72.34
	R	60.83	66.41	64.98	**68.70**	67.50	68.16	67.83
	F	64.65	69.89	68.27	**70.16**	69.83	68.72	70.01
BioELECTRA	P	83.88	**85.54**	83.97	84.02	83.97	84.02	84.95
	R	85.44	87.41	88.29	88.07	88.29	88.07	**89.60**
	F	84.66	86.47	86.08	86.00	86.08	86.00	**87.22**

Table 3: BioRed-Disease fine-tuning scores. Precision (P), Recall (R), and F1-score (F) evaluation metrics. The best scores are shown in bold and the second best ones are underlined.

Model	Metrics	INS	ISNS	ENS (0.3)	ENS (0.5)	ENS(0.9)	Vanilla	WeLT
BioBERT	P	87.93	88.73	**90.56**	88.37	89.54	85.00	88.57
	R	80.77	84.11	87.18	**87.31**	86.91	84.77	86.91
	F	84.20	86.36	**88.84**	87.84	88.21	84.89	87.73
PubMedBERT	P	89.03	88.30	88.72	89.55	89.04	89.55	**90.57**
	R	82.37	**87.71**	86.11	85.84	85.71	85.84	85.98
	F	85.57	88.01	87.39	87.66	87.34	87.66	**88.21**
BlueBERT	P	86.63	87.18	87.29	**89.05**	88.36	86.42	88.80
	R	86.51	88.11	**90.78**	90.12	88.25	90.12	88.91
	F	86.57	87.64	89.00	**89.58**	88.30	88.23	88.85
SciBERT	P	73.86	80.88	79.84	82.37	81.60	**82.53**	81.48
	R	49.79	58.74	**67.69**	66.75	63.95	64.35	67.55
	F	59.48	68.05	73.26	73.74	71.70	72.31	**73.86**
BioELECTRA	P	**89.64**	86.27	88.11	61.55	85.97	61.55	89.43
	R	77.43	85.58	84.11	53.00	**86.78**	53.00	84.77
	F	83.09	85.92	86.06	56.95	86.37	56.95	**87.04**

Table 4: BioRed-Chemical fine-tuning scores. Precision (P), Recall (R), and F1-score (F) evaluation metrics. The best scores are shown in bold and the second best ones are underlined.

Model	Metrics	INS	ISNS	ENS ($\theta = 0.3$)	ENS ($\theta = 0.5$)	ENS($\theta = 0.9$)	Vanilla	WeLT
BioBERT	P	**93.39**	92.59	92.56	93.09	92.79	92.71	92.83
	R	90.64	92.94	92.72	92.88	93.31	92.92	**93.63**
	F	91.99	92.77	92.64	92.99	93.05	92.82	**93.23**
PubMedBERT	P	**93.39**	92.59	92.32	92.83	92.79	79.68	93.31
	R	90.64	**92.94**	89.74	90.99	89.93	80.10	91.19
	F	91.99	**92.77**	91.01	91.90	91.34	79.89	92.24
BlueBERT	P	**89.45**	88.10	86.72	86.28	86.98	86.28	87.68
	R	73.53	79.62	80.94	**81.65**	81.16	**81.65**	80.68
	F	80.71	83.65	83.73	83.90	83.97	83.90	**84.04**
SciBERT	P	87.95	89.28	88.82	**90.03**	89.51	87.01	89.22
	R	79.62	81.89	82.95	83.93	83.67	**84.62**	83.06
	F	83.58	85.43	85.78	**86.88**	86.49	85.80	86.03
BioELECTRA	P	**95.11**	94.66	94.28	94.00	94.28	94.00	94.07
	R	91.92	93.92	94.33	94.26	94.33	94.26	**94.65**
	F	93.49	94.29	94.30	94.13	94.30	94.13	**94.36**

Table 5: BC5CDR-Chemical fine-tuning scores. Precision (P), Recall (R), and F1-score (F) evaluation metrics. The best scores are shown in bold and the second best ones are underlined.

(b) pre-training approach, and (c) class imbalance percentage.

According to the statistics in Table (11), BioRED is the smallest dataset. WeLT has the best performance for fine-tuning experiments except for PubMedBERT on BioRED-Disease, as shown in Table (3). The same applies to BioRED-Chemical in Table (4) except for BioBERT and BlueBERT. On the other hand, BC4CHEMD is the largest dataset. WeLT achieves the second-best score for all the experiments in Table (7), except for BioBERT.

Regarding the class distribution percentage in Table (1), Linnaeus is the highest skewed dataset, and NCBI is the least imbalanced one. WeLT has the highest score for the experiments related to Linnaeus as shown in Table (2), except for the second-best score while fine-tuning BioBERT. Similarly, WeLT has the best score for all the experiments using NCBI except for BlueBERT, as shown in Table (9).

WeLT has the best score for the BC5CDR-Chemical experiments in Table (5) except for fine-tuning SciBERT and the second-best score for PubMedBERT. BC5CDR-Disease experiments in Table (6) show that WeLT has the best score for all the experiments, except fine-tuning SciBERT, and got the second-best score for BioBERT. Finally, BC2GM experiments in Table (8) show that WeLT outperformed except for being second-best score for fine-tuning SciBERT.

We believe that WeLT shows advantages over the chosen existing weighting schemes, because it considers the overall dataset class distribution. None of these weighting schemes considers the overall class distribution while calculating the re-scaled weights. WeLT presents a generic approach to calculate re-scaled weights based on their normalized inverse effect in the dataset. In other words, major and minor classes have smaller and higher weights, respectively, in WeLT. This gives higher loss costs for minor classes while fine-tuning.

We highlight special patterns and give some intuitions about the successful and failed cases related to the performance of WeLT:

1. Fine-tuning WeLT on the largest dataset (BC4CHEMD) has shown the second-best score for all the experiments but failed for fine-tuning BioBERT. Based on our observation, ENS variants have the best performance. We believe that considering the overall class distribution for calculating the re-scaled weights degrades the performance. Further investigations should be considered by adding data size as an additional factor. Despite ENS variants outperformance, we emphasize that adding extra hyper-parameters in ENS is problematic and expensive since the appropriate β factor is unknown.

2. WeLT has the best fine-tuning scores for the highly skewed dataset (Linnaeus) on all the experiments, except for BioBERT. We believe that the calculation of new re-scaled weights has a positive impact on performance. The rest of the other weighting schemes focus only the number of class samples and not on the overall class distribution.

3. Fine-tuning BioELECTRA using WeLT has

the best scores except for BC4CHEMD. BioELECTRA is the biomedical version of ELECTRA. ELECTRA's pre-training strategy uses a more efficient approach called "replaced token detection". It learns from all input tokens rather than just the small subset that was masked out like in BERT models. It uses another neural network that aims to trick the model by replacing random tokens with fake tokens.

4. Regarding fine-tuning PubMedBERT as domain-specific pre-trained language model, WeLT has the best score for these four datasets as follows: (a) Linnaeus, (b) BioRED-Chemical, (c) BC5CDR-Disease, and (d) BC2GM. ENS variants has the best score for fine-tuning BioRED-Disease dataset. For the three other remaining datasets, WeLT has the second-best score. We believe that WeLT and different variants of ENS should be considered when fine-tuning PubMedBERT. The same applies for SciBERT experiments.

6 Error Analysis

We have further analyzed different types of BioNER mismatches for error analysis. Based on our observations, we noticed that still trivial BioNER errors occur. Yet, addressing the class imbalance before fine-tuning has a positive impact on the overall performance and the sequence labeling evaluation as shown in Table (10). As a proof-of-concept, we further evaluated the tagging quality outputs of each fine-tuning approach on the Linnaeus datasets. We report the F1-scores using seqeval with strict mode and FairEval with fair mode. WeLT has the best score for fine-tuning BlueBERT, SciBERT and BioELECTRA fine-tuned models and got the second best score for BioBERT and PubMedBERT.

The observed mismatches occur due to the following three types of errors:

- **Type-1:** an entity predicted by the NER model but not annotated in the gold-standard datasets. For instance, *"S"* is detected by BioPLMs since it is an abbreviation for *"Sulphur"*. However, it was not annotated by human experts in the BC5CDR gold standard dataset.

- **Type-2:** an entity annotated in the gold-standard datasets but not predicted by the NER model. The main issue behind the misclassification are abbreviated entities. For example, *"PAN"*, which is an abbreviation for *"Peroxyacetyl nitrate"*, is not recognized as a chemical entity.

- **Type-3:** an entity correctly predicted but has overlapping spans errors. For example, BioPLMs recognizes two chemical entities separately such as *"amphotericin B-"* and *"sodium deoxycholate"*. However, the gold-standard annotation is *"amphotericin B-sodium deoxycholate"*.

We believe that the first two mismatch error types require knowledge enrichment and usage of Active Learning approaches to tackle such semantic annotation issues. The third mismatch error can be fixed by enhancing the post-processing method to better merge tokens that are part of a recognized entity.

7 Conclusion and Future Directions

In this work, we propose a weighted loss trainer (WeLT) as a cost-sensitive approach that addresses the class imbalance problem of Biomedical gold-standard datasets before fine-tuning models. We have evaluated WeLT using five different BioPLMs, including general-domain and domain-specific pre-training. We do not only focus on Biomedical BERT variants, but also BioELECTRA. We have conducted **280** experiments in total (8 datasets, 5 BioPLMs, and 7 different fine-tuning approaches). Furthermore, we have extensively evaluated WeLT's performance against other fine-tuning approaches. Moreover, we present a thorough error analysis. We only focus on BioNER as a downstream task but we believe that WeLT can be adapted to other tasks with a class imbalance problem. WeLT can also be applied to non-biomedical domain applications. In the future, we plan to evaluate WeLT's performance for fine-tuning larger datasets (i.e., more than 30,683 training examples). More precisely, in this work, BC4CHEMD is the largest dataset with 30,683 training instances.

Limitations

We propose WeLT as an approach to address the class imbalance problem. We have only evaluated WeLT for BioNER tasks, although we hypothesize that it can be adapted to any application / domain that has skewed dataset. We further want to point

out that our method was only evaluated on English datasets. Yet, we argue that it can be applied to other languages as well. Finally, we have not assessed WELT's performance on larger training datasets (i.e., more than 30,683 training examples).

Ethics Statement

We see no impending risk in the development of our proposed and potential applicability to other non-biomedical applications. However, we would like to point out the aforementioned limitations (see the previous section). We report all the required hyper-parameters, implementation details, and hardware specifications to reproduce our results. We release all the fine-tuned models on the Hugging Face Hub.

Acknowledgments

WM and GM wish to acknowledge base funding by the HITS and the Klaus Tschira Foundation, KTS, supporting this work. This paper is supported by European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement PoLiMeR, No 812616

References

Abbas Akkasi, Ekrem Varoğlu, and Nazife Dimililer. 2018. Balanced undersampling: a novel sentence-based undersampling method to improve recognition of named entities in chemical and biomedical text. *Applied Intelligence*, 48:1965–1978.

Sultan Alrowili and Vijay Shanker. 2021. BioM-transformers: Building large biomedical language models with BERT, ALBERT and ELECTRA. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 221–227, Online. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. 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.

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.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Charles Elkan. 2001. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd.

Martin Gerner, Goran Nenadic, and Casey M Bergman. 2010. Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):1–17.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. BioELECTRA:pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online. Association for Computational Linguistics.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):1–17.

Robert Leaman, Rezarta Islamaj, Virginia Adams, Mohammed A Alliheedi, João Rafael Almeida, Rui Antunes, Robert Bevan, Yung-Chun Chang, Arslan Erdengasileng, Matthew Hodgskiss, et al. 2023. Chemical identification and indexing in full-text articles: an overview of the NLM-Chem track at BioCreative VII. *Database*, 2023.

Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. 2019. Mixout: Effective regularization to finetune large-scale pretrained language models. *arXiv preprint arXiv:1909.11299*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. BioRED: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*. Bbac282.

Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. 2020. Cost-sensitive bert for generalisable sentence classification with imbalanced data. *arXiv preprint arXiv:2003.11563*.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

Hiroki Nakayama. 2018. seqeval: A Python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Katrin Ortmann. 2022. Fine-grained error analysis and fair evaluation of labeled spans. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1400–1407, Marseille, France. European Language Resources Association.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In NIPS 2017 Autodiff Workshop: The Future of Gradient-based Machine Learning Software and Techniques.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Hong Shen and Anoop Sarkar. 2005. Voting between multiple data representations for text chunking. In *Advances in Artificial Intelligence*, Lecture Notes in Computer Science, page 389–400, Berlin, Heidelberg. Springer.

Larry Smith, Lorraine K Tanabe, Cheng-Ju Kuo, I Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, Manabu Torii, et al. 2008. Overview of BioCreative II gene mention recognition. *Genome biology*, 9(2):1–19.

Manan Suri. 2022. Pickle at semeval-2022 task 4: Boosting pre-trained language models with task specific metadata and cost sensitive learning. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 464–472.

Taha ValizadehAslani, Yiwen Shi, Jing Wang, Ping Ren, Yi Zhang, Meng Hu, Liang Zhao, and Hualou Liang. 2022. Two-stage fine-tuning: A novel strategy for learning class-imbalanced data. *arXiv preprint arXiv:2207.10858*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 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.

Yang Zhao, Zoie Shui-Yee Wong, and Kwok Leung Tsui. 2018. A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection. *Journal of healthcare engineering*, 2018.

A Hyper-parameters

We report all the hyper-parameters that were used to train the 280 models in the WeLT GitHub Repository.

B Comparable Weighting Schemes

Inverse of Number of Samples (INS)

$$weight[class] = \frac{1}{n_c} \tag{5}$$

where n_c is the corresponding number of samples in the class.

Inverse of Square Root of Number of Samples (ISNS)

$$weight[class] = \frac{1}{\sqrt{n_c}} \tag{6}$$

The main difference between INS and ISNS is that INS increases the recall by decreasing the values of false negatives while having low precision as the weights of majority class has been diminished. Therefore, ISNS overcomes this by taking the square root of class frequencies so that the class weights would be larger than INS.

Effective Number of Samples (ENS)

$$weight[class] = \frac{1-\beta}{1-\beta^{n_c}} \tag{7}$$

where β is a hyper-parameter $\in [0, 1]$.

C Dataset statistics and other results

Due to page limitations, we report the rest of our experimental results and the gold-standard datasets statistics here.

Model	Metrics	INS	ISNS	ENS ($\theta = 0.3$)	ENS ($\theta = 0.5$)	ENS($\theta = 0.9$)	Original	WeLT
BioBERT	P	85.13	84.14	85.14	84.70	84.91	84.70	**85.61**
	R	85.03	**86.25**	86.18	86.12	85.37	86.12	85.53
	F	85.08	85.18	**85.66**	85.40	85.14	85.40	85.57
PubMedBERT	P	78.05	79.05	80.22	79.32	79.99	79.32	80.67
	R	74.68	77.28	77.41	**78.05**	77.89	**78.05**	77.28
	F	76.33	78.15	78.79	78.68	78.92	78.68	**78.94**
BlueBERT	P	77.13	77.39	76.97	77.00	75.72	77.00	**78.12**
	R	70.32	75.15	76.76	**77.19**	76.92	**77.19**	76.67
	F	73.57	76.26	76.86	77.09	76.31	77.09	**77.38**
SciBERT	P	**79.45**	78.50	**79.45**	78.49	79.04	78.49	79.19
	R	69.75	74.95	76.58	76.74	**77.35**	76.74	76.55
	F	74.28	76.68	77.99	77.60	**78.19**	77.60	77.85
BioELECTRA	P	86.35	86.97	85.83	85.15	85.62	85.15	**87.58**
	R	86.55	87.20	87.81	87.74	**89.01**	87.74	87.68
	F	86.45	87.08	86.81	86.42	87.28	86.42	**87.63**

Table 6: BC5CDR-Disease fine-tuning scores. Precision (P), Recall (R), and F1-score (F) evaluation metrics. The best scores are shown in bold and the second best ones are underlined.

Model	Metrics	INS	ISNS	ENS ($\theta = 0.3$)	ENS ($\theta = 0.5$)	ENS($\theta = 0.9$)	Vanilla	WeLT
BioBERT	P	**92.49**	91.97	91.84	91.78	91.48	91.78	91.70
	R	88.14	89.80	90.54	90.21	**90.93**	90.21	90.45
	F	90.26	90.88	91.18	90.99	**91.20**	90.99	91.07
PubMedBERT	P	**91.69**	81.36	91.21	90.75	90.06	90.75	91.36
	R	85.59	80.52	88.63	88.43	**88.87**	88.43	88.38
	F	88.54	80.94	**89.90**	89.57	89.46	89.57	89.84
BlueBERT	P	**89.55**	89.21	89.07	88.67	88.88	88.67	89.22
	R	82.81	84.62	85.68	**85.85**	**85.85**	**85.85**	85.46
	F	86.05	86.85	**87.34**	87.24	**87.34**	87.24	87.30
SciBERT	P	**81.17**	80.18	80.05	80.20	80.90	79.71	79.71
	R	68.13	72.88	**75.03**	74.43	73.67	74.35	74.99
	F	74.08	76.36	**77.46**	77.21	77.12	76.93	77.28
BioELECTRA	P	**93.19**	92.71	92.87	92.57	92.70	92.57	93.01
	R	89.80	91.02	91.70	91.85	**92.00**	91.85	91.61
	F	91.46	91.86	92.28	92.21	**92.35**	92.21	92.30

Table 7: BC4Chem fine-tuning scores. Precision (P), Recall (R), and F1-score (F) evaluation metrics. The best scores are shown in bold and the second best ones are underlined.

Model	Metrics	INS	ISNS	ENS ($\theta = 0.3$)	ENS ($\theta = 0.5$)	ENS($\theta = 0.9$)	Vanilla	WeLT
BioBERT	P	83.09	82.89	82.95	82.62	82.57	82.62	**83.34**
	R	82.43	82.51	83.35	83.44	**83.47**	83.44	83.30
	F	82.76	82.70	83.15	83.03	83.02	83.03	**83.32**
PubMedBERT	P	**84.72**	84.32	74.27	73.74	73.66	83.47	83.99
	R	83.46	84.60	73.96	74.30	73.80	84.90	**85.48**
	F	84.08	84.46	74.11	74.02	73.73	84.18	**84.73**
BlueBERT	P	83.66	83.63	83.96	83.92	83.36	83.92	**84.56**
	R	82.67	83.60	84.42	84.14	**84.45**	84.14	83.93
	F	83.16	83.61	84.19	84.03	83.90	84.03	**84.24**
SciBERT	P	72.31	72.96	73.15	72.55	**73.61**	72.98	73.05
	R	71.60	73.09	75.08	75.35	**75.77**	74.86	75.35
	F	71.95	73.02	74.10	73.92	**74.68**	73.90	74.18
BioELECTRA	P	**83.89**	83.28	83.34	83.25	83.47	83.25	83.73
	R	83.58	84.77	85.13	84.79	85.05	84.79	**85.29**
	F	83.74	84.02	84.23	84.01	84.25	84.01	**84.50**

Table 8: BC2GM-Gene fine-tuning scores. Precision (P), Recall (R), and F1-score (F) evaluation metrics. The best scores are shown in bold and the second best ones are underlined.

Model	Metrics	INS	ISNS	ENS ($\theta = 0.3$)	ENS ($\theta = 0.5$)	ENS($\theta = 0.9$)	Original	WELT(ours)
BioBERT	P	85.75	85.41	86.74	85.23	85.33	86.12	**86.87**
	R	87.81	88.43	88.64	87.81	87.91	87.91	**88.95**
	F	86.77	86.89	87.68	86.50	86.60	87.01	**87.90**
PubMedBERT	P	80.73	81.36	79.70	79.68	81.58	79.68	**82.45**
	R	77.70	**80.52**	79.79	80.10	80.31	80.10	79.79
	F	79.19	80.94	79.75	79.89	80.94	79.89	**81.10**
BlueBERT	P	**86.76**	86.47	86.52	86.17	86.36	86.17	86.40
	R	88.75	89.27	90.31	88.95	**91.04**	88.95	90.00
	F	87.74	87.85	88.37	87.54	**88.64**	87.54	88.16
SciBERT	P	86.38	85.77	84.96	85.95	**86.73**	85.95	86.34
	R	88.54	89.16	88.33	89.27	88.54	89.27	**89.58**
	F	87.44	87.43	86.61	87.58	87.62	87.58	**87.93**
BioELECTRA	P	86.55	87.26	85.74	85.65	85.65	85.65	**87.66**
	R	83.85	87.81	88.33	88.33	88.33	88.33	**88.85**
	F	85.18	87.53	87.01	86.97	86.97	86.97	**88.25**

Table 9: NCBI fine-tuning scores. Precision (P), Recall (R), and F1-score (F) evaluation metrics. The best scores are shown in bold and the second best ones are underlined.

Model	Metrics	INS	ISNS	ENS (0.3)	ENS (0.5)	ENS(0.9)	Vanilla	WeLT
BioBERT	Seqeval	85.01	86.88	87.86	88.49	**89.15**	88.49	88.70
	FairEval	86.76	88.37	89.16	89.72	**90.08**	89.72	89.92
PubMedBERT	Seqeval	82.04	84.89	86.79	**88.79**	86.98	84.89	86.45
	FairEval	84.36	86.62	88.17	**89.19**	88.17	86.62	88.18
BlueBERT	Seqeval	65.56	75.42	76.14	76.14	76.14	88.33	**88.73**
	FairEval	66.33	76.25	76.52	76.52	76.52	89.58	**89.74**
SciBERT	Seqeval	61.35	74.16	76.39	75.41	76.63	75.41	**77.40**
	FairEval	63.02	74.83	76.84	76.06	76.85	76.06	**77.55**
BioELECTRA	Seqeval	74.49	78.91	80.20	81.95	82.20	81.95	**83.38**
	FairEval	78.27	81.32	82.21	84.06	84.33	84.06	**85.64**

Table 10: Sequence labeling evaluation F1-score for species entities in Linnaeus using seqeval with strict mode and FairEval with fair mode. The best scores are in bold and second best ones ae underlined.

Dataset	num_training	num_validation	num_test
NCBI	5433	924	941
BC5CDR-Disease	4561	4582	4798
BC5CDR-Chemical	4561	4582	4798
BC4CHEMD	30683	30640	26365
BC2GM	12575	2520	5039
BioRED-Chemical	4432	1140	1108
BioRED-Disease	4432	1140	1108
Linnaeus	11936	4079	7143

Table 11: Number of sentences in biomedical ground-truth datasets for training, development, and test data.