DALLMi: Domain Adaption for LLM-Based Multi-label Classifier

(Article begins on next page)

26 November 2024

DALLMi: Domain Adaption for LLM-based Multi-label Classifier

Miruna Bețianu[1][0000−1111−2222−3333], Abele Mălan✉[1,2][0009−0002−4493−7439],
Marco Aldinucci[3][0000−0001−8788−0829], Robert Birke[3][0000−0003−1144−3707], and
Lydia Chen[1,2][0000−0002−4228−6735]

[1] Delft University of Technology, Delft, Netherlands.
Email: m.betianu@student.tudelft.nl
[2] University of Neuchâtel, Neuchâtel, Switzerland.
Email: {abele.malan, yiyu.chen}@unine.ch
[3] University of Turin, Torino, Italy.
Email: {marco.aldinucci, robert.birke}@unito.it

Abstract. Large language models (LLMs) increasingly serve as the backbone for classifying text associated with distinct domains and simultaneously several labels (classes). When encountering domain shifts, e.g., classifier of movie reviews from IMDb to Rotten Tomatoes, adapting such an LLM-based multi-label classifier is challenging due to incomplete label sets at the target domain and daunting training overhead. The existing domain adaptation methods address either image multi-label classifiers or text binary classifiers. In this paper, we design DALLMi, **D**omain **A**daptation **L**arge **L**anguage **M**odel **i**nterpolator, a first-of-its-kind semi-supervised domain adaptation method for text data models based on LLMs, specifically BERT. The core of DALLMi is the novel variation loss and MixUp regularization, which jointly leverage the limited positively labeled and large quantity of unlabeled text and, importantly, their interpolation from the BERT word embeddings. DALLMi also introduces a label-balanced sampling strategy to overcome the imbalance between labeled and unlabeled data. We evaluate DALLMi against the partial-supervised and unsupervised approach on three datasets under different scenarios of label availability for the target domain. Our results show that DALLMi achieves higher mAP than unsupervised and partially-supervised approaches by 19.9% and 52.2%, respectively.

Keywords: Large language models · Domain adaptation · Multi-label classification · Semi-supervised learning

1 Introduction

Text classification is a fundamental task in Natural Language Processing with a diverse range of applications, including sentiment analysis [14], spam detection [6], and document categorization [11]. Large Language Models (LLMs) are the state-of-the-art backbone for such classification tasks [22]. By pre-training on large and heterogeneous corpora, they achieve good baseline performance across
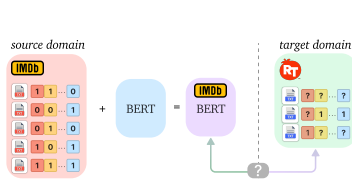
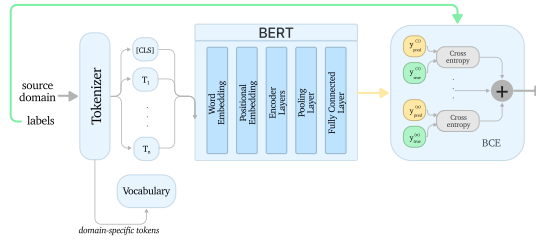Fig. 1: Adapting *BERT* from IMDb to Rotten Tomatoes.

Fig. 2: *BERT* multi-label classification flow.

various domains, even in cases where multiple labels[4] may be present simultaneously. When encountering a new domain with a sufficient amount of labeled data, fine-tuning LLMs can further increase the accuracy of classification tasks.

In real life scenarios, there is only limited labeled data in many target domains, making regular fine-tuning ineffective and requiring a domain shift from source domain. Fig. 1 outlines an example of such a challenge in classifying text into multiple labels/classes. Assume we have an LLM trained to predict the genres of a movie from an IMDb dataset, as the source domain. Each movie is associated with multiple genres/labels. We then have to perform the same task on a new dataset from Rotten Tomatoes, as the target domain, which contains incomplete labels. The annotation of the target domain is a time-consuming and resource-intensive task, nonetheless unfeasible. The research challenge is how to utilize the knowledge of the source domain and the limited labels for the target domain to enhance the performance of LLM-based multi-label classifiers.

Existing text classification adaptation research [21] focuses on the multi-class case, where each sample corresponds to only one label from the available set, like in the ubiquitous sentiment analysis task or the even simpler binary one. The multi-label scenario, where samples correspond to any number of labels, poses an increased challenge, further increased by possibly underrepresented labels in the available data. Most prior multi-label domain adaptation efforts lie within the modality of image data [15,20]. Moreover, only few unsupervised approaches exist to finetune LLMs [18,17], which are usually limited to single label classification.

This paper introduces `DALLMi`, a novel semi-supervised technique for LLM adaptation to different textual domains. We combine supervised fine-tuning on a source text domain with semi-supervised fine-tuning on a partially labeled text target domain. Specifically, we focus on scenarios where there are limited positive samples for each label, and the remaining samples are unlabeled, as shown in Fig. 1. `DALLMi` introduces a variational loss that leverages labeled and unlabeled information to maximize the knowledge extracted from all samples. To further boost its classification ability, `DALLMi` augments the target dataset with synthetic samples generated by mixing labeled and unlabeled ones. We extensively evaluate `DALLMi` under various settings for multiple datasets, including ablations on relevant features. To recap, we make the following contributions:

---

[4] each label represents a possible class

- We design `DALLMi`, the first semi-supervised LLM domain adaptation framework for multi-label text classification.
- We design a novel variational loss for the model to learn from both labeled and unlabeled parts of target instances.
- We interpolate representations of labeled and unlabeled samples into new synthetic ones to regularize our variational loss.
- `DALLMi` outperforms partial fine-tunning and unsupervised approaches by 19.9% and 52.2%, respectively.
- We make the code available at `https://github.com/mirunabetianu/DALLMi`.

2 Language Model and Domain Adaptation

BERT for multi-label classification. Fig. 2 shows a high-level overview of the architecture of BERT's multi-label classifier, which assumes a full set of labels. The input sentence is initially passed through the tokenizer to obtain tokens, paddings, and truncation to prepare it for BERT. This process may also include the addition of new tokens to BERT's vocabulary. The encoded sentence proceeds through BERT's forward method, which incorporates embedding and encoding layers, to obtain the internal hidden representations. A pooling layer decreases the hidden representation's dimension, from which a classification head (usually a single fully connected layer) obtains the final logits. An activation function may be applied to determine predicted labels based on a threshold. The default training approach uses multiple binary cross entropy losses to treat the multi-label classification as several binary problems.

Prior Art Domain Adaptation. Domain adaptation addresses the dataset shift [16] between the source and target domains. It can be split into four main categories: fully supervised, semi-supervised, weakly supervised, and fully unsupervised. Fully supervised methods rely heavily on many target labels, aiming to generalize the model to both domains [13]. When the target domain labels are limited, the problem becomes semi-supervised. Common approaches combine source and domain data to improve model training [27]. Weakly supervised methods aim to adapt the model by considering the uncertainty and limitations of the target domain [7]. Unsupervised domain adaptation aims to adapt the model without any information about the target labels, approaches including batch normalization tuning [24], feature alignment [8], or whole network training [12]. Most methods operate within image domains. Nonetheless, some techniques exist for LLMs, with fine-tuning being the most common adaptation method [9,10,2]. Other methods expand the LLM's vocabulary, enhancing the corpus with specific domain tokens [19], or its architecture, adding new adapter layers [5]. In terms of classification adaptation, the majority of LLM-based approaches are designed for sentiment analysis, using adversarial adaptation with knowledge distillation [17] or supervised fine-tuning [18]. Our method, however, targets general classification workloads and employs semi-supervised fine-tuning.

3 `DALLMi`

Adapting a fine-tuned LLM to partially labeled target data is a challenging task, primarily due to the need for a substantial amount of labels to ensure that the model gives reliable predictions. To address the challenge posed by the scarcity of the target domain, we introduce a new semi-supervised fine-tuning method based on Positive Unlabeled (PU) learning and data augmentation. Our method takes inspiration from VPU [4], estimating separate losses for positive and unlabeled samples and combining positive-unlabeled pairs to generate artificial samples to compensate for the limited number of labels.

Formally, we aim to learn a multi-label classifier, $\Phi : X^t \rightarrow Y$, where $X^t \in \mathbb{R}^d$ is the input data from the target domain, and $Y = [y_1 \ldots y_L], \forall y_l \in [1,0]$ is the value vector for each label from 1 to L. Given some label l, $y_l = 1$ denotes the presence of a positive example, while $y_l = 0$ denotes the absence of an annotation. For each possible label $l \in \{0 \ldots L\}$, we separate X^t into two sets: the unlabeled set $\mathcal{U}_N^l$ with samples s_u and the positive set $\mathcal{P}_N^l$ with samples s_p.

Primer on VPU. Positive unlabeled learning [1] focuses on the scenarios where labeled samples are always a subset of the positive ones, and unlabeled ones may be positive or negative. The goal of the trained model is to distinguish between positive and negative instances, as in the classic fully labeled case [1]. Rooted in the variational principle, VPU [4] uses a novel variation loss function to approximate the ideal **binary** classifier without the class prior, given unlabeled and positive samples. The variational loss function aims to quantify the empirical difference between the predicted outcomes for unlabeled and positive samples. In addressing the lack of positive samples, the VPU algorithm introduces a consistency regularisation term that incorporates the principles of MixUp, mitigating the problem of overfitting and increasing the robustness of the model. The MixUp term uses data augmentation techniques to generate synthetic samples by interpolating between positive and unlabeled instances. In addition, the algorithm quantifies the consistency between the model's predictions and the predictions made on these interpolated samples by the mean squared logarithmic error. We note that PU-MLC considers the problem of expanding VPU for multi-label classification [25], but in the image analysis domain.

Overview of `DALLMi`. To jointly leverage both positive labeled and unlabeled data, `DALLMi` fine-tunes the LLM-based multi-label classifier in a semi-supervised manner. We extend the design idea of VPU, combining the variation loss per label and MixUp regularization based on data interpolation, for a multi-label classifier. In contrast to image data or tabular data, such an interpolation is not straightforward for text data, which is represented as tokens in the LLM. We thus need to design a novel loss function and interpolation scheme to compute the MixUp, and then train the multi-label classifier and backbone BERT.

`DALLMi` consists of three key novelties, namely label-balanced sampling, variational loss and MixUp regularization. Fig. 3 shows the primary flow of `DALLMi`
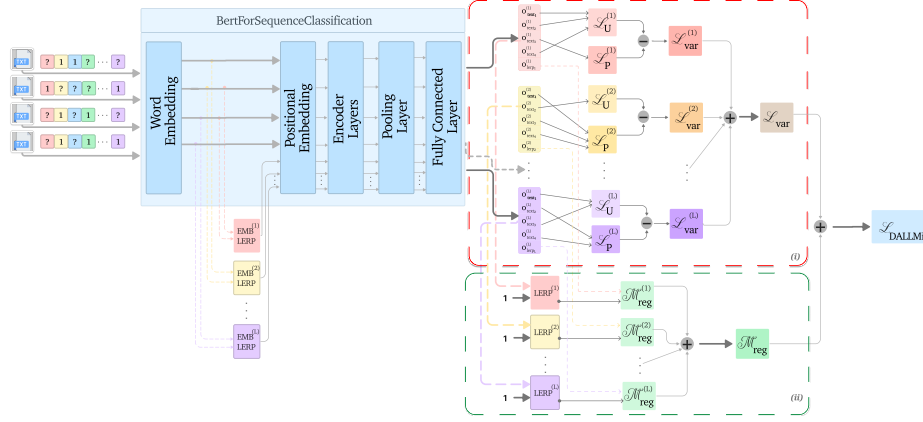
Fig. 3: `DALLMi` flow: Unlabeled (?) and positive (1) samples from the target domain are fed through BERT, generating label-specific output logits. The logits are used to compute partial per-label variational losses for unlabeled and positive samples (dashed box (i)). The MixUp regularization combines per label linear interpolations (LERPs) applied to both inputs and outputs (dashed box (ii)).

for computing the proposed variational loss and MixUp regularization for an example set of 4 samples. The target domain is pre-processed to ensure compatibility with the BERT classifier (blue shaded box). After the initial word embeddings layer, a distinctive input interpolation mechanism is introduced for each label. Here, unlabeled and positive embeddings are paired to generate new artificial embeddings via a linear interpolation block (LERP). After interpolation, the original and newly generated embeddings pass through the remaining layers of the BERT classifier. The final layer features a classification head represented as a single fully connected layer outputting L logit scores. These logit scores are used to compute a specialised multi-label variational loss, $\mathcal{L}_{var}$, which sums up all per-label variation losses (dashed block (i)). The MixUp regularization $\mathcal{M}_{reg}$ sums up per label combinations of input interpolations and output interpolations, which are determined by linearly interpolating the logits derived from the input interpolation and the true output (dashed block (ii)).

Label-balanced Sampling. In multi-label classification, where labels are representable as a sparse matrix, the challenge is creating label-balanced batches, particularly as the label count increases. Since, for any label there are often more unlabeled samples than positive ones, we need to overcome this disparity by ideally boosting the number of labeled samples from each label. Therefore, the first step in `DALLMi` involves an efficient and balanced sampler that ensures at least one positive sample from each label exists in every batch.

To achieve such balance, we introduce a cycle sampler that iterates through every label and retrieves one positive sample, repeating the process as many as needed until to reach the desired batch size. By ensuring the presence of at least

one positive sample for each label, we obtain a fair and accurate estimation of the variational loss and allow calculating the linear interpolation for the MixUp.

Variational Loss per Label. Having at least one positive sample from each label allows the loss function to make reasonable estimations across the sample set. In this context, we introduce the variational loss; see the red dashed box in Fig. 3. For each label, the loss, grounded in the variational principle, uses positive and unlabeled samples to minimize the divergence between the positive distributions of the modeled and ideal classifiers as a proxy for the original task. The difference between the log of the expected sigmoids for unlabeled data and the expected value of the log sigmoids for labeled data serves as the approximation for the classifier's positive distribution. Interestingly, through empirical analysis, our findings diverge from this conventional approach. Specifically, we observe that replacing the log operations with the norm yields significant performance improvements. This norm-based approach is particularly advantageous when working within a specific domain. In addition, using the logarithm can exacerbate errors in certain situations, especially when dealing with labels with very low probabilities. This problem becomes more pronounced as the width of the datasets increases. In both cases, aggregating the values across all possible labels gives the final variational loss:

$$\mathcal{L}_{var}^{l}(\Phi) = \frac{1}{|\mathcal{U}_N^l|} \sum_{s_u \in \mathcal{U}_N^l} \sigma(\Phi(s_u)) - \frac{1}{|\mathcal{P}_N^l|} \sum_{s_p \in \mathcal{P}_N^l} ||\sigma(\Phi(s_p))||, \qquad (1)$$

where $\sigma(\cdot)$ is the sigmoid function, and $||\cdot||$ represents the norm of a vector.

Although the effectiveness of variational loss is proven [4,25], it has limitations, notably the need for a substantial number of positive samples in the dataset. To mitigate this limitation, we use data augmentation.

MixUp Regularization. We compute the MixUp [26,23] term based on synthetically augmented data. The primary purpose of the MixUp is to generate additional samples by interpolating between positive and unlabeled samples, thus compensating for the lack of positive labels. The interpolations are weighted combinations using the Beta distribution to facilitate a smooth transition. In addition to effectively expand the dataset size, the MixUps quantify the consistency between the predicted and interpolated outputs.

LLM inputs consist of sequences of tokens that are not necessarily closer in value if they are closer in meaning. Therefore, simple token-level interpolation could result in problematic combinations that may hinder the model's learning process. For the interpolation to be possible, we need to delve into the BERT model layers to identify which internal representations of the text data would be suitable [23]. Fig. 4 illustrates three possible locations where to extract internal representations to perform MixUps and generate artificial samples. The first location is after the tokens pass the embedding layer of BERT. We extract

---

Algorithm 1 `DALLMi` – Loss Calculation

Input: Label-balanced sampled data batch B, BERT model Φ, hyperparameters $\alpha/\beta/\lambda$
Output: Loss value for samples in B

1: **for** each label l **do**
2:      Select unlabeled samples U_N^l and positive samples P_N^l from batch B
3:      Calculate variational loss by Eq. 1
4:      Select one pair of unlabeled and positive inputs, s_u and s_p
5:      Sample $\mu \sim Beta(\alpha, \beta)$
6:      $e_u, e_p = \Phi.\text{embeddings}(s_u), \Phi.\text{embeddings}(s_p)$
7:      Compute interpolation by Eq. 4
8:      Compute artificial model prediction by Eq. 5
9:      Compute MixUp term by Eq. 2
10: **end for**
11: Compute the overall loss $\mathcal{L}_{\text{final}}(\Phi) = \sum_{l=1}^{L} \mathcal{L}_{var}^l(\Phi) + \lambda \mathcal{M}_{reg}^l(\Phi)$
12: Backpropagate $\mathcal{L}_{\text{final}}(\Phi)$

---

the word embeddings for the chosen unlabeled and positive sentences and perform interpolation using the Beta distribution (see Eq. 2). The resulting new embedding undergoes the subsequent layers of BERT to generate predictions for this new sample. Similarly, the other two locations extract, interpolate and feed back the embeddings at the deeper encoding and pooling layers of the model. Our empirical evaluation of the all interpolation locations favours the word embedding variant. We then compute the MixUp via a norm-based mean squared error to determine the consistency between the predicted outcome for unlabeled samples and of interpolated one:

$$\mathcal{M}_{reg}^l(\Phi) = (||\sigma(\tilde{\Phi})|| - ||\sigma(\Phi(\tilde{e}))||)^2, \tag{2}$$

$$\text{where } \mu \overset{iid}{\sim} Beta(\alpha, \beta) \tag{3}$$

$$\tilde{e} = \mu \cdot e_p + (1 - \mu) \cdot e_u \tag{4}$$

$$\tilde{\Phi} = \mu \cdot 1 + (1 - \mu) \cdot \Phi(s_u) \tag{5}$$

where e_p and e_u represent the embedding corresponding to a positive s_p and unlabeled s_u sample, respectively; the hyperparameters α and β control the shape of the Beta distribution.

Algo. 1 summarises the overall computation of `DALLMi` loss. Following the steps displayed in Fig. 3, a balanced batch is selected and used for the calculation of per-label variational loss. Additionally, the algorithm creates input and output interpolations to compute per-label MixUp values. The last step of the algorithm combines the variational loss and MixUp regularization terms weighted via the λ hyperparameter. The gradients of the final loss are computed using backpropagation, which are then used to update the BERT classifier parameters and weights, gradually guiding the model towards better performance.
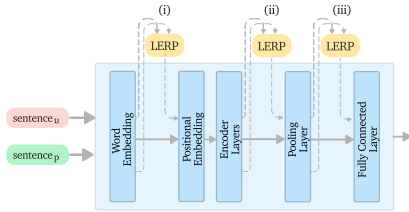
Fig. 4: Examples of possible MixUp strategies by linear interpolation in LLM hidden representations: (i) word embedding, (ii) encoding, (iii) sentence embedding.

Table 1: Evaluation datasets source problem domains & properties.

Dataset	Source/Target	#Samples	#Labels
PubMed	Male/Female	5757	14
arXiv	Old/New	7000	14
Movies	Wiki/IMDb	1000	20

4 Experiments

We assess the effectiveness of `DALLMi` against different LLM fine-tuning methods on three multi-label datasets: *PubMed*, *ArXiv*, and *Movies*.

Evaluation Setup. We use three datasets from different textual categories: healthcare, academia, and entertainment. *PubMed* is a medical dataset consisting of research articles from the PubMed[5] repository. The articles' subheadings denote the source and target domains, namely *female* and *male* patients. The labels represent different biological categories. *ArXiv* provides a collection of research paper abstracts, where labels represent subjects into which to categorize the abstract. Here, the target and source domains are old and new articles scrapped from the ArXiv repository[6]. The *Movies* dataset contains a collection of movie summaries that may belong to different genres. The source and target domains for the movie overviews are Wikipedia and IMDb, respectively. Table 1 shows an overview of the datasets' properties. Each dataset from both domains is split inot a 80% training and 20% testing using random partitioning. To simulate missing labels we discard 50%, 70%, and 90% of the labels for each sample.

Training Details. We use pretrained BERT *bert-base-uncased*[7]. Specifically, we employ the *BertForSequenceClassification* model and fine-tune it using the *Trainer* class for easier reproducibility, efficient GPU memory utilization, and simplified workflow. All runs use batch size equal to 64, learning rate equal to 5e-5, and 12 training epochs. We set the λ, α, β to 1, 0.3, 0.3. We run each experiment thrice and report the average mAP score [3] on the target domain. Furthermore, we do not freeze BERT's layers during the fine-tuning.

Baselines. We compare `DALLMi` with supervised fine-tuning, no fine-tuning, unsupervised domain adaptation method AAD [18], and an asymptotic upper bound using fully supervised fine-tuning without removing any labels. For AAD, we follow the authors' work, freezing BERT's layers and training for three epochs.

---

[5] https://pubmed.ncbi.nlm.nih.gov

[6] https://arxiv.org

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

Table 2: mAP scores under different ratios of available labels per column for: target domain data of source model, regular fine-tuning (supervised), AAD[18] (unsupervised), and DALLMi (semi-supervised).

Dataset	None	Method & Ratio Available Labels				AAD	DALLMi		
	None	Fine-Tuning				AAD	DALLMi		
	-	100%	50%	30%	10%	-	50%	30%	10%
PubMed	57.1	$62.0_{\pm.28}$	$50.6_{\pm0.40}$	$45.3_{\pm.36}$	$42.5_{\pm.05}$	$53.7_{\pm.40}$	$58.9_{\pm.48}$	$58.2_{\pm0.38}$	$52.2_{\pm.57}$
arXiv	21.8	$25.0_{\pm.24}$	$18.5_{\pm.12}$	$15.7_{\pm.45}$	$13.7_{\pm.19}$	$21.5_{\pm.03}$	$24.5_{\pm.32}$	$23.2_{\pm.60}$	$20.5_{\pm.48}$
Movies	20.7	$24.7_{\pm.39}$	$15.8_{\pm.55}$	$13.2_{\pm.09}$	$12.6_{\pm.00}$	$17.6_{\pm.12}$	$26.5_{\pm.21}$	$26.2_{\pm.48}$	$26.0_{\pm.36}$

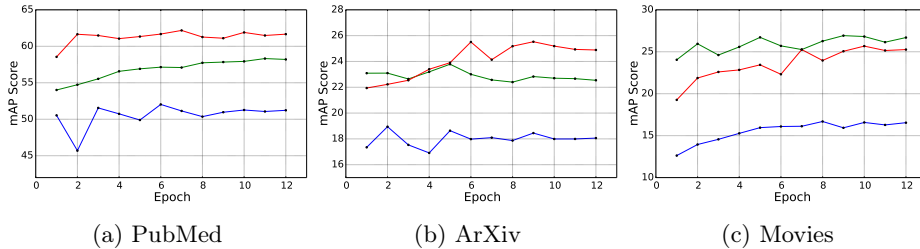(a) PubMed   (b) ArXiv   (c) Movies

Fig. 5: mAP scores/epoch for: supervised fine-tuning w/ 50% labels (blue), DALLMi w/ 50% labels (green), and supervised fine-tuning w/ 100% labels (red).

Empirical Results. Table 2 reports the mAP results of DALLMi compared to the baseline methods for different ratios of discarded labels per column. We observe that DALLMi consistently outperforms the supervised fine-tuning technique among all the dataset and label ratio combinations. Our method outperforms the unsupervised AAD in almost all cases, except for scenarios with only 10% label availability, where it remains competitive. Consistent with expectations, the supervised fine-tuning, having all labels, outperforms our proposed algorithm. However, in the case of the Movies dataset, our method achieves marginally higher performance than fully supervised fine-tuning. The variation can be explained by the dataset's low sample count and the wide nature of the dataset. We argue that our approach's effectiveness is due to its incorporation of multiple interpolations and the generation of artificial samples, which contribute to higher results in such data settings. Fig. 5 presents an overview of how the mAP score varies over the 12 epochs in the case of 50% label removal for each dataset within a fixed seed. Our method's performance fits between that of the fully supervised fine-tuning with and without label removal. Here, in Fig. 5c, we can better observe the phenomenon from Table 2 in Movies. Our method achieved slightly better performance than supervised fine-tuning without label removal.

Ablation Studies. To establish the effectiveness of DALLMi, we conduct three ablation studies. First, we analyze the impact of varying interpolation

Table 3: Relative changes in mAP score for `DALLMi`'s Table 2 results when changing (i) the word-level MixUp to encoding/sentence-level and (ii) the Norm-based loss to the Log-based formulation (negative percentage means score decrease).

Dataset	Ablation & Ratio Available Labels					
	(i) Encoding/Sentence MixUp			(ii) Log-based Loss		
	50%	30%	10%	50%	30%	10%
PubMed	-0.018%	-0.167%	-0.937%	-20.94%	-21.80%	-16.73%
arXiv	-0.497%	-0.455%	-1.234%	-24.94%	-23.34%	-18.16%
Movies	+0.547%	-0.966%	-0.935%	-39.76%	-44.47%	-46.75%

methods. Second, we compare our norm-based approach to the original log-based VPU. Third, we evaluate the efficiency of our sub-sampling technique.

Table 3(i) illustrates the differences in mAP scores resulting from the different interpolation methods, as opposed to the base word embedding interpolation. The reference mAP scores for the comparative analysis are those from Table 2. In particular, the encoding and sentence embeddings both show analogous scores, having slightly lower performance compared to the word embeddings in the PubMed and ArXiv datasets. In contrast, using encoding or sentence-based interpolations in the Movies dataset results in a marginal mAP score increase. This finding highlights the adaptability of the interpolation mechanism, indicating its potential for customization to the specifics of individual datasets. The Movies dataset constitutes a unique case study, where our method is validated under conditions characterized by a limited number of samples in a wide dataset.

Table 3(ii) shows the changes in mAP scores obtained by our method using a log-based variational approach. Exclusively, these scores display a decline in performance across all three datasets and even fall below those achieved by supervised fine-tuning with label removal. We argue that the complexities imposed by the internal representations of textual data make log-based methods overly strict with respect to deviations from learning, leading to a pronounced amplification of errors and, consequently, to poorer performance. In contrast, our proposed method consistently outperforms the baselines, with competitive results in the few cases where it does. This contrast highlights the robustness and validity of our approach over logarithmic variational strategies.

Table 4 illustrates the differences in training time and mAP score associated with different batch sampling strategies. In the first version of the `DALLMi`, we use a standard unweighted sampler, where the batches have a higher risk of being imbalanced, leading to a poorer performance, even if the training time is low. In the second version, we use nested batching to overcome the imbalance, pairing a large batch of unlabeled samples with smaller batches associated with positive samples for each specific label. Despite showing comparable results to our suggested method, nested batching significantly increases training time. In our final approach, we use a cycle sampler whereby each batch contains at least

Table 4: Training time and mAP score w/ 50% labels for: unweighted sampler, and nested batching. `DALLMi` achieves training time similar to the unweighted sampler with performance on the level of nested batching.

Dataset	Training time (s)		mAP Score	
	Unweighted Sample	Nested Batch	Unweighted Sample	Nested Batch
PubMed	1793.45	2607.62	58.13	58.86
arXiv	1553.97	2796.19	22.92	23.61
Movie Genres	369.54	631.78	25.36	25.22

one positive sample from every label. This adapted sampling technique maintains quality predictions while substantially reducing the training duration.

5   Conclusion

To tackle the domain shift in classifying unlabeled text data, we propose a semi-supervised domain adaptation approach, `DALLMi`, specifically for LLM-based multi-label classifiers. By incorporating a novel variational loss, MixUp regularizer, and label-balanced sampling, `DALLMi` effectively exploits the limited positively labeled and abundant unlabeled text data during domain adaptation. We evaluate `DALLMi` against partial-supervised and unsupervised methods across three datasets, accounting for various label availability scenarios, demonstrating its efficiency. Importantly, our method achieves higher mAP scores, exceeding unsupervised and partial-supervised approaches by 19.9% and 52.2%, respectively. Our results showcase `DALLMi`'s effectiveness and resilience of in enhancing multi-label classification tasks in the presence of domain shifts and scarce labels.

References

1. Bekker, J., Davis, J.: Learning from positive and unlabeled data: a survey. Mach. Learn. **109**(4), 719–760 (2020)
2. Buonocore, T.M., Crema, C., Redolfi, A., Bellazzi, R., Parimbelli, E.: Localizing in-domain adaptation of transformer-based biomedical language models. J. Biomed. Informatics **144**, 104431 (2023)
3. Cartucho, J., Ventura, R., Veloso, M.: Robust object recognition through symbiotic deep learning in mobile robots. In: IEEE/RSJ IROS. pp. 2336–2341 (2018)
4. Chen, H., Liu, F., Wang, Y., Zhao, L., Wu, H.: A variational approach for learning from positive and unlabeled data. In: NeurIPS. vol. 33, pp. 14844–14854 (2020)
5. Chronopoulou, A., Peters, M.E., Dodge, J.: Efficient hierarchical domain adaptation for pretrained language models. In: NAACL. pp. 1336–1351 (2022)

6. Crawford, M., Khoshgoftaar, T.M., Prusa, J.D., Richter, A.N., Najada, H.A.: Survey of review spam detection using machine learning techniques. J. Big Data **2**, 23 (2015)

7. Deng, A., Wu, Y., Zhang, P., Lu, Z., Li, W., Su, Z.: A weakly supervised framework for real-world point cloud classification. Comput. Graph. **102**, 78–88 (2022)

8. Eastwood, C., Mason, I., Williams, C.K.I., Schölkopf, B.: Source-free adaptation to measurement shift via bottom-up feature restoration. In: ICLR (2022)

9. Grangier, D., Iter, D.: The trade-offs of domain adaptation for neural language models. In: ACL. pp. 3802–3813 (2022)

10. Guo, Y., Rennard, V., Xypolopoulos, C., Vazirgiannis, M.: Bertweetfr: Domain adaptation of pre-trained language models for french tweets. In: W-NUT. pp. 445–450 (2021)

11. Lee, L.H., Wan, C.H., Rajkumar, R., Isa, D.: An enhanced support vector machine classification framework by using euclidean distance function for text document categorization. Appl. Intell. **37**(1), 80–99 (2012)

12. Liu, H., Long, M., Wang, J., Wang, Y.: Learning to adapt to evolving domains. In: NeurIPS. vol. 33, pp. 22338–22348 (2020)

13. Motiian, S., Piccirilli, M., Adjeroh, D.A., Doretto, G.: Unified deep supervised domain adaptation and generalization. In: IEEE ICCV. pp. 5716–5726 (2017)

14. Nasukawa, T., Yi, J.: Sentiment analysis: capturing favorability using natural language processing. In: K-CAP. pp. 70–77 (2003)

15. Pham, D.D., Koesnadi, S.M., Dovletov, G., Pauli, J.: Unsupervised adversarial domain adaptation for multi-label classification of chest x-ray. In: IEEE ISBI. pp. 1236–1240 (2021)

16. Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D.: Dataset shift in machine learning. MIT Press (2022)

17. Rietzler, A., Stabinger, S., Opitz, P., Engl, S.: Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In: LREC. pp. 4933–4941 (2020)

18. Ryu, M., Lee, G., Lee, K.: Knowledge distillation for BERT unsupervised domain adaptation. Knowl. Inf. Syst. **64**(11), 3113–3128 (2022)

19. Sachidananda, V., Kessler, J.S., Lai, Y.: Efficient domain adaptation of language models via adaptive tokenization. In: SustaiNLP@EMNLP. pp. 155–165 (2021)

20. Singh, I.P., Ghorbel, E., Kacem, A., Rathinam, A., Aouada, D.: Discriminator-free unsupervised domain adaptation for multi-label image classification (2023)

21. Singhal, P., Walambe, R., Ramanna, S., Kotecha, K.: Domain adaptation: Challenges, methods, datasets, and applications. IEEE Access **11**, 6973–7020 (2023)

22. Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., Wang, G.: Text classification via large language models (2023)

23. Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., Bengio, Y.: Manifold mixup: Better representations by interpolating hidden states. In: ICML. vol. 97, pp. 6438–6447 (2019)

24. Wang, D., Shelhamer, E., Liu, S., Olshausen, B.A., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: ICLR (2021)

25. Yuan, Z., Zhang, K., Huang, T.: Positive label is all you need for multi-label classification (2023)

26. Zhang, H., Cissé, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: ICLR (2018)

27. Zhang, Y., Zhang, H., Deng, B., Li, S., Jia, K., Zhang, L.: Semi-supervised models are strong unsupervised domain adaptation learners (2021)