Personalized LoRA for Human-Centered Text Understanding

You Zhang[1], Jin Wang[1*], Liang-Chih Yu[2*], Dan Xu[1], Xuejie Zhang[1]

[1]School of Information Science and Engineering, Yunnan University, Yunnan, P.R.China
[2]Department of Information Management, Yuan Ze University, Taiwan
{yzhang0202, wangjin}@ynu.edu.cn, lcyu@saturn.yzu.edu.tw

Abstract

Effectively and efficiently adapting a pre-trained language model (PLM) for human-centered text understanding (HCTU) is challenging since user tokens are million-level in most personalized applications and do not have concrete explicit semantics. A standard and parameter-efficient approach (e.g., LoRA) necessitates memorizing numerous suits of adapters for each user. In this work, we introduce a personalized LoRA (PLoRA) with a plug-and-play (PnP) framework for the HCTU task. PLoRA is effective, parameter-efficient, and dynamically deploying in PLMs. Moreover, a personalized dropout and a mutual information maximizing strategies are adopted and hence the proposed PLoRA can be well adapted to few/zero-shot learning scenarios for the cold-start issue. Experiments conducted on four benchmark datasets show that the proposed method outperforms existing methods in full/few/zero-shot learning scenarios for the HCTU task, even though it has fewer trainable parameters. For reproducibility, the code for this paper is available at: https://github.com/yoyo-yun/PLoRA.

Introduction

Human-centered text understanding (HCTU) aims to capture potential mental states in texts according to user preferences where user historical written texts are informative sources for user understanding (Crisan et al. 2022; Lynn et al. 2017; Capel and Brereton 2023). For users with different preferences and unique needs, similar texts might express various and diverse understandings, e.g., in personalized sentiment analysis (Zhang et al. 2021; Tang, Qin, and Liu 2015; Chen et al. 2016). With the thriving of pre-trained language models (PLMs) in natural language processing (NLP), the importance of personalization has been highlighted in recent works (Wu et al. 2023; Min et al. 2021; Liu, Zhang, and Gulla 2023).

Traditional personalized sentiment analysis methods typically use personalized knowledge injection (PKI) techniques (Zhong et al. 2021; Houlsby et al. 2019; Hu et al. 2021) to embed user attributes/preferences into dense representations and then inject them into neural networks, as shown in Fig. 1(a). However, these models require full-model fine-tuning (FFT) and sophisticated structures to couple task-specific transferring and personalized injecting, which is not suitable for adapting large-scale PLMs to personalized sentiment analysis.

To address the problem, parameter-efficient fine-tuning (PEFT) techniques such as the adapter, prompt-tuning, and low-rank adaptation (LoRA) (Wu et al. 2018; Zhang, Wang, and Zhang 2021) have been proposed. These techniques only require adding and fine-tuning a few parameters in PLMs, and thus can effectively and efficiently fine-tune large-scale PLMs for downstream tasks. To use PEFT for personalized sentiment analysis, as shown in Fig. 1(b), the user attributes/preferences can be considered as an adapter (dashed rectangles) to fine-tune large-scale PLMs through, for example, LoRA, thus avoiding fine-tuning the whole model parameters. However, each user is associated with an adapter and a full model copy is not practicable because user tokens are always million-level in real-world applications (e.g., Amazon). In addition, this may also suffer from the under-fitting problem due to limited training examples for each user. To leverage the advantage of both PKI and PEFT, this study proposes a personalized low-rank adaptation (PLoRA) mechanism by combining PKI and LoRA, as shown in Fig. 1(c). The PKI is used to embed all user attributes/preferences into a unified embedding, followed by LoRA to adapt large-scale PLMs to the HCTU task without FFT.

Moreover, we further use the Plug-and-Play (PnP) framework (Sun et al. 2022; Zhang et al. 2023b) to extend the proposed PLoRA mechanism so that it can be more flexible and easier to be deployed for the cold-start issue including both zero-shot and few-shot learning scenarios (Wu et al. 2023; Dathathri et al. 2020), as shown in Fig. 1(d). The zero-shot learning scenario occurs when new and anonymous users due to privacy considerations request a service. In this circumstance, directly applying personalized models to such a zero-shot learning scenario may degrade the prediction performance because the models have no idea about the new users (Zhang et al. 2023a). Conversely, the extended PLoRA can use personalized dropout (PDropout) (Srivastava et al. 2014) and mutual information maximization (MIM) (Krause, Perona, and Gomes 2010) to remove personalized information such that the model's generalization ability can be increased to handle anonymous new users.

[*]Corresponding authors.

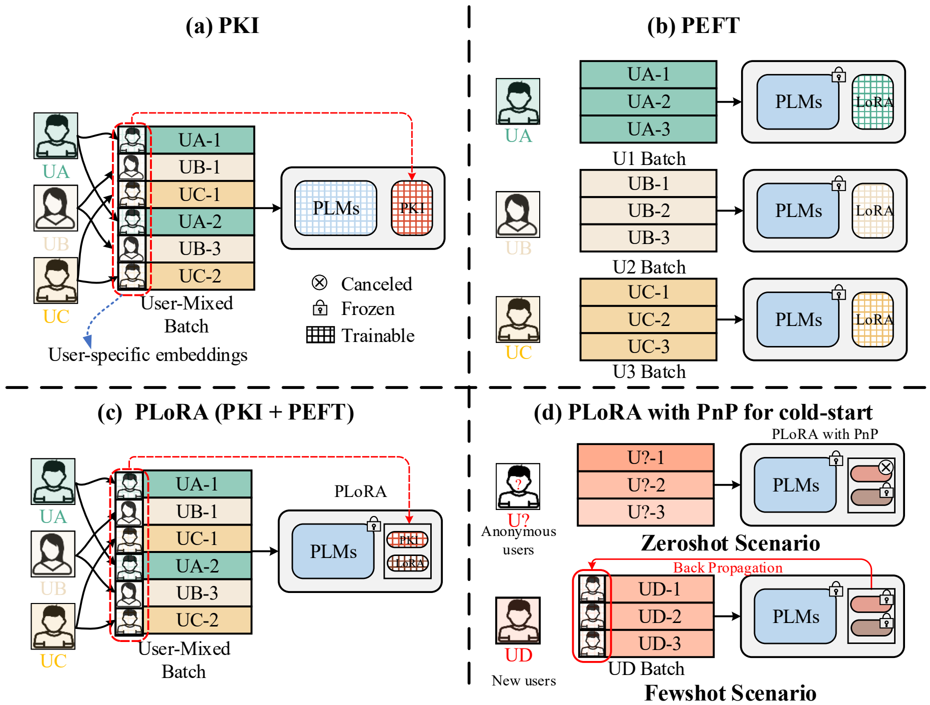

Figure 1: Different methods for human-centered text understanding.

The few-shot learning scenario occurs when a new user with a few samples requests a service. Unlike traditional methods that need to retrain the model, both PLoRA and PLMs parameters are frozen. The extended PLoRA is used as a knowledge extractor to produce personalized information for each new user via a backpropagation optimization, and then perform the PnP to complete online training and prediction.

Similar to LoRA, the proposed PLoRA does not increase additional sequence lengths when it handles text input and has no additional inference latency in comparison to other PEFT methods such as adapter (Pfeiffer et al. 2020), prefix tuning (Li and Liang 2021), Prompt-tuning (Lester, Al-Rfou, and Constant 2021), and P-tuning (Liu et al. 2021). Moreover, PLoRA is agnostic to neural networks and orthogonal to many prior methods, hence it can be easily deployed in various PLMs and combined with other technologies such as prompt-tuning and prefix-tuning (Houlsby et al. 2019; Guo, Rush, and Kim 2021). Experiments conducted on four benchmark datasets show that the proposed method outperforms existing methods in full/few/zero-shot learning scenarios for the HCTU task. The main contributions in this work are as follows.

- We proposed the PLoRA mechanism by combining the PKI and LoRA to inject personalized information into PLMs without full-model fine-tuning.
- We further extend PLoRA using the PnP framework to increase the model's generalization ability and enable online training and prediction for the code-start scenario.
- Experimental results show that PLoRA is effective, lightweight, easy-to-deploy in PLMs for HCTU tasks.

The remainder of this paper is structured as follows. Section 2 describes the details of the proposed method. Extensive experiments are analyzed in Section 3 and conclusions are finally drawn in Section 4. What's more, the technical appendix provided related work, further proofs, detailed settings and additional experiments.

Methodology

Problem Formulation

The HCTU task considers every text data sample belonging to a certain user and captures textual representations according to the user preferences where a collection of written text records is associated with each user. These records facilitate models to understand user preferences and provide personalization services. Consequently, HCTU can be formulated as follows. Given an input text $\mathbf{x} = [x_1, x_2, ..., x_N]$, the goal of human-centered models is to generate an output y for the user u, where N represents the input length and y associates task targets such as sentiment scores. These tasks can be modeled as $\arg\max q(y|\mathbf{x}, u)$. For each user u, a collection of data is denoted as $\mathcal{D}_u = \sum_i (\mathbf{x}_i, y_i, u)$ on which user preference can be generated.

To simulate cold-start issues including zero-shot and few-shot learning scenarios, a couple of data ($\mathcal{D}^{\mathrm{A}}$ and $\mathcal{D}^{\mathrm{B}}$) are provided for each dataset $\mathcal{D} = \sum_u \sum_i (\mathbf{x}_i, y_i, u) = \mathcal{D}^{\mathrm{A}} + \mathcal{D}^{\mathrm{B}}$ where $(\forall u^{\mathrm{A}} \in \mathcal{D}^{\mathrm{A}}) \cap (\forall u^{\mathrm{B}} \in \mathcal{D}^{\mathrm{B}}) = \emptyset$. $\mathcal{D}^{\mathrm{A}}$

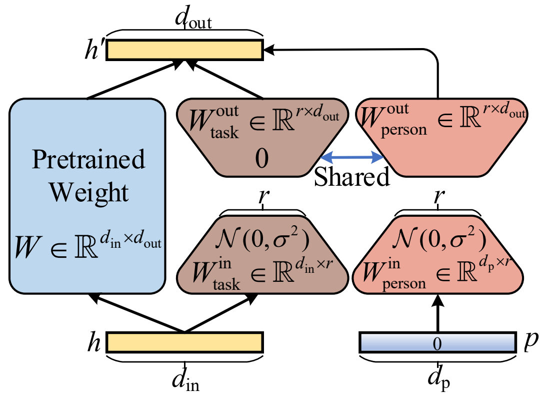

Figure 2: An illustration of PLoRA.

is used for full-shot learning and $\mathcal{D}^{\text{B}}$ is provided for cold-start evaluation. For zero-shot learning, it requires models $q(y|\mathbf{x}, u)$ learned from $\mathcal{D}^{\text{A}}$ to show superior generalization performance $q(y|\mathbf{x})$ on $(\mathbf{x}, y) \in \mathcal{D}^{\text{B}}$ without user u^{B} as input. For few-shot learning, the user preference $p = f(u)$ is learned from a few data $\mathcal{D}_u$ and then used for personalization services $q(y|\mathbf{x}, u)$ where $(\mathbf{x}, y, u) \in \mathcal{D}_u$ and $\mathcal{D}_u \in \mathcal{D}^{\text{B}}$.

Personalized LoRA

In this section, we elaborate on PLoRA for personalization in PLMs, as illustrated in Figure 2, which combines task-specific LoRA and user-specific PKI.

Task-specific LoRA A neural network such as PLMs contains many matrix multiplications. LoRA provides a PEFT that facilitates weight matrices to capture intrinsic rank for downstream tasks adaptation. For a weight matrix $W \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$ in PLM, lightweight matrices of $W^{\text{in}}_{\text{task}} \in \mathbb{R}^{d_{\text{in}} \times r}$ and $W^{\text{out}}_{\text{task}} \in \mathbb{R}^{r \times d_{\text{out}}}$ with a low rank of $r \ll \min(d_{\text{in}}, d_{\text{out}})$ are additionally applied to update W as $W + W^{\text{in}}_{\text{task}} W^{\text{out}}_{\text{task}}$. Given $h \in \mathbb{R}^{d_{\text{in}}}$ as input textual representation, the output $h' \in \mathbb{R}^{d_{\text{out}}}$ is calculated with task-specific LoRA via:

$$h' = hW + hW^{\text{in}}_{\text{task}} W^{\text{out}}_{\text{task}} \,. \tag{1}$$

In the training phase, W is fixed and only $W^{\text{in}}_{\text{task}}$ and $W^{\text{out}}_{\text{task}}$ are learned.

Personalization-specific PKI To adapt a generic model to personalization, most of the existing works are dedicated to incorporating personalized knowledge $p = f(u) \in \mathbb{R}^{d_p}$ with textual representation via:

$$h' = hW + pW_{\text{person}} \,, \tag{2}$$

where $f(u)$ means user preferences or embeddings of u. Unfortunately, such methods require an optimization of both W and W_{person} in an FFT method so as to coordinate with personalization and downstream tasks, making their deployment cumbersome in practice.

PLoRA To compose a parameter-efficient and personalized adapter in PLMs, PLoRA takes h and p as inputs and outputs h' via:

$$\begin{aligned} h' &= hW + hW^{\text{in}}_{\text{task}} W^{\text{out}}_{\text{task}} + pW_{\text{person}} \\ &= hW + hW^{\text{in}}_{\text{task}} W^{\text{out}}_{\text{task}} + pW^{\text{in}}_{\text{person}} W^{\text{out}}_{\text{person}} \\ &= hW + (hW^{\text{in}}_{\text{task}} + pW^{\text{in}}_{\text{person}}) W^{\text{out}}_{\text{task}} \end{aligned} \,. \tag{3}$$

Initially, a low-rank hypothesis is applied to PKI with $W_{\text{person}} = W^{\text{in}}_{\text{person}} W^{\text{out}}_{\text{person}}$ for capture intrinsic rank features of injection weights. To further facilitate the couples with task-specific adaptation and personalization, we share $W^{\text{out}}_{\text{task}}$ with $W^{\text{out}}_{\text{person}}$. Similar to LoRA, only $W^{\text{in}}_{\text{task}}$, $W^{\text{out}}_{\text{task}}$ (or $W^{\text{out}}_{\text{person}}$), and $W^{\text{in}}_{\text{person}}$ are trainable parameters that receive gradient updates.

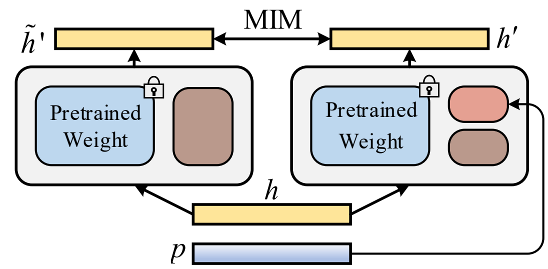

Figure 3: A diagram of MIM on PLoRA for zero-shot learning.

Plug-and-Play for Cold Start

Large-scale PLMs trained on a large number of corpora show excellent text-understanding capabilities. Instead of FFT or modifying model architecture for task-specific domains, the LoRA module can be regarded as a PnP module that guides PLMs to be adapted to target domains such as sentiment analysis. Similarly, PKI can also be considered as a PnP module that guides PLMs to serve different users. Note that, when PnP modules are applied, PLMs will not be trained and thus their architecture will not be modified and their pre-training power will not be deteriorated.

To build a PnP framework and apply PLoRA for complex cold-start issues, a reparameterization strategy is proposed. Especially, $W^{\text{in}}_{\text{task}}$ and $W^{\text{in}}_{\text{person}}$ are randomly in Gaussian distributions; $W^{\text{out}}_{\text{task}}$ and p are zeros. Hence, $W^{\text{in}}_{\text{task}} W^{\text{out}}_{\text{task}}$ is zero at the beginning of the training phase and $pW^{\text{in}}_{\text{person}} W^{\text{out}}_{\text{person}}$ is zero when PLMs first meet a user u. As a result, PLoRA is a PnP module that is orthogonal to PLMs and gradually grows with capabilities of guiding PLMs to be adapted to task-specific and human-centered domains.

Although high-performance PLoRA gains, it is still difficult for zero-shot learning and few-shot learning scenarios (See Section Experiments).

Zero-shot learning scenarios It is challenging in zero-shot learning since poor decomposition of generic and human-centered features makes PLoRA degrading performance for unseen or anonymous users. To address this problem, we propose PDropout and MIM methods to help

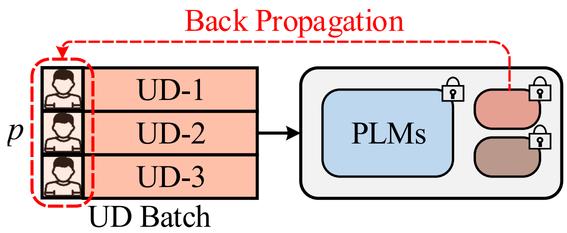

Figure 4: A diagram of PnP framework on few-shot learning scenarios for user D (UD).

PLoRA to remain its generalization performance for zero-shot learning scenarios.

- **PDropout**. During the training, we randomly mask out users in a batch of samples with a dropout ratio of $\omega \in [0, 1]$. Consequently, with the capability of adapting to personalization, the PLoRA framework remains accessible to generic task performance.
- **Mutual Information Maximization**. Inspired by (Zhang et al. 2023a), we introduce an MIM method to align the distance between task-specific and human-centered representations ($\tilde{h}'$ and h'), as shown in Figure 3. During the training, human-centered representation regarded as a teacher states guides better generic task-specific performance where MIM can be considered as a knowledge distillation mechanism. By contrast, MIM is also a regularization term to leverage generalization performance on human-centered representation. In practice, both mean square error (MSE) and Kullback-Leibler (KL) divergence (Kullback and Leibler 1951) are prevalent for instantiating MIM.

Few-shot learning scenarios For few-shot learning as shown in Figure 4, we fix all model parameters and inject a zeroed embedding p referring to a new user u. With a few samples associated with the user, HCTU models update the user preference p through a backpropagation mechanism so that the user could participate in personalized service with updated p as an access token.

Train and Inference

In principle, the proposed PLoRA can be applied to any subset of linear projectors in neural networks to provide PEFT and personalization. In this work, we limit our method to multi-head attention mechanisms (i.e., query and value projectors) in the Transformer structure for downstream tasks, following LoRA study (Hu et al. 2021).

For optimization in the training phase, the loss function of full-shot learning is defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{Fullshot}} &= \mathcal{L}_{\text{CE}} + \alpha \mathcal{L}_{\text{MIM}} \\ \mathcal{L}_{\text{CE}} &= \text{CE}(q(\tilde{y}|\mathbf{x}, \text{PDropout}(u, \omega); \theta), y) \\ \mathcal{L}_{\text{MIM}} &= \sum_{(\tilde{h}', h') \in q(\tilde{y}|\mathbf{x}, u)} \text{MIM}(\tilde{h}', h') \end{aligned} , \quad (4)$$

where $(\mathbf{x}, y, u) \in \mathcal{D}^{\text{A}}$ refers to each sample; $\text{PDropout}(\cdot)$ and $\text{MIM}(\cdot)$ are the proposed PDropout and MIM methods; θ presents lightweight trainable parameters; α is balance ratio for MIM term.

For few-shot learning, the training objective is formulated as:

$$\begin{aligned} f(u) &= \arg\min_{p=f(u)} \mathcal{L}_{\text{Fewshot}} \\ \mathcal{L}_{\text{Fewshot}} &= \sum_{(\mathbf{x}, y, u) \in \mathcal{D}^{\text{B}}} \text{CE}(q(\tilde{y}|\mathbf{x}, u; \hat{\theta}), y) \end{aligned} , \quad (5)$$

where $\hat{\theta}$ presents well-trained parameters in full-shot learning scenarios and updated $f(u)$ stand for the semantics of user u.

Similar to LoRA, the proposed PLoRA still has no additional inference latency in deployment. We can explicitly compute $W = W + W^{\text{in}}_{\text{task}} W^{\text{out}}_{\text{task}}$ and $b = b + f(u) \cdot W^{\text{in}}_{\text{person}} W^{\text{out}}_{\text{person}}$ to update a linear projector for personalized services for the user u where the linear projector contains weight matrix W and bias vector b. When we need to switch to pure (or generic) task-specific tasks or anonymous users, we can recover b by subtracting $f(u) \cdot W^{\text{in}}_{\text{person}} W^{\text{out}}_{\text{person}}$. Furthermore, we can also add different $f(u') \cdot W^{\text{in}}_{\text{person}} W^{\text{out}}_{\text{person}}$ terms to switch services for the user u'. As a consequence, PLoRA can also switch to other tasks with very little memory overhead.

Experiments

To investigate the effectiveness and efficiency of the proposed methods for HCTU, extensive experiments were conducted and analyzed on personalized sentiment analysis.

Datasets and Evaluation

Datasets We took personalized sentiment analysis as a test bed since sentiment is a critical and sensitive evaluation for human subjective expression. The used datasets include IMDB, YELP, GDRD, and PPR, where a collection of data is associated with different users. To simulate complex and practical situations in real-world applications, all datasets are individually divided into two parts $\mathcal{D}^{\text{A}}$ and $\mathcal{D}^{\text{B}}$ where $\mathcal{D}^{\text{A}}$ contains much larger samples than $\mathcal{D}^{\text{B}}$ and $\mathcal{D}^{\text{B}}$ aims to simulate cold-start scenarios, as formulated in Section Methodology. Either $\mathcal{D}^{\text{A}}$ or $\mathcal{D}^{\text{B}}$, it splits into train, dev, and test data for experiments. More detailed statistics of datasets were listed in Appendices.

Evaluation To measure the effectiveness, Accuracy (Acc), mean squared error (MSE), and macro F_1 (F_1) were adopted as evaluation metrics. Note that, higher F_1 and Acc (%) and lower MSE mean better results.

Experimental Setup

To perform personalized sentiment analysis using PLMs, we applied PLoRA to multi-head attention including query and value projectors, following (Hu et al. 2021), where introduced PLMs involved BERT ($\mathcal{B}$) (Devlin et al. 2019), RoBERT ($\mathcal{R}$) (Liu et al. 2019), and Flan-T5 ($\mathcal{FT}5$) (Chung et al. 2022). For the reproduction of experiments, more implementation details of hyperparameters were reported in Appendices.

Datasets	IMDB-A			YELP-A			GDRD-A			PPR-A			TP(%)
	Acc	MSE	F_1	Acc	MSE	F_1	Acc	MSE	F_1	Acc	MSE	F_1	
Generic sentiment analysis													
NSC	47.7	1.729	44.7	65.2	0.452	62.6	-	-	-	-	-	-	100x
BERT	$\underline{51.7}$	1.398	$\underline{50.3}$	67.6	$\underline{0.390}$	65.1	$\underline{57.0}$	$\underline{0.582}$	$\underline{53.7}$	$\underline{83.1}$	$\underline{0.192}$	45.7	100x
$\mathcal{B}$-LoRA	51.0	$\underline{1.349}$	48.9	$\underline{67.8}$	0.400	$\underline{66.2}$	56.8	0.590	53.2	82.6	0.198	$\underline{47.8}$	2.1x
$\mathcal{R}$-LoRA	**52.5**	1.252	50.5	69.6	0.354	68.9	58.4	0.566	56.0	83.6	0.188	**55.5**	2.3x
$\mathcal{FT}$5-LoRA	**52.5**	**1.218**	**50.8**	**70.0**	**0.352**	**69.3**	**58.1**	**0.553**	**55.9**	**84.5**	**0.168**	53.8	2.7x
Human-centered sentiment analysis													
NSC+U	51.1	1.460	47.1	67.3	0.444	64.9	-	-	-	-	-	-	100x
$\mathcal{B}$-MAA	55.4	$\underline{1.129}$	53.4	71.6	0.352	69.8	63.6	0.427	58.3	84.5	0.173	41.5	150x
$\mathcal{B}$-PKI	55.6	1.200	53.2	70.1	0.377	68.5	64.8	0.470	60.8	84.7	0.173	48.3	2.1x
$\mathcal{B}$-UserAdapter	55.7	1.131	53.5	70.8	0.357	68.5	64.7	0.465	60.0	84.4	0.175	48.0	100x
$\mathcal{B}$-PLoRA$_{2S}$	54.5	1.246	52.0	70.7	0.362	68.9	63.9	0.493	59.9	84.5	0.174	47.8	3.1x
$\mathcal{B}$-PLoRA	$\underline{57.0}$	1.151	$\underline{54.9}$	$\underline{72.1}$	$\underline{0.338}$	$\underline{70.3}$	$\underline{65.0}$	$\underline{0.463}$	$\underline{61.4}$	$\underline{85.1}$	$\underline{0.168}$	$\underline{49.8}$	3.1x
$\mathcal{R}$-PLoRA	58.0	1.008	55.8	72.9	0.318	71.3	65.8	0.457	61.8	85.5	0.158	**64.1**	3.2x
$\mathcal{FT}$5-PLoRA	**58.7**	**0.980**	**56.1**	**73.3**	**0.314**	**71.5**	**66.7**	**0.447**	**63.3**	**86.7**	**0.144**	60.4	3.8x

Table 1: Comparative test results on full-shot learning scenarios. TP presents a trainable parameter (including user embeddings) ratio during the optimization. All figures are averaged over five runs. The underscored and black-face figures mean the best scores for only $\mathcal{B}$ and all experiments in each group, respectively.

Datasets	IMDB-B			YELP-B			GDRD-B			PPR-B			TP(%)
	Acc	MSE	F_1	Acc	MSE	F_1	Acc	MSE	F_1	Acc	MSE	F_1	
Few-shot learning													
$\mathcal{B}$-PLoRA*	50.2	1.722	47.3	67.1	0.479	64.8	58.5	0.632	54.8	83.7	0.188	44.3	2.9x
$\mathcal{B}$-MAA (FS)	54.0	1.652	51.8	71.0	0.379	68.9	63.8	0.535	59.4	84.0	0.188	40.2	0.2x
$\mathcal{B}$-UserAdapter (FS)	52.9	$\underline{1.471}$	50.9	70.9	0.383	69.2	**$\underline{65.2}$**	**$\underline{0.502}$**	60.0	83.8	0.168	$\underline{44.8}$	0.2x
$\mathcal{B}$-PLoRA$_{2S}$ (FS)	54.2	1.641	52.1	70.1	0.410	68.1	61.6	0.611	57.1	84.1	0.168	42.6	0.1x
$\mathcal{B}$-PLoRA (FS)	$\underline{55.1}$	1.559	$\underline{53.5}$	$\underline{71.2}$	$\underline{0.362}$	$\underline{69.3}$	64.0	0.536	**$\underline{60.5}$**	$\underline{84.5}$	$\underline{0.164}$	43.7	0.1x
$\mathcal{R}$-PLoRA (FS)	56.5	1.541	54.7	72.6	0.355	71.3	63.9	0.528	59.8	84.7	0.171	57.8	0.1x
$\mathcal{FT}$5-PLoRA (FS)	**58.0**	**1.396**	**55.8**	**73.2**	**0.338**	**72.3**	63.1	0.514	59.6	**86.1**	**0.148**	**59.4**	0.1x
Zero-shot learning													
$\mathcal{B}$-LoRA*	$\underline{47.2}$	2.232	44.4	65.0	0.497	62.7	52.3	0.687	48.3	81.4	0.218	34.5	2.1x
$\mathcal{B}$-UserAdapter	44.3	2.123	41.9	65.6	0.418	65.1	53.2	0.681	51.5	82.1	0.185	$\underline{43.6}$	-
$\mathcal{B}$-PLoRA$_{2S}$	44.1	1.889	41.6	$\underline{67.3}$	0.416	65.4	53.5	0.676	51.5	82.3	0.183	41.5	-
$\mathcal{B}$-PLoRA (ZS)	46.9	$\underline{1.632}$	$\underline{45.1}$	$\underline{67.3}$	$\underline{0.412}$	$\underline{65.8}$	$\underline{56.2}$	$\underline{0.640}$	$\underline{53.7}$	$\underline{83.3}$	$\underline{0.180}$	42.8	-
$\mathcal{R}$-PLoRA (ZS)	47.6	1.553	46.0	68.6	0.383	67.5	**56.6**	0.630	**55.4**	83.1	0.187	53.7	-
$\mathcal{FT}$5-PLoRA (ZS)	**48.1**	**1.521**	**46.5**	**70.7**	**0.352**	**68.8**	54.3	**0.618**	52.8	**84.7**	**0.159**	**57.4**	-

Table 2: Comparative test results on few/zero-shot learning scenarios. * presents corresponding models optimized only from $\mathcal{D}^{\mathrm{B}}$. FS means corresponding models are first learned in $\mathcal{D}^{\mathrm{A}}$ and are then adapted to $\mathcal{D}^{\mathrm{B}}$ in a few-shot learning strategy, i.e., Eq (5). ZS denotes corresponding models optimized with PDropout or MIM in $\mathcal{D}^{\mathrm{A}}$ are directly applied for $\mathcal{D}^{\mathrm{B}}$ without user inputs.

We compare our methods with the previous high-performance models, including neural sentiment classification (NSC) (Chen et al. 2016), multi-attribute attention (MAA) from MA-BERT (Zhang et al. 2021), and UserAdatper (Zhong et al. 2021). Moreover, several comparative methods derived from our motivations were adopted, including PKI (only PKI tuned in optimization), LoRA, and two-stage (2S). Note that 2S means human-centered models were learned from generic data in advance and then adapted to various users via the few-shot learning strategy in Eq. (5) so that updated models were well capable of providing generic and personalized services.

Comparative Results and Analysis

Full-shot learning Table 1 reported comparative results for both generic (the first group) and human-centered (the second group) scenarios in $\mathcal{D}^{\mathrm{A}}$. From the first group, applying the LoRA method to PLMs facilitated general language knowledge learned from large corpora to be adapted to sentiment analysis tasks with a few trainable parameters. Accordingly, LoRA-based PLMs achieved comparable performance in comparison to FFT models.

Against the first group, models in the second group gained better results on three metrics. This is because the introduction of personalized knowledge helps these models to accurately locate users' implicate sentiments where sentiment preferences from different users might differ. Based on the same PLMs, i.e., BERT, the proposed method of $\mathcal{B}$-PLoRA was on par with previous best-performance models, i.e., $\mathcal{B}$-MAA and $\mathcal{B}$-UserAdapter concerning their effectiveness. However, MAA and UserAdapter necessitate FFT optimization for task adaptations. Moreover, $\mathcal{B}$-PKI showed relatively lower scores, indicating FFT is essential for PLM

Models	IMDB-A	IMDB-B		YELP-A	YELP-B	
	Full-shot	Few-shot	Zero-shot	Full-shot	Few-shot	Zero-shot
$\mathcal{B}$-PLoRA	54.9	53.5	45.1	70.3	69.3	66.1
w/o PKI	48.9	-	41.6	66.2	-	65.4
w/o LoRA	53.2	52.0	-	68.5	67.7	-
w/o PDropout	54.2	51.4	45.0	69.8	68.9	65.7
w/o MIM	53.9	51.7	44.0	69.4	68.6	65.8
w/o PDropout & MIM	54.2	51.7	41.9	69.2	68.1	59.9

Table 3: Ablation study of test F_1 scores on IMDB and YELP with respect to full/few/zero-shot learning scenarios.

adapted to downstream tasks. By contrast, the proposed model is more efficient due to the dynamic combination between LoRA and PKI. Not only encoder-based PLMs, but decoder-based T5 can load PLoRA for downstream tasks, revealing an easy-to-deploy capability of PLoRA on wide applications.

In real-world applications, cold-start issues are serious in personalized services due to user-specific domain adaptation and data sparsity. We simulated cold-start issues as few-shot learning and zero-shot learning scenarios and conducted corresponding experiments, as reported in Table 2.

Few-shot Learning To make human-centered models capable of serving on unseen users where the users were from $\mathcal{D}^{\mathrm{B}}$ while out of $\mathcal{D}^{\mathrm{A}}$, we tested the proposed PLoRA and previous works via the introduced few-shot learning strategy from Eq. (5). From the table, it can be found that these models outperformed $\mathcal{B}$-LoRA* that was directly optimized on datasets $\mathcal{D}^{\mathrm{B}}$ in a full-shot way since these models updated from $\mathcal{D}^{\mathrm{A}}$ in advance could store robust performance and then be easily deployed to unseen users with a few samples.

Zero-shot Learning Theoretically, PKI-based methods with sophisticated structures, i.e., NSC+U and $\mathcal{B}$-MAA, were hard to directly handled pure textual data as they were difficult to distill pure textual representation from human-centered features (Zhang et al. 2023a). Fortunately, UserAdatper could directly discard user tokens at input layer and performed well when they met unseen users in zero-shot learning scenarios. $\mathcal{B}$-PLoRA$_{2S}$ separately and step-by-step optimized task-specific and use-specific plugins so that it competitively performed in both zero-shot learning and few-shot learning scenarios. However, lacking mutual learning between LoRA (for task adaptation) and PKI (for personalization) made $\mathcal{B}$-PLoRA$_{2S}$ method underperform our proposed PLoRA (ZS) that applies PDropout and MIM strategies to decouple task-specific and user-specific knowledge in the full-shot learning scenarios.

Model Analysis

Ablation study Results of an ablation study on IMDB and YELP were reported in Table 3, where the ablation target included: 1) PKI and LoRA to investigate the effectiveness of PLoRA in task-specific and user-specific knowledge fusion; 2) PDropout and MIM to validate the PnP performance of knowledge decoupling. From the first group in the table, the performance of PLoRA degraded with the elimination of PKI or LoRA, demonstrating that both task-specific and user-specific adaptations were crucial for the deployment of PLMs in HCTU applications.

From the last group, it can be found that, without PDropout or MIM, PLoRA slightly dropped F_1 score in all three scenarios; while it performed catastrophic descents in zero-shot learning scenarios when neither PDropout nor MIM was adopted. This is because PLoRA have coupled task-specific and user-specific knowledge in the full-shot learning procedure, leading to inferior performance of uncoupling task-specific knowledge for generic sentiment analysis if without zero-shot learning strategies. This phenomenon also indicated that either PDropout or MIM could facilitate the proposed PLoRA effectively adapted to generic scenarios. Note that, a dynamic combination of PDropout and MIM is beneficial for better adaptation.

Sparsity To reveal the robustness of PLoRA, we explored its dataset B-dev Acc in few-shot learning scenarios with different degrees of data sparsity or the number of few-shot samples in B-training sets, as shown in Figure 5. It can be found that the improvements of PLoRA increased with the number of shot samples increasing or the data sparsity degrees decreasing. For both our used IMDB and YELP datasets, 15-shot learning relatively gained competitive dev performance. It indicated that, with sufficient user-oriented data, the proposed method could extract user-specific knowl-

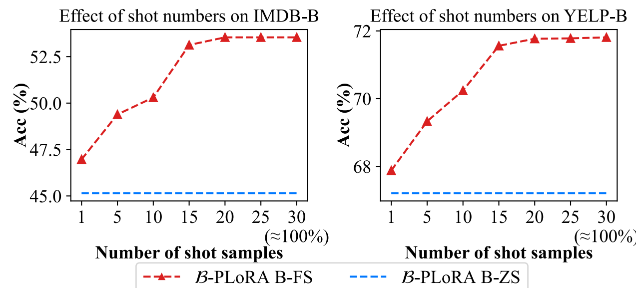

Figure 5: Effect of PLoRA with different data sparsity. FuS, FS, and ZS in figures (including the following figures) means using full/few/zero-shot learning methods, respectively. $\{\mathcal{M}\}$-PLoRA $\{$A/B$\}$-$\{$FuS/ZS/FS$\}$ corresponds dev figures of PLM $\mathcal{M}$-based PLoRA applied for datasets (A or B) with FuS/ZS/FS methods. $\approx 100\%$ means almost the full training dataset in $\mathcal{D}^{\mathrm{B}}$ is used and also presents no data sparsity for every user in $\mathcal{D}^{\mathrm{B}}$.

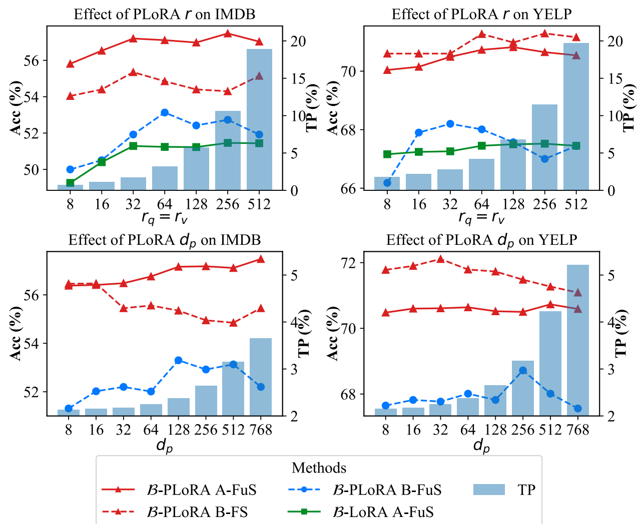

Figure 6: Performance on dev datasets with diverse parameters of PLoRA.

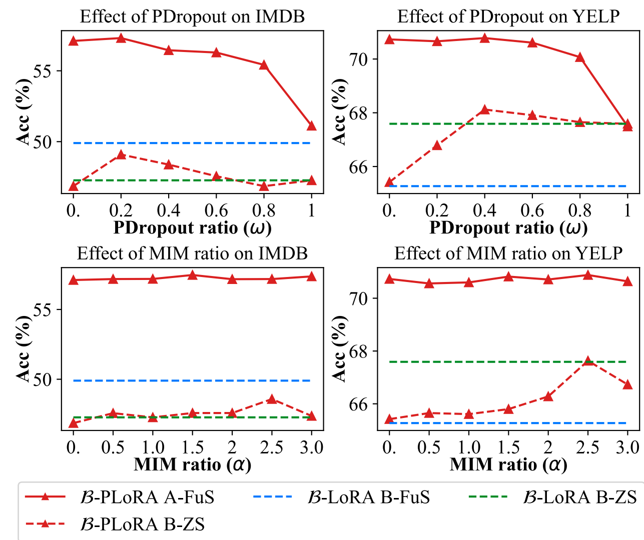

Figure 7: Performance on dev datasets with diverse parameters of PDropout and MIM. Note that, one of them is investigated with the other being not applied.

edge which then helps sentiment models to be adapted to such users in turn. Even in a few samples used (such as 1, 5, and 10), PLoRAs with the few-shot learning strategy (i.e., $\mathcal{B}$-PLoRA B-FS) still outperformed those (i.e., $\mathcal{B}$-PLoRA B-ZS) that were directly applied to zero-shot learning scenarios, indicating the effect of the few-shot learning strategy.

The sensitivity of hyperparameters. To further investigate how each component of PLoRA affects the final performance, we conducted several experiments on the sensitivity of hyperparameters, as shown in Figure 6-7.

Among the four subfigures in Figure 6, lower rank dimensionalities of applied weights and lower user embeddings for user semantics would not support complete performance in all scenarios. By contrast, larger figures of PLoRA configuration might saturate model performance and desire exponential-enlarged trainable parameters. It indicated that, in practice, empirical explorations of an appropriate PLoRA configuration and considerations of limited compute budget were critical for satisfactory services.

To further investigate how PDropout and MIM affect the effectiveness in zero-shot learning scenarios, we conducted analytical experiments in Figure 7. From the figure, it can be first found that, without PDropout nor MIM ($\omega = \alpha = 0$), $\mathcal{B}$-PLoRA B-ZS did not perform well in zero-shot scenarios, even worse than $\mathcal{B}$-LoRA B-ZS (green dashed lines or $\omega = 1$) that was learned from dataset $\mathcal{D}^{\mathrm{A}}$ without PKI and was then directed applied for dataset $\mathcal{D}^{\mathrm{B}}$ in zero-shot scenarios. Secondly, with appropriate configurations of ω and α, $\mathcal{B}$-PLoRA could effectively mitigate the above degradation, in consistent with the analysis in Table 3.

Discussions

In summary, the priority of the proposed PLoRA was located at 1) a dynamic combination of PEFT-based LoRA and knowledge-injected PKI for task-specific and human-centered adaptation; 2) an introduction of the few-shot learning strategy to adapt well-trained PLMs to unseen users in the training phase; 3) a PnP framework that couples and decouples task-specific and human-centered knowledge for cold-start issues.

We conducted experiments with PLMs including BERT, RoBERTa, and Flan-T5, which does not rashly verify that PLoRA could be only applied to these models. As discussed in Section Methodology, the proposed PLoRA can be adapted to a broader range of neural networks as long as containing linear projectors. Note that we do not want to limit the capability of our work to only sentiment analysis. Instead, our explorations could be extended to text-generative tasks and much broader applications as we introduced the text-generative paradigm of Flan-T5 to handle classification tasks. It sheds light on the promising future direction of PEFT and human-centered reasoning in large-scale language models.

Conclusions

We proposed PLoRA, a human-centered PEFT approach that successfully demonstrated the effectiveness in enhancing transfer learning from pre-training to downstream tasks for PLMs. By adopting a PnP framework, PLoRA significantly improves its adaptative ability to the cold-start issues in real-world applications. The experiments conducted on diverse personalized sentiment analysis tasks validated the effectiveness and efficiency of our method. Moreover, this work not only contributes to the improvements of the performance of PLMs on text understanding tasks but also sheds light on future works, including the explorations of PLoRA on LayerNorm layer, CNN, and other hybrid components, and the service extension of applications.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61966038 and 62266051, the Yunnan Ten Thousand Talents Program and Yunling Scholars Special Project under Grant YNWR-YLXZ-2018-022, the Ministry of Science and Technology (MOST), Taiwan, ROC, under Grant No. MOST110-2628-E-155-002, and the Yunnan Postdoctoral Science Foundation under Grant No. C615300504048. The authors would like to thank the anonymous reviewers for their constructive comments.

References

Capel, T.; and Brereton, M. 2023. What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2023)*, 1–23. ISBN 9781450394215.

Chen, H.; Sun, M.; Tu, C.; Lin, Y.; and Liu, Z. 2016. Neural Sentiment Classification with User and Product Attention. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2016)*, 1650–1659.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Castro-Ros, A.; Pellat, M.; Robinson, K.; Valter, D.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V.; Huang, Y.; Dai, A.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2022. Scaling Instruction-Finetuned Language Models. *arXiv eprint arXiv: 2210.11416*.

Crisan, A.; Drouhard, M.; Vig, J.; and Rajani, N. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FACCT 2022)*, 427–439. ISBN 9781450393522.

Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; and Liu, R. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *International Conference on Learning Representations (ICLR 2020)*.

Devlin, J.; Chang, M. W.; Lee, K.; and Toutanova, K. 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 (NAACL-HLT 2019)*, 4171–4186.

Guo, D.; Rush, A.; and Kim, Y. 2021. Parameter-Efficient Transfer Learning with Diff Pruning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, 4884–4896.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (ICML-2019)*, volume 97, 2790–2799. PMLR.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR 2021)*.

Krause, A.; Perona, P.; and Gomes, R. 2010. Discriminative Clustering by Regularized Information Maximization. In Lafferty, J.; Williams, C.; Shawe-Taylor, J.; Zemel, R.; and Culotta, A., eds., *Advances in Neural Information Processing Systems (NIPS)*, volume 23. Curran Associates, Inc.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP-2021)*, 3045–3059.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP-2021)*, 4582–4597.

Liu, P.; Zhang, L.; and Gulla, J. A. 2023. Pre-train, Prompt and Recommendation: A Comprehensive Survey of Language Modelling Paradigm Adaptations in Recommender Systems. *arXiv eprint arXiv:2302.03735*.

Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; and Tang, J. 2021. GPT Understands, Too. *arXiv eprint arXiv: 2103.10385*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Lynn, V.; Son, Y.; Kulkarni, V.; Balasubramanian, N.; and Schwartz, H. A. 2017. Human Centered NLP with User-Factor Adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, 1146–1155. Stroudsburg, PA, USA.

Min, B.; Ross, H.; Sulem, E.; Veyseh, A. P. B.; Nguyen, T. H.; Sainz, O.; Agirre, E.; Heinz, I.; and Roth, D. 2021. Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey. *arXiv eprint arXiv:2111.01243*.

Pfeiffer, J.; Rücklé, A.; Poth, C.; Kamath, A.; Vulić, I.; Ruder, S.; Cho, K.; and Gurevych, I. 2020. AdapterHub: A Framework for Adapting Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP-2020)*, 46–54.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of machine learning research*, 15(1): 1929–1958.

Sun, T.; Shao, Y.; Qian, H.; Huang, X.; and Qiu, X. 2022. Black-Box Tuning for Language-Model-as-a-Service. In *Proceedings of the 39th International Conference on Machine Learning (ICML-2022)*, volume 162, 20841–20855.

Tang, D.; Qin, B.; and Liu, T. 2015. Learning Semantic Representations of Users and Products for Document Level Sentiment Classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-2015)*, 1014–1023.

Wu, Y.; Xie, R.; Zhu, Y.; Zhuang, F.; Zhang, X.; Lin, L.; and He, Q. 2023. Personalized Prompt for Sequential Recommendation. *arXiv eprint arXiv:2205.09666*.

Wu, Z.; Dai, X. Y.; Yin, C.; Huang, S.; and Chen, J. 2018. Improving review representations with user attention and product attention for sentiment classification. In *Proceedings of The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 5989–5996. ISBN 9781577358008.

Zhang, Y.; Wang, J.; Yu, L.-C.; Xu, D.; and Zhang, X. 2023a. Domain Generalization via Switch Knowledge Distillation for Robust Review Representation. In *Findings of the Association for Computational Linguistics (ACL-2023)*, 12812–12826.

Zhang, Y.; Wang, J.; Yu, L.-C.; and Zhang, X. 2021. MA-BERT: Learning Representation by Incorporating Multi-Attribute Knowledge in Transformers. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP 2021)*, 2338–2343.

Zhang, Y.; Wang, J.; and Zhang, X. 2021. Personalized sentiment classification of customer reviews via an interactive attributes attention model. *Knowledge-Based Systems*, 226: 107135.

Zhang, Z.; Zeng, Z.; Lin, Y.; Wang, H.; Ye, D.; Xiao, C.; Han, X.; Liu, Z.; Li, P.; Sun, M.; and Zhou, J. 2023b. Plug-and-Play Knowledge Injection for Pre-trained Language Models. *arXiv eprint arXiv:2305.17691*.

Zhong, W.; Tang, D.; Wang, J.; Yin, J.; and Duan, N. 2021. UserAdapter: Few-shot user learning in sentiment analysis. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP-2021)*, 1484–1488.