Cross-lingual Social Misinformation Detector based on Hierarchical Mixture-of-Experts Adapter

Haofang Fan[1]*, Xiran Hu[2]*, Geng Zhao[2]*†
[1]Interdisciplinary Center for Scientific Computing, Heidelberg University
haofang.fan@stud.uni-heidelberg.de
[2]Department of Computational Linguistics, Heidelberg University, Heidelberg, Germany
xiran.hu@stud.uni-heidelberg.de, zhao@cl.uni-heidelberg.de

Abstract

The spread of social misinformation has been a global concern, particularly affecting non-native speaker users who are more susceptible to misinformation on foreign social media platforms. In light of this, this study focuses on mitigating the challenges faced by social misinformation detectors in quickly regaining their capability after crossing linguistic borders, especially for non-native users with only monolingual social media histories. By integrating sentiment analysis as an auxiliary, less sensitive task, we transform the challenging cross-lingual transfer into a manageable multi-task framework. Then, we propose HierMoE-Adpt, a novel, cost-effective, parameter efficient fine-tuning method based on hierarchical mixture-of-experts adaptation, to enhance cross-lingual social misinformation detection. HierMoE-Adpt includes a hierarchical routing strategy and an expert-mask mechanism, effectively merging knowledge about understanding posts in a new language with misinformation detection capabilities, contributing to the recovery of personal misinformation detectors' performance in sync with the dynamics of international travel.

1 Introduction

The spread of misinformation in the wake of breaking news is a global phenomenon that poses significant challenges for non-native speakers on social media platforms (Shu et al., 2017; Wu et al., 2019), leading to serious misleading effects. Previous studies indicate that a considerable proportion of international travelers and expatriates feel more susceptible to being influenced or deceived by local social media and news upon landing in a new country (Pérez-Rosas and Mihalcea, 2014; Levitan et al., 2015). Unfortunately, for many such sensitive audiences, a significant portion of the social media platforms and mobile devices they use operate primarily in their native languages. As a result, the built-in social misinformation detectors on these platforms and devices tend to rely heavily on and be limited to achieving accurate veracity checking and flagging against content in user's native language (Wen et al., 2018; Chu et al., 2021). Therefore, it is crucial to develop automatic approaches that enable detectors trained on monolingual misinformation datasets to efficiently adapt to new linguistic environments, allowing social misinformation detectors to cross national borders in sync with the movement of individuals. Furthermore, minor language communities on international social media platforms are also susceptible to misinformation (Kwon et al., 2016; Cruz et al., 2020; Yang et al., 2021), such as the notably high rates of propagation of rumors and fraud in the Korean or Simplified Chinese Twitter communities. These phenomena demonstrate the demand for cross-lingual detectors on global SNS, both when users physically traverse international borders or engage virtually.

Figure 1: Let the built-in social misinformation detector cross the language borders, following your physical or virtual footprints.

Existing works attempt to enhance the cross-lingual capabilities of detectors by constructing multilingual misinformation datasets (Nielsen and McConville, 2022; Gupta and Srikumar, 2021). However, collecting sufficient data of this kind is costly. Besides, some related studies (Tian et al.,

*All authors contributed equally to this work.
†Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 7253–7265
January 19–24, 2025.

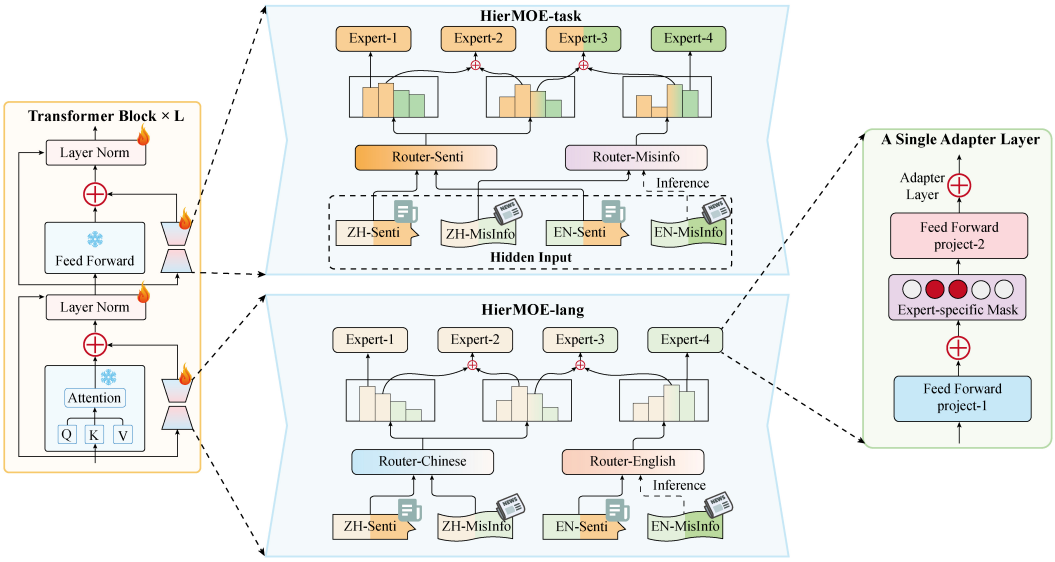

Figure 2: The framework of HierMoE-Adpt(middle), with illustration of Expert-Masking Mechanism(right).

2021; Du et al., 2021) directly fine-tune pre-trained language models (PLMs) (Devlin et al., 2018) on zero-shot/few-shot social misinformation detection tasks. However, these studies merely regard the problem as a text classification task with a heavy end-task-specific fine-tuning stage. This method deviates from the inherent target of adaptation, as it can potentially reduce models' effectiveness for related tasks for which the models were not specifically fine-tuned. Moreover, it is prone to catastrophic forgetting of the model's capability to detect misinformation in the source language environment. Furthermore, it involves unfreezing and reconfiguring the base model's pre-trained components, such as multi-head attention (MHA) or feed-forward networks (FFN), which may not be suitable for lightweight adaptation.

From a broader perspective, however, recent works have developed various novel adapters for cross-lingual and/or cross-task model adaptation. In the natural language understanding (NLU) domain, sentiment analysis, as it has rich open-sourced data resources across all major languages, is introduced by a number of researchers as an auxiliary task(Bhutani et al., 2019; Zhou et al., 2021; Dementieva and Panchenko, 2021), alongside their primary or objective tasks to conduct multi-task learning. This beneficial process is called "external knowledge sharing". Interestingly, this form of "external knowledge sharing" exactly closely aligns with research findings in the field of multilingual language models: under a cross-lingual multi-task setting, combining languages and tasks that the model has encountered at least once (into an unseen combination) for inference leads to better adaptation effects and improved few-shot performance compared to hard cross-lingual transfer in a single-task setting.

Thus, intuitively, we consider using sentiment classification datasets as a cost-effective auxiliary resource to aid models in adapting to target languages. During the model adaptation process, we hope to design a novel PEFT(parameter-efficient fine-tuning adapter) that successfully learns knowledge relevant to understanding the target language (e.g., writing styles) as well as knowledge specific to misinformation detection, thereby achieving promising results in nearly zero-shot misinformation detection in the target language environment. We propose HierMoE-Adpt[1] (**Hierarchical Mixture-of-Expert Adapter for Cross-Lingual Social Misinformation Detection**), a novel hierarchical mixture-of-experts PEFT method for cross-lingual adaptation of social misinformation detectors.

Specifically, HierMoE-Adpt uses a source language social misinformation dataset and bilingual sentiment analysis datasets as input resources for this simplified multi-language multi-task setting. It innovatively introduces a soft hierarchical routing strategy to facilitate layered learning of "content understanding" and "task adaptation", allowing each expert to preferentially learn language or task knowledge. Through information sharing, our

[1]The code will be released at https://github.com/remake-dark/HierMoE-Adapt

method not only learns essential specialized knowledge but also common knowledge in cross/multilingual content understanding and multi-task adaptation. Moreover, we propose a continual-learning styled expert-mask method to enhance the personalization of experts, encouraging the early emergence of experts proficient in target language understanding and misinformation detection. During the inference period, the well-trained HierMoE-Adpt can be used for approximate zero-shot or few-shot flagging of social misinformation instances in the target language. **Our main contributions are as follows**:

• We propose HierMoE-Adpt, a novel PEFT model-agnostic adapter for cross-lingual adaptation of social misinformation detectors, demonstrating its performance advantage over baselines through approximate zero-shot and few-shot experiments.

• To the best of our knowledge, HierMoE-Adpt is the first adaption method dedicated to cross-lingual social misinformation detection.

2 Related Works

2.1 Social Misinformation Detection

Generally, social misinformation detection (SMD) conducts binary classification to distinguish between real and fake news or posts on SNS. The technical routine is also shared by rumor detection and fake news detection. Among these, content-based SMD is a sub-field commonly used to identify fake or inaccurate posts on social media platforms, treating all textual elements of an item as inputs without considering any given social contexts or attributes (Sheng et al., 2021). To alleviate the propagation of misinformation, in the early stages, some complex, high-budget but well-performing ideas were proposed. MDFEND (Nan et al., 2021) focuses on multi-domain challenge, using domain gates to design a benchmark for judgment, addressing difficulties in multi-domain scenarios. Visual features of suspected posts, particularly visual entities, e.g. landmarks, have also been incorporated into SMD studies (Qi et al., 2021). Moreover, some works (Giachanou et al., 2019; Zhang et al., 2021) achieve the perception of the emotions expressed by audiences by extracting long-term emotional signals and using dual emotion.

However, with the rising demand for addressing complex scenarios and domain shifts in recent years, more challenges, e.g. robustness, generalisability, cross-lingual capability, have emerged. Nielsen (Nielsen and McConville, 2022; Gupta and Srikumar, 2021) focus on bulding multilingual social misinformation datasets as a benchmark. To enhance the ability of cross-lingual SMD, Du (Du et al., 2021) proposes a novel framework that jointly encodes cross-lingual news texts and uses factual information from one language to detect misinformation in another language. Chu (Chu et al., 2021) compares the differences in textual features between Chinese and English, then test the applicability of cross-lingual models. Some works (Lin et al., 2023; Wu et al., 2023) focus on using prompt learning to address cross-lingual issues, alleviating the challenges posed by low annotation rates and achieving efficient prevention in the early stages of misinformation spread. However, methods allowing zero/few-shot settings and lightweight training remain rare (Lin et al., 2023; Tian et al., 2021; Panda and Levitan, 2022; Ozcelik et al., 2023).

2.2 Cross-lingual Adaption for Language Models

In recent times, numerous Pretrained Language Models (PLMs) and Large Language Models (LLMs) trained on multilingual corpora have demonstrated impressive multilingual capabilities (Pires et al., 2019; Lai et al., 2023). However, the real-world application and bootstrapping of NLP models in specialized tasks with insufficient end-task data continue to pose significant challenges (Wang et al., 2019). In this paper, our main task can essentially be regarded as a branch of multi-task cross-lingual transfer (Schuster et al., 2018; Pfeiffer et al., 2020), where parameter-efficient fine-tuning (PEFT) methods, known for their high parameter efficiency and modularity, emerge as methodologies. Many previous works have demonstrated competitive performance in this scenario. For instance, P-tuning (Liu et al., 2022) improves the model's ability to understand the attributes of inputs and context relative to task demands. HyperLoRA (Xiao et al., 2023) deploys hypernetworks in LoRA (Hu et al., 2021) to enhance transfer between closely-related or mutually-intelligible languages. Empirically, in task scenarios still employing PLMs as the backbone model, adapters remain the status of the most practical technical approach: Bottleneck Adapter and Parallel Adapter (Houlsby et al., 2019; He et al., 2021) introduce the concept of adapters, which can be uniformly described as "tiny feed-forward layers inserted after pre-trained MHAs and

FFNs." While these classical methods are widely applicable, they exhibit disadvantages regarding the disentanglement of knowledge between specific languages and tasks, leading to more frequent catastrophic forgetting and performance drops. To address these issues, MAD-X (Pfeiffer et al., 2020) proposes task-specific and language-specific layers to facilitate model inference. Hyper-X (Üstün et al., 2022) employs a hypernetwork block to further disentangle the hidden representations of languages and tasks. CPT (Ke et al., 2022) allows the model to continually learn OOD data in a continual learning paradigm. Recently, the concept of Mixture-of-Experts (MoE) (Riquelme et al., 2021; Shazeer et al., 2017; Li et al., 2022; Hu et al., 2023) has revitalized research into cross-lingual adapters. Adamix (Wang et al., 2022) deploys multiple experts in a bottleneck layer with a random routing strategy to achieve language/task-agnostic fine-tuning. MoEBERT (Zuo et al., 2022) departs from traditional PEFTs, proposing importance-guided masking and fine-tuning of neurons within the FFNs. By capturing transferrable knowledge between tasks, HyperMoE and PEMT (Zhao et al., 2024; Lin et al., 2024) show excellent performance in multi-task fine-tuning and is poised for application in multi-task cross-lingual transfer. Furthermore, our method, as a hierarchical MoE adapter, is distinctive in its ability to convert hard approximate zero-shot cross-lingual inference into multi-task cross-lingual transfer inference under a highly challenging language set (only two distantly related languages).

3 Methodology

Assume that the dataset for HierMoE-Adpt training is a combination of source language misinformation detection and bilingual sentiment classification. x is a arbitrary instance from the dataset(the framework is shown in Figure 2).

3.1 HierMoE-Adpt

As shown in Figure 2, we propose HierMoE-Adpt. In HierMoE-Adapt, we strategically and respectively set a group of parallel Mixture of Expert-Adapter (MoE-Adpt) for the multi-head attention layer and the feed-forward Neural Network (FFN) layer. Inspired by existing studies (Shazeer et al., 2017), we assign the MoE group closest to the input layer (MHA-parallel) to undertake the task of transcending linguistic barriers and comprehending the content of news articles and social media posts. Conversely, the other MoE group situated closer to the FFN is designed to specialize in adapting to tasks such as misinformation detection and sentiment analysis (using hidden representations after cross-lingual understanding as inputs). These two sets of MoE-adapters are respectively termed HierMoE-lang for cross-language understanding and HierMoE-task for cross-task learning.

In the context of cross-lingual understanding, HierMoE-lang incorporates two cosine routers designed to guide instances from the source and target language domains, irrespective of the task at hand, towards the most suitable K experts. The routers are denoted as g^{tgt} for the target language and g^{src} for the source language. Specifically, when an instance is fed into its respective router, the router generates an affinity vector, where each element signifies the score associated with a corresponding expert. The $top-K$ experts, determined by their scores, are then selected to process the instance, culminating in an aggregation phase.

Given a suspicious news or post instance x, the process can be written as:

$$h(x) = \sum_{i=1}^{N_e} g^s(x)_i e_i(x),$$
$$\text{where } g^s(x) = \text{top}_K\left(\text{softmax}\left(\frac{E^{lang}W_r^s x}{\tau\|W_r^s\|\|E^{lang}\|}\right)\right), \tag{1}$$

where $s \in \{src, tgt\}$, it depends on the language (source or target) the content of x belonging to. $E^{lang} \in \mathbb{R}^{N_e \times d_e}$ is the learnable identity embeddings of experts, W_r^s denotes the personalized inner weight matrix of the router. τ is a temperature hyperparameter. $h(\cdot)$ denotes the output of HierMoE-lang, while $e_i(\cdot)$ denotes the output of expert i.

Similarly, within HierMoE-task, we introduce two distinct cosine routers, $g^{misinfo}$ for social misinformation detection and g^{senti} for sentiment classification. Hidden states of instances are routed without consideration of language (denoted as h), following a dispatch method similar to that of the MoE-lang, to the K most apt experts, which is written as:

$$u(h) = \sum_{j=1}^{N_e} g^p(h)_j e_j(h),$$
$$\text{where } g^p(h) = top_K\left(softmax\left(\frac{E^{task}W_r^p h}{\tau\|W_r^p\|\|E^{task}\|}\right)\right), \tag{2}$$

where $p \in \{misinfo, senti\}$, it depends on the task (misinformation detection or sentiment classification) that h is intended for. The meaning of all other symbols can be matched one-to-one with the symbols mentioned above.

Therefore, in HierMoE-lang, experts show varying degrees and orientations of proficiency and preference towards understanding the two languages. Some display a specialty for a singular language (receiving instances predominantly in one language) while others excel in capturing commonalities between the source and target languages (receiving a more balanced mix of instances from both languages). This duality aligns with the cornerstone of cross-lingual Natural Language Understanding (NLU). A similar diversity in knowledge is mirrored in HierMoE-task.

To facilitate task-specific classification, we deploy two softmax prediction heads, complemented by a widely-used expert-balancing loss (proposed by SMoE (Riquelme et al., 2021)) to encourage a more equitable distribution of instances across experts. Additionally, drawing from prior research, we employ a Masked Language Model (MLM) loss as an auxiliary term, enhancing the language adapter's (HierMoE-lang in this paper) ability to grasp the language and uncover hidden patterns in the narratives of fake news or posts. The loss function is written as:

$$loss = l_{main} + \lambda_1 l_{balance} + \lambda_2 l_{mlm}, \tag{3}$$

where $l_{main} = l_{misinfo} + \alpha l_{senti}$ represents a combination of the two text classification tasks.

3.2 Inference Period

Our HierMoE-Adpt architecture, featuring hierarchical routers, not only bifurcates the learning of content understanding and task adaptation but also plays an important role during model inference. For unseen instances (target language + misinformation detection), the selection of g^{tgt} and $g^{misinfo}$ as the components of the inference-oriented hierarchical router suffices for inference and evaluation. Viewing the selected experts as a collective, they are expected to possess vast knowledge pertinent to the target language and misinformation detection, alongside transferable and useful insights related to the source language and sentiment classification task.

3.3 Personalized Expert-Masking Mechanism

To ensure the existence of experts specialized for target languages and tasks during model inference, it is necessary to guarantee each expert's knowledge and function specialization. Although it can be gradually achieved through hierarchical routers, we hope to ensure it more efficiently. To this end, inspired by a continual learning method (Ke et al., 2022), we proposed a personalized expert-mask mechanism (PerEM). In PerEM, although the training overhead for each expert is similar to that of a parallel bottleneck adapter, the architecture is set to a two-layer fully connected network with an expanded hidden dimension. Specifically, we deploy a "warm-up" stage, using 50% of the training set and a non-MoE configuration for just 2 training epochs at the beginning. Then, the "warm-up" parameters are used to initialize all experts. Meanwhile, each expert is assigned a unique learnable expert-mask embedding. We deploy a sigmoid as a pseudo gate to generate soft expert-masks from expert-embeddings and use them between the first and second fully connected layers. Afterwards, for each expert, the masks undergo binarization, truncating forward propagation for those neurons that should be activated by other experts, according to the value distribution of the masks. It can be written as:

$$m_i = \sigma\left(e_i^{mask}/\tau_1\right), \tag{4}$$

where $\sigma(\cdot)$ denotes the sigmoid function, e_i represents the mask embedding of expert i. In formula writing, we omit a hard binarization assignment on mask m_i. Then, the mask is used to personalize the learning direction of experts:

$$o_i = e_i^{in} \otimes m_i, \tag{5}$$

where e_i^{in} denotes the hidden state produced by the first feed-forward-project layer of expert i. $\otimes$ represents an element-wise multiply. o_i is the masked output and would be fed into the next layer. The current masks of all experts are stored in the mask set E^{mask}. Meanwhile, we implement a gradient flow blocking operation during back propagation. Specifically, for each expert, we perform a MaxPool accumulation on the expert-masks of all other experts to adjust the gradients of the relevant layers in the expert:

$$grad_i^{'} = grad_i \otimes \left(1 - MaxPool\left(E^{mask} - m_i\right)\right). \tag{6}$$

As we see, gradient components corresponding to the '1' values in the MaxPool operation are reset to '0', while the remaining elements are left unchanged. Note that in source code implementation, we expand the MaxPooled accumulation vector to align with the dimensions of the gradient matrix (omitted in the formula writing). Although the size of each model increases, the application of masks and blocking make the computational overhead nearly equivalent to that of K standard bottleneck adapters.

To summarize, the non-MoE-based "warm-up" training equips the HierMoE-Adpt with fundamental knowledge of both language and task, also serving as an anchor for subsequent personalized learning, while PerEM restricts each expert to learning directions not already occupied by others, achieving the desired function differentiation among experts. An intuitive additional explanation of our motivation in this section is that experts are expected to possess specialized capabilities for understanding specific languages or adapting to particular tasks. However, considering the commonalities in knowledge required across different languages and tasks (for instance, signs of intense emotional expression are critical for both sentiment classification and authenticity checking), while it is desirable for each expert to acquire distinct knowledge, their individualized knowledge should not be entirely personalized, but should softly allow experts to also learn cross-lingual or cross-task commonality knowledge (as evidenced by some works on topic of zero-shot/meta SMD (Lin et al., 2023; Tian et al., 2021)). Notably, in the work that inspired us, expert-masking is designed to prevent knowledge about different tasks from mixing and causing catastrophic forgetting during continuous learning (markedly different from our study, yet the positive side effects are also inherited in our apporach). Our further innovated version aims to ensure that each expert's knowledge instantaneously and synchronously diverges towards specific languages or tasks. The effectiveness of PerEM is further probed and analyzed in the following experimental sections. Conclusively, our method allows the hierarchical routers to more easily and effectively find the suitable experts and combine them. This approach contrasts with the rigid assembly of task and language adapters in other Modular NNs for inference. Our hierarchical router captures the pattern of layer-collaboration during training, achieving smoother inference. The implementation of masking accelerates differentiation and reduces computational costs. Moreover, although we do not specify the exact number of activated neurons for each expert, the outcomes of the model training indicate an approximately equal distribution of neuron activation rights among the experts, consistent with our expectations.

4 Experiments

To standardize the pipeline, we use XLM-R (Conneau et al., 2020) as the base model. The following 8 methods will be evaluated and compared as baselines: (1) Adapter (Houlsby et al., 2019); (2) Parallel Adpt (He et al., 2021); (3) P-Tuning (Liu et al., 2022). (4) MAD-X (Pfeiffer et al., 2020); (5) CPT (Ke et al., 2022); (6) Hyper-X (Üstün et al., 2022); (7) AdaMix (Wang et al., 2022); (8) Ours. All selected method are rigorously for adaption-tuning and all MHAs and FFNs are unfrozen, they're updated with a extremely small learning rate . We conduct our experiments in two cross-lingual multitask settings: 'from Chinese (ZH) to English' and 'from English to Chinese'. Both social misinformation and auxiliary sentiment datasets are collected from Weibo and the English datasets are from Twitter. Full details of baselines, implementations and datasets are shown in Appendix A.

4.1 Comprehensive Evaluations

4.1.1 Main Evaluation

In this section, we set two main scenarios of adaption for cross-lingual social misinformation detection: Chinese-to-English (trained on Weibo, Weibo-Senti-100k, and SA; tested on Twitter) and English-to-Chinese (trained on Twitter, Weibo-Senti-100k, and SA; tested on Weibo). In the approximate zero-shot setting, we provide only 20 visible posts as a small prompt. For the few-shot setting (marked with $*$ in Table 1), we additionally feed 200 in-distribution instances for training. We report the experimental results (average over 3 runs) for both the zero-shot and the few-shot setting in Table 1. The findings are as follows:

(1). HierMoE-Adpt achieved comprehensive performance leadership across four experimental groups. It holds 1.8 percentage points absolute advantage in approximate zero-shot setting, while the advantage increases to approximately 2.8 percentage points in the few-shot setting. This indicates that our method possesses excellent cross-lingual misinformation detection capabilities. (2):

Table 1: Comprehensive Evaluation on ZH-EN and EN-ZH language crossing, "†" represents few-shot settings. $p \leq 0.05$ (*) and $p \leq 0.005$ (**) indicate our paired t-tests vs the best baseline.

	Twitter-to-Weibo				Weibo-to-Twitter			
#metrics	$MacroF1$	Acc	$MacroF1^{\dagger}$	$Acc^{\dagger}$	$MacroF1$	Acc	$MacroF1^{\dagger}$	$Acc^{\dagger}$
XML-R	0.5217	0.5532	0.5896	0.6205	0.5091	0.5370	0.6009	0.6267
Adapter	0.5349	0.5737	0.6192	0.6381	0.5126	0.5561	0.6175	0.6478
Parallel-Adpt	0.5269	0.5829	0.6328	0.6741	0.5303	0.5674	0.6178	0.6609
P-Tuning	0.5489	0.5728	0.6425	0.6692	0.5418	0.5633	0.6346	0.6593
MAD-X	0.6582	0.7019	0.7041	0.7290	0.6332	0.6811	0.6753	0.7305
CPT	0.6446	0.6922	0.6845	0.7322	0.6041	0.6550	0.6830	0.7243
Hyper-X	0.6051	0.6753	0.6520	0.7128	0.5847	0.6485	0.6462	0.7083
AdaMix	0.6612	0.6981	0.7120	0.7559	0.6422	0.6815	0.6710	0.7243
Ours	**0.6781***	**0.7234****	**0.7461****	**0.7897****	**0.6625***	**0.7070***	**0.6841**	**0.7587****

Table 2: Recovery Test under Source Language on ZH-EN and EN-ZH language crossing (Average of 3 Runs).

	Twitter-to-Weibo		Weibo-to-Twitter	
Method	$MacroF1$	Acc	$MacroF1$	Acc
Adapter	0.8417	0.8649	0.8405	0.8491
Parallel-Adpt	0.8395	0.8662	0.8328	0.8470
MAD-X	0.8743	0.8795	0.8519	0.8603
CPT	0.8615	0.8820	0.8561	0.8570
AdaMix	**0.8846**	0.8881	0.8575	0.8655
Ours	0.8835	**0.8976**	**0.8647**	**0.8731**

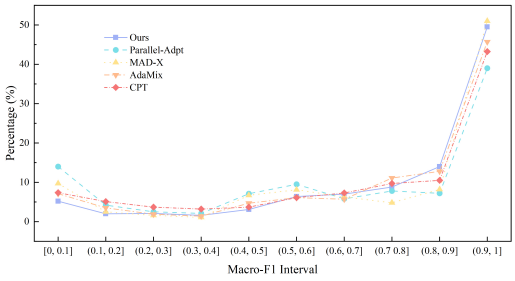

Figure 3: Distribution of Probabilistic Scores (All Groundtruth Labels are "Fake"; Metric: Macro-F1).

Our method demonstrates a more stable advantage under the English-to-Mandarin scenario. In the English-to-Mandarin groups, our method shows an amplified performance advantage of about 30%. When more target samples are provided to the detector, this figure increases further to approximately 45%. A possible reason is that the content and semantics of Twitter instances are richer, and our method is better suited for crossing borders from high-resource to low-resource languages. (3): In the approximate zero-shot settings, our method shows a more significant advantage in terms of the score for positive samples. This experimental phenomenon clearly aligns with our initial intention in designing cross-lingual social misinformation detection. Note that this conclusion is in comparison with other baselines. In nearly all cases of cross-domain or lingual SMD, the macro-f1 score is lower than the Acc.

4.1.2 Recovery Test

After adapting to the target language, the detector is expected to efficiently re-adapt to the source language environment with only a brief period of post-training, continuing to perform the initial SMD (Social Misinformation Detector) tasks. The proposed test represents a significant challenge and a crucial application requirement. It also essentially reflects whether the detector, while crossing linguistic borders, retains its capability to detect misinformation in the source language. In other words, it evaluates whether the cross-lingual adaptation of the social misinformation detector is inherently modular (like a plugin). To test this aspect of performance, in this section, we select the four most competitive baselines (Houlsby et al., 2019; Wang et al., 2022; Ke et al., 2022; Pfeiffer et al., 2020; He et al., 2021) and our method. We randomly sample 500 SMD task samples in the source language and conduct continual training on the models (few-shot groups) post-main evaluation. In terms of experimental details, we reuse all experimental settings and implementation details from the main evaluation. The results are reported in Table 2. Experimental results show that our method has an average performance advantage of 0.31 percentage points over the best-performing competitive baseline. This indicates that our approach can more effectively retain the capability to detect misinformation in social media posts in the source language while crossing linguistic barriers. Furthermore, it demonstrates the ability to rapidly and lightly recover this capability with just a few-shot tuning.

Table 3: Extensive comprehensive evaluation on multi-language settings crossing across four main baselines

	German-Japanese		Japanese-German		Chinese-Japanese		English-German	
#metrics	$MacroF1$	Acc	$MacroF1$	Acc	$MacroF1$	Acc	$MacroF1$	Acc
XML-R	0.5195	0.5492	0.5246	0.5560	0.5780	0.5991	0.5542	0.5893
Adapter	0.5444	0.5781	0.5596	0.5942	0.6245	0.6471	0.6110	0.6297
MAD-X	0.6398	0.6752	0.6613	0.6771	0.6705	0.6874	0.6679	0.6840
AdaMix	0.6503	0.6829	0.6570	0.6814	0.6808	0.7019	0.6690	0.6975
Ours	**0.6645**	**0.7092**	**0.6690**	**0.6883**	**0.7071**	**0.7299**	**0.6994**	**0.7183**

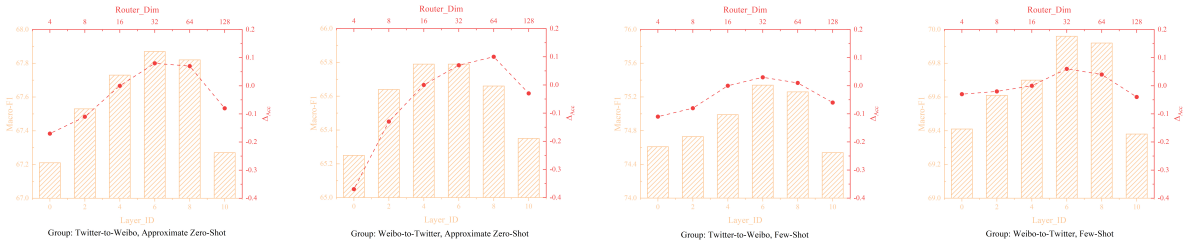

Figure 4: Sensitivity analysis: the Location of Adapter Insertion and Routers' Hidden Dimension.

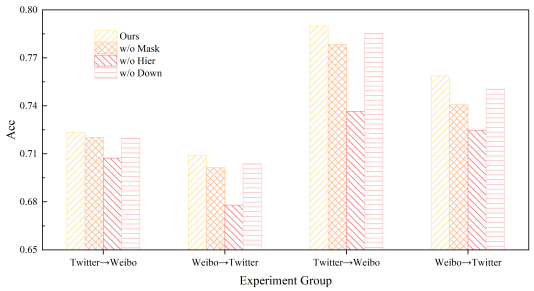

Figure 5: Ablation Study (Task: All Groups, Metric: Acc).

4.1.3 Multilingual Test

To validate the effectiveness of our proposed HierMoE-Adpt across multiple language settings, we respectively collect and pre-process private German and Japanese datasets to conduct a extensive multi-language evaluation. The results are shown in 3.

From the results, we are delighted to find that the advantage of our proposed method is consistent with that in the setting of main evaluations, which furthermore demonstrate that our proposed method can be transferred to multilingual scenarios.

4.2 Extensive Analysis

4.2.1 Segmentation Analysis

To intuitively analyze the performance improvement of HierMOE-Adpt over several competitive baseline methods, we segment the performance for positive instances based on the probabilistic score, using 0.1 as the interval for statistical analysis. The results are demonstrated in Figure 3. Specifically, among the compared methods, we select the three best-performed baselines according to the comprehensive evaluation, i.e., MAD-X, CPT and AdaMix, and add Parallel-Adpt as the most basic method. We use the Twitter-to-Weibo few-shot setting as the example scenario and Macro-F1 as the metrics.

The result shows that our method achieves holistic improvements across nearly all intervals. Specifically, the proportion of items where the model appears "insufficiently confident" remains at a lower level. Although our method performs lower confidence than MAD-X, the overall performance remains the best. Post items in each segmented interval shows a trend of movement towards better interval compared to the results of the Parallel Adapter. These findings suggest that our method contributes to a more optimal and stable probability distribution in the detector's output.

4.2.2 Sensitivity Analysis

In this part, we test the impact of the hidden dimension of routers as well as the location (Layer-ID) where we start to add our HierMoE adapter for each block, continuing to the last block. Specifically, we respectively set the starting Layer-ID for the HierMoE adapter as {0, 2, 4, 6, 8, 10}, and try the hidden dimension of routers across {4, 8, 16, 32, 64, 128}. Experiment results, reported in Figure 4, show that, considering the budget, 16 and 32 are the best hidden dimensions. Additionally, our method works best when inserted into only the last

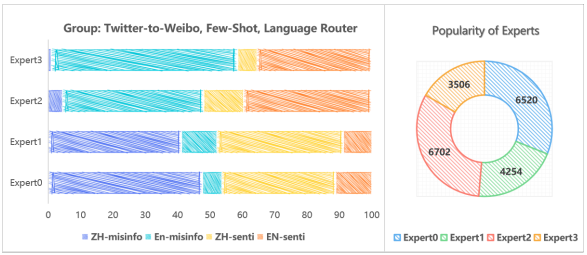

Figure 6: Details of Sample Dispatch on Language Level (Task: Few-shot Setting).

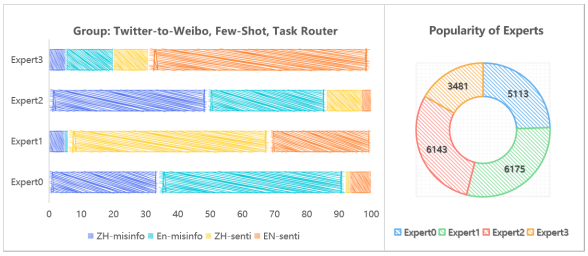

Figure 7: Details of Sample Dispatch on Task Level (Task: Few-shot Setting).

6 blocks (as all-layer insertion has negative effect). The results also indicate that, in our approach, unlike other MoE-pretrain methods and MoE-based frameworks, the value of the router's embedding dimension should be set slightly larger, rather than significantly larger, than the number of experts. (previous works suggest either half the number of experts or significantly higher.) Furthermore, inserting our proposed module only in the relatively later layers proves more efficient, yielding approximately a 0.3–0.6 percentage point performance advantage. This is particularly meaningful in approximate zero-shot settings.

4.2.3 Ablation Study

We conduct simple ablation studies on the four settings. To build the three downgraded versions, we respectively removed the Expert-Mask, hierarchical routers (replacing them with dual-layer sparse MoE adapters), and placed both sets of MoE adapters downstream of the FFN (similar to MAD-X, there is no adapter layers in parallel with MHA layers.). The experimental results shown in Figure 5 demonstrate that hierarchical routers are indispensable for effective inference. When the hierarchical routers are removed, performance drops sharply by an average of 2.48 percentage points. Both the Expert-Mask and the position of the adapters are indispensable instruments, with an average impact on performance of 1.01 and 0.43 percentage points, respectively. In conclusion, the Hierarchical-MoE strategy constitutes the primary factor in improving the performance. Furthermore, rather than stacking all adapters downstream of transformer blocks, our adapter placement strategy allows more direct positive effects.

4.2.4 Inner Experts

We probe the dispatch details of samples across experts (the composition(%) of the sample set dispatched to an arbitrary expert) and count the popularity of experts (the total numbers of dispatched samples). We select the Twitter-to-Weibo + Few-shot as the experiment group.

As shown in Figure 6 and Figure 7, our hierarchical strategy proves efficient. Some experts clearly demonstrate specialization in the target language and the main task (SMD), while the capacity of experts is substantially balanced. Specifically, for language routers, the least and most popular experts are allocated a sample quantity roughly 30% below and 17% above the average, respectively. Such disparity in dispatch is acceptable in MoE deployments. For Task routers, this gap is mitigated to a certain extent. Moreover, experts specializing in social misinformation detection exhibit higher popularity (more post items are input). Conversely, routing results for items from sentiment classification task are more dispersed, with a significant portion dispatched to experts not specialized in the task at hand (more so than the reverse). Nevertheless, experts for understanding Chinese and English items demonstrated better balance (with the difference in their popularity aligning closely with the actual data volume ratio between the two languages, only a 5.74% disparity). Overall, our routers also perform quite well in dispatching.

5 Conclusions

In this work, we propose HierMoE-Adpt, an innovative PEFT adaption method for cross-lingual social misinformation detection, which additionally leverages bilingual sentiment analysis knowledge. It incorporates a hierarchical routing strategy and an expert-masking mechanism. HierMoE-Adpt advances cross-lingual SMD and mitigates the challenges related to protecting non-native users from misinformation on SNS. Experimental results indicate that HierMoE-Adpt outperforms the baselines and exhibits flexibility in crossing language borders back and forth through post-training.

Limitations

Considering the preceding sections have thoroughly explained the motivation, innovations, specific algorithm, and experimental design of HierMoE-Adpt, this section primarily discusses the limitations identified in our study, as well as challenges that await mitigation in the future.

Inference Period

We are in search of a better inference method that can assemble and integrate the knowledge of various experts based on masks, thereby achieving non-MoE inference. Although intuitively, utilizing the two hierarchical routers to perform inference seems an elegant and logical approach, previous works confirm that, after extensive training, merge-based inference methods yield the best end-task performance for MoEs. Such a proposal looks more promising considering that our masking mechanism can ensure low negative interference among experts.

Transfer Application

Since HierMoE–Adpt is not a specific architecture but a broadly applicable adaption method, it can be directly applied to tasks like improving well-known models that do not rely on PLM backbones, such as MDFEND (Nan et al., 2021). Thus, we hope to further investigate whether HierMoE–Adpt can serve as a plugin to enhance existing SOTA SMD frameworks (cross-lingual or cross-domain).

References

Bhavika Bhutani, Neha Rastogi, Priyanshu Sehgal, and Archana Purwar. 2019. Fake news detection using sentiment analysis. In *2019 twelfth international conference on contemporary computing (IC3)*, pages 1–5. IEEE.

Samuel Kai Wah Chu, Runbin Xie, and Yanshu Wang. 2021. Cross-language fake news detection. *Data and Information Management*, 5(1):100–109.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jan Christian Blaise Cruz, Julianne Agatha Tan, and Charibeth Cheng. 2020. Localization of fake news detection via multitask transfer learning. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2596–2604.

Daryna Dementieva and Alexander Panchenko. 2021. Cross-lingual evidence improves monolingual fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 310–320.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jiangshu Du, Yingtong Dou, Congying Xia, Limeng Cui, Jing Ma, and S Yu Philip. 2021. Cross-lingual covid-19 fake news detection. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 859–862. IEEE.

Anastasia Giachanou, Paolo Rosso, and Fabio Crestani. 2019. Leveraging emotional signals for credibility detection. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 877–880.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682.

Yasser H. 2024. Twitter tweets sentiment dataset. https://www.kaggle.com/datasets/yasserh/twitter-tweets-sentiment-dataset.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ke Hu, Bo Li, Tara N Sainath, Yu Zhang, and Francoise Beaufays. 2023. Mixture-of-expert conformer for streaming multilingual asr. *arXiv preprint arXiv:2305.15663*.

Zixuan Ke, Haowei Lin, Yijia Shao, Hu Xu, Lei Shu, and Bing Liu. 2022. Continual training of language models for few-shot learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10205–10216.

K Hazel Kwon, C Chris Bang, Michael Egnoto, and H Raghav Rao. 2016. Social media rumors as improvised public opinion: semantic network analyses of twitter discourses during korean saber rattling 2013. *Asian Journal of Communication*, 26(3):201–222.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Sarah I Levitan, Guzhen An, Mandi Wang, Gideon Mendels, Julia Hirschberg, Michelle Levine, and Andrew Rosenberg. 2015. Cross-cultural production and detection of deception from speech. In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, pages 1–8.

Bo Li, Yifei Shen, Jingkang Yang, Yezhen Wang, Jiawei Ren, Tong Che, Jun Zhang, and Ziwei Liu. 2022. Sparse mixture-of-experts are domain generalizable learners. *arXiv preprint arXiv:2206.04046*.

Hongzhan Lin, Pengyao Yi, Jing Ma, Haiyun Jiang, Ziyang Luo, Shuming Shi, and Ruifang Liu. 2023. Zero-shot rumor detection with propagation structure via prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5213–5221.

Zhisheng Lin, Han Fu, Chenghao Liu, Zhuo Li, and Jianling Sun. 2024. Pemt: Multi-task correlation guided mixture-of-experts enables parameter-efficient transfer learning. *arXiv preprint arXiv:2402.15082*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347.

Dan S Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3141–3153.

Oguzhan Ozcelik, Arda Sarp Yenicesu, Onur Yildirim, Dilruba Sultan Haliloglu, Erdem Ege Eroglu, and Fazli Can. 2023. Cross-lingual transfer learning for misinformation detection: Investigating performance across multiple languages. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 549–558.

Subhadarshi Panda and Sarah Ita Levitan. 2022. Improving cross-domain, cross-lingual and multi-modal deception detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 383–390.

Verónica Pérez-Rosas and Rada Mihalcea. 2014. Cross-cultural deception detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 440–445.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1212–1220.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2018. Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Qiang Sheng, Xueyao Zhang, Juan Cao, and Lei Zhong. 2021. Integrating pattern-and fact-based fake news

detection via model preference learning. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 1640–1650.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

SophonPlus. 2024. Weibo sentiment analysis 100k dataset. https://github.com/SophonPlus/ChineseNlpCorpus/tree/master/datasets/weibo_senti_100k.

Lin Tian, Xiuzhen Zhang, and Jey Han Lau. 2021. Rumour detection via zero-shot cross-lingual transfer learning. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part I 21*, pages 603–618. Springer.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, Gertjan van Noord, and Sebastian Ruder. 2022. Hyper-x: A unified hypernetwork for multi-task multilingual transfer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7934–7949.

Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan, and Jianfeng Gao. 2022. Adamix: Mixture-of-adaptations for parameter-efficient model tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5744–5760.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Weiming Wen, Songwen Su, and Zhou Yu. 2018. Cross-lingual cross-platform rumor verification pivoting on multimedia content. *arXiv preprint arXiv:1808.04911*.

Jiaying Wu, Shen Li, Ailin Deng, Miao Xiong, and Bryan Hooi. 2023. Prompt-and-align: prompt-based social alignment for few-shot fake news detection. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2726–2736.

Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter*, 21(2):80–90.

Zedian Xiao, William Held, Yanchen Liu, and Diyi Yang. 2023. Task-agnostic low-rank adapters for unseen english dialects. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7857–7870.

Chen Yang, Xinyi Zhou, and Reza Zafarani. 2021. Checked: Chinese covid-19 fake news dataset. *Social Network Analysis and Mining*, 11(1):58.

Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2019. Jointly embedding the local and global relations of heterogeneous graph for rumor detection. In *2019 IEEE international conference on data mining (ICDM)*, pages 796–805. IEEE.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*, pages 3465–3476.

Hao Zhao, Zihan Qiu, Huijia Wu, Zili Wang, Zhaofeng He, and Jie Fu. 2024. Hypermoe: Towards better mixture of experts via transferring among experts. *arXiv preprint arXiv:2402.12656*.

Xianbing Zhou, Yang Yong, Xiaochao Fan, Ge Ren, Yunfeng Song, Yufeng Diao, Liang Yang, and Hongfei Lin. 2021. Hate speech detection based on sentiment knowledge sharing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7158–7166.

Simiao Zuo, Qingru Zhang, Chen Liang, Pengcheng He, Tuo Zhao, and Weizhu Chen. 2022. Moebert: from bert to mixture-of-experts via importance-guided adaptation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1610–1623.

A Experiment Setting Details

A.1 Implementation Details

In our experiments, we set the adapter's hidden dimension to 64 (Only CPT and Ours have 256 bdue to the hard mask mechanism). For our method, the number of experts per group is set to 4, with each router selecting $K = 2$ experts, and $\{\lambda_1, \lambda_2, \alpha\} = \{0.01, 0.08, 0.5\}$. We train all adapters (CPT and Ours) for 100 epochs, with a batch size of 64 per device, a learning rate of 5e-5 (all backbone components of XLM-R remain frozen except LayerNorm and Classification Heads), and a weight decay of 0.01. Other hyperparameters include $\beta_1 = 0.9$, $\beta_2 = 0.98$, a router hidden dimension of 16, and a maximum sequence length of 200. The dimension of mask-embedding is set to 256. For training datasets, we split the training and validation sets in a 90%:10% ratio and use the remaining posts in the target language social misinformation detection dataset as the test set. We report accuracy and macro-averaged F1. Experiments are conducted using 1 NVDIA A100-40G GPUs (2 $*$ NVDIA 4090 GPUs with batch-size-per-device as 32 can be a lower substitute plan).

A.2 Datasets

In this paper, we conduct our experiments in two cross-lingual, multitask settings oriented towards social misinformation detection: 'from Chinese to English' and 'from English to Chinese'. We selected Twitter as the social media platform for the English-language environment and Weibo, the Chinese equivalent of Twitter, for the Mandarin-language environment. Specifically, For ZH-Misinformation, we adopt the most widely-used Weibo dataset (Ma et al., 2016) (consisting of 5,189 post items). For EN-Misinformation, we combine the Twitter15 and Twitter16 datasets (Yuan et al., 2019) (totaling 5,803 post items). For ZH-Sentiment, we randomly sample 5,000 posts from Weibo-Senti-100K (SophonPlus, 2024), a large-scale sentiment polarity dataset containing over 100K labelled user posts. For EN-Sentiment, we randomly sample 5,000 posts from Sentiment Analysis (SA) – Kaggle 2021 dataset (H, 2024), which contains over 30k labeled user posts on a wide range of topics.

A.3 Baseline Introductions

To standardize the pipeline, we use XLM-R (Conneau et al., 2020) as the base model. The following 8 methods are evaluated and compared as baselines: (1) Adapter (Houlsby et al., 2019); a bottleneck layer on top of FFNs. (2) Parallel Adpt (He et al., 2021); a parallel version, both for MHA and FFN. (3) P-Tuning (Liu et al., 2022). (4) MAD-X (Pfeiffer et al., 2020); with language-specific and task-specific layers, the most commonly used cross-lingual multi-task transfer adapter. (5) CPT (Ke et al., 2022); a cross-domain adapter based on continual learning, we set the training sequence as "source-language-senti, source-language-misinfo, target-language-senti". (6) Hyper-X (Üstün et al., 2022); a competitor based on hypernetworks; (7) AdaMix (Wang et al., 2022); the only MoE-based adapter competitor, without hierarchical mechanism. (8) Ours. All selected method are rigorously used for adaption-tuning, with all MHAs and FFNs unfrozen during training.