MetaAdapt: Domain Adaptive Few-Shot Misinformation Detection via Meta Learning

Zhenrui Yue **Huimin Zeng** **Yang Zhang** **Lanyu Shang** **Dong Wang**
University of Illinois Urbana-Champaign
{zhenrui3, huiminz3, yzhangnd, lshang3, dwang24}@illinois.edu

Abstract

With emerging topics (e.g., COVID-19) on social media as a source for the spreading misinformation, overcoming the distributional shifts between the original training domain (i.e., source domain) and such target domains remains a non-trivial task for misinformation detection. This presents an elusive challenge for early-stage misinformation detection, where a good amount of data and annotations from the target domain is not available for training. To address the data scarcity issue, we propose MetaAdapt, a meta learning based approach for domain adaptive few-shot misinformation detection. MetaAdapt leverages limited target examples to provide feedback and guide the knowledge transfer from the source to the target domain (i.e., learn to adapt). In particular, we train the initial model with multiple source tasks and compute their similarity scores to the meta task. Based on the similarity scores, we rescale the meta gradients to adaptively learn from the source tasks. As such, MetaAdapt can learn how to adapt the misinformation detection model and exploit the source data for improved performance in the target domain. To demonstrate the efficiency and effectiveness of our method, we perform extensive experiments to compare MetaAdapt with state-of-the-art baselines and large language models (LLMs) such as LLaMA, where MetaAdapt achieves better performance in domain adaptive few-shot misinformation detection with substantially reduced parameters on real-world datasets.

1 Introduction

Recently, significant progress has been made in misinformation detection due to the improvements in developing machine learning-based methods (Wu et al., 2019; Shu et al., 2020c; Wu et al., 2022b). Such methods include large language models (LLMs), which can be fine-tuned for detecting and responding to rumors on social media platforms (Jiang et al., 2022; He et al., 2023; Touvron et al., 2023). However, misinformation on emerging topics remains an elusive challenge for existing approaches, as there exists a large domain gap between the training (i.e., source domain) and the target distribution (i.e., target domain) (Yue et al., 2022). For instance, existing models often fail to detect early-stage misinformation due to the lack of domain knowledge (see Figure 1).

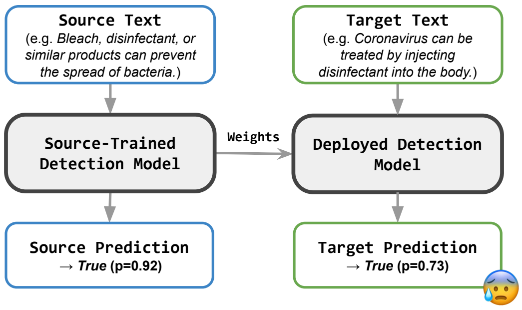

Figure 1: Existing models (from source domain) fail to detect rumors on emerging topics (target domain).

With the increase of emerging topics (e.g., COVID-19) on social media as a source of misinformation, the failure to distinguish such early-stage misinformation can result in potential threats to public interest (Roozenbeek et al., 2020; Chen et al., 2022). To tackle the problem of cross-domain early misinformation detection, one possible solution is crowdsourcing, which collects domain knowledge from online resources (Medina Serrano et al., 2020; Hu et al., 2021b; Shang et al., 2022a; Kou et al., 2022b). Another alternative approach is to transfer knowledge from labeled source data to unlabeled target data with domain adaptive methods (Zhang et al., 2020; Li et al., 2021; Suprem and Pu, 2022; Shu et al., 2022; Yue et al., 2022; Zeng et al., 2022). However, the former methods use large amounts of human annotations while the latter approaches require extensive unlabeled examples. As such, existing methods are less effective for detecting cross-domain early misinformation, where neither large amounts of annotations nor tar-

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5223–5239
July 9-14, 2023 ©2023 Association for Computational Linguistics

get domain examples can be provided for training.

Despite the insufficiency of early misinformation data, limited target examples and annotations can often be achieved at minimal costs (Kou et al., 2021, 2022a; Shang et al., 2022c). Nevertheless, previous approaches are not optimized to learn from the source data under the guidance of limited target examples (Zhang et al., 2020; Lee et al., 2021; Mosallanezhad et al., 2022; Yue et al., 2022). Such methods are often unaware of the adaptation objective and thus fail to maximize the transfer of source domain knowledge. To fully exploit existing data from different domains, we consider a cross-domain few-shot setting to adapt misinformation detection models to an unseen target domain (Motiian et al., 2017; Zhao et al., 2021; Lin et al., 2022). That is, given a source data distribution and access to limited target examples, our objective is to maximize the model performance in the target domain. An example of such application can be adapting a model from fake news detection to COVID early misinformation detection, where abundant fake news from existing datasets can be used for training the model under the guidance of limited COVID misinformation examples.

In this paper, we design MetaAdapt, a few-shot domain adaptation approach based on meta learning for early misinformation detection. Specifically, we leverage the source domain examples (i.e., source task) and train a model to obtain the task gradients. Then, we evaluate the updated model on the few-shot target examples (i.e., meta task) to derive second-order meta gradients w.r.t. the original parameters. We additionally compute the similarity between the task gradients and meta gradients to select more 'informative' source tasks, such that the updated model adaptively learns (from the source data) to generalize even with a small number of labeled examples. In other words, the meta model learns to reweight the source tasks with the objective of optimizing the model performance in the target domain. Therefore, the resulting model can optimally adapt to the target distribution with the provided source domain knowledge. To show the efficacy of our meta learning-based adaptation method, we focus on the early-stage misinformation of COVID-19 and demonstrate the performance of MetaAdapt on real-world datasets, where MetaAdapt can consistently outperform state-of-the-art methods and large language models by demonstrating significant improvements.

We summarize our contributions as follows[1]:

1. We propose a few-shot setting for domain adaptive misinformation detection. Here, the labeled source data and limited target examples are provided for the adaptation process.
2. We propose MetaAdapt, a meta learning-based method for few-shot domain adaptive misinformation detection. Our MetaAdapt 'learns to adapt' to the target data distribution with limited labeled examples.
3. MetaAdapt can adaptively learn from the source tasks by rescaling the meta gradients. Specifically, we compute similarity scores between the source and meta tasks to optimize the learning from the source distribution.
4. We show the effectiveness of MetaAdapt in domain adaptive misinformation detection on multiple real-world datasets. In our experiments, MetaAdapt consistently outperforms state-of-the-art baselines and LLMs.

2 Related Work

2.1 Misinformation Detection

Existing misinformation detection methods can be categorized into the following: (1) content-based misinformation detection: such models are trained to perform misinformation classification upon input claims. For example, pretrained transformer models are used to extract semantic or syntactic properties to detect misinformation (Karimi and Tang, 2019; Das et al., 2021; Yue et al., 2022; Jiang et al., 2022). Moreover, multimodal input is used to learn text and image features that improve detection performance (Khattar et al., 2019; Shang et al., 2021; Santhosh et al., 2022; Shang et al., 2022b); (2) social-aware misinformation detection: user interactions can be used to evaluate online post credibility (Jin et al., 2016). Similarly, patterns on propagation paths help detect misinformation on social media platforms (Monti et al., 2019; Shu et al., 2020b). Social attributes like user dynamics enhance misinformation detection by introducing context (Shu et al., 2019). Combined with content-based module, misinformation detection systems demonstrate improved accuracy (Mosallanezhad et al., 2022; Lin et al., 2022); (3) knowledge-based

[1]We adopt publicly available datasets in the experiments and release our implementation at https://github.com/Yueeeeeeee/MetaAdapt.

misinformation detection: external knowledge can be leveraged as supporting features and evidence in fact verification and misinformation detection (Vo and Lee, 2020; Liu et al., 2020; Brand et al., 2021). Knowledge graphs or crowdsourcing approaches can derive additional information for explainability in misinformation detection (Cui et al., 2020; Hu et al., 2021b; Koloski et al., 2022; Kou et al., 2022a; Shang et al., 2022a; Wu et al., 2022a). Yet existing methods focus on improving in-domain performance or explainability, few-shot misinformation detection in a cross-domain setting is not well researched. Hence, we study domain adaptive few-shot misinformation detection using content-based language models in our work.

2.2 Domain Adaptive Learning

Domain adaptive learning aims to improve model generalization on an unseen domain given a labeled source domain. Such methods are primarily studied in image and text classification problems (Li et al., 2018; Kang et al., 2019; Sicilia et al., 2021). In image classification, existing methods minimize the representation discrepancy between source and target domains to learn domain-invariant features and transfer source knowledge to the target domain (Kang et al., 2019; Na et al., 2021). Similarly, domain-adversarial and energy-based methods adopt additional critique, with which domain-specific features are regularized (Sicilia et al., 2021; Xie et al., 2022). Class-aware contrastive learning is proposed for fine-grained alignment, which regularizes the inter-class and intra-class distances to achieve domain-invariant yet class-separating features (Li et al., 2018; Shen et al., 2022).

In text classification, various approaches are proposed to improve the target domain performance in cross-domain settings (Silva et al., 2021; Li et al., 2021; Ryu et al., 2022; Nan et al., 2022). For instance, domain-adversarial training is used to learn generalizable features to detect cross-domain multimodal misinformation (Wang et al., 2018; Lin et al., 2022; Shu et al., 2022). Reinforcement learning and contrastive adaptation are also adopted for fine-grained domain adaptation in misinformation detection (Mosallanezhad et al., 2022; Yue et al., 2022). Nevertheless, domain-adaptive misinformation detection is not well studied in the few-shot learning setting. Therefore, we combine both settings and develop a method tailored for few-shot domain adaptation in misinformation detection: MetaAdapt. By leveraging knowledge transfer via the proposed meta objective, our approach shows significant improvements on out-of-domain misinformation using only a few labeled examples.

2.3 Few-Shot Learning

Few-shot learning aims to learn a new task with a few labeled examples (Wang et al., 2020). Existing few-shot learning approaches (e.g., prototypical networks) learn class-wise features in the metric space to rapidly adapt to new tasks (Vinyals et al., 2016; Snell et al., 2017). Meta learning methods search for the optimal initial parameters for unseen few-shot tasks via second-order optimization (Finn et al., 2017; Rajeswaran et al., 2019; Zhou et al., 2021). In computer vision, few-shot domain adaptation is studied in image classification to transfer knowledge to an unseen target domain (Motiian et al., 2017; Tseng et al., 2019; Zhao et al., 2021). For language problems, meta learning is proposed to improve the few-shot performance in language modeling and misinformation detection (Sharaf et al., 2020; Han et al., 2021; Salem et al., 2021; Zhang et al., 2021; Lei et al., 2022).

To the best of our knowledge, few-shot domain adaptive misinformation detection via meta learning is not studied in current literature. Moreover, the mentioned few-shot setting can be helpful in real-world scenarios (e.g., detecting rumors on emerging topics). As such, we propose meta learning-based MetaAdapt for misinformation detection. MetaAdapt leverages limited target examples and adaptively exploits the source domain knowledge via task similarity, and thus improves the few-shot misinformation detection performance in the unseen target domain.

3 Preliminary

We consider the following problem setup for domain adaptive few-shot misinformation detection: labeled source data and k-shot target examples (i.e., k examples per class) are available for training. The objective of our framework is to train a misinformation detection model $\boldsymbol{f}$ that is optimized for the target domain performance using both source and few-shot target examples.

Data: Our research is defined within the scope of *single-source* adaptive misinformation detection (i.e., we study the adaptation problem from a single source domain to the target domain). We denote $\mathcal{D}_s$ as the source domain and $\mathcal{D}_t$ as the (differ-

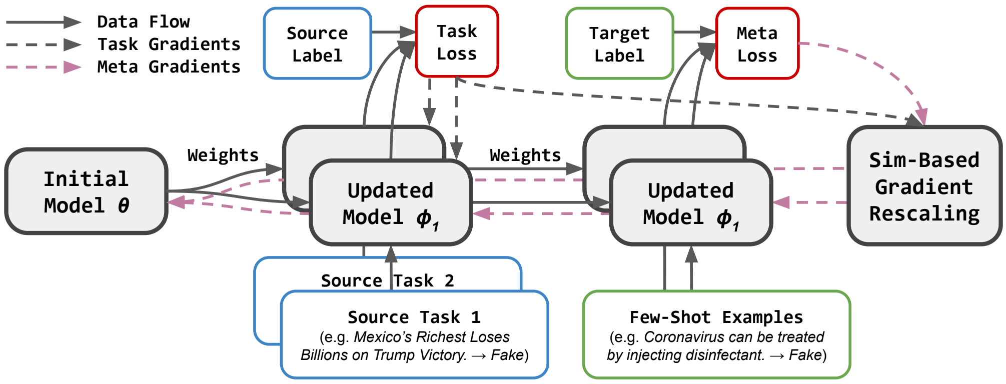

Figure 2: The proposed MetaAdapt, we illustrate the computation of task and meta gradients via task similarity.

ent) target domain. In our setting, labeled source data and limited target examples can be used for training. The few-shot adaptation is performed in two-fold: (1) an initial model is updated upon multiple batches of sampled source data examples (i.e., source tasks) respectively; (2) the updated models are evaluated on the few-shot target examples respectively (i.e., meta task) to compute the meta loss, followed by updating the initial parameters using the derived second-order derivatives. The input data is defined as follows:

- *Labeled source data*: source training data $\boldsymbol{X}_s$ is provided by source domain $\boldsymbol{\mathcal{D}}_s$. Here, each sample $(\boldsymbol{x}_s^{(i)}, y_s^{(i)}) \in \boldsymbol{X}_s$ is a tuple comprising of input text $\boldsymbol{x}_s^{(i)}$ and label $y_s^{(i)} \in \{0, 1\}$ (i.e., false or true). During training, source data batches are sampled as different 'source tasks' and used to optimize the initial model.

- *Few-shot target data*: we assume limited access to the target domain $\boldsymbol{\mathcal{D}}_t$. In other words, only k-shot subset $\boldsymbol{X}_t'$ from $\boldsymbol{X}_t$ is provided for training. Target samples are provided in the same label space with $(\boldsymbol{x}_t^{(i)}, y_t^{(i)}) \in \boldsymbol{X}_t'$, while the size of $\boldsymbol{X}_t'$ is constrained with k examples in each label class (i.e., 10 in our experiments). $\boldsymbol{X}_t'$, or 'meta task' is used to compute the meta loss and meta gradients w.r.t. the initial parameters.

Model & Objective: The misinformation detection model is represented by a function $\boldsymbol{f}$ parameterized by $\boldsymbol{\theta}$. $\boldsymbol{f}$ takes textual statements as input and predicts the probability of input as true information, i.e., $y^{(i)} = \arg\max(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{x}^{(i)}))$. For optimization, our objective is to maximize the model performance on target data $\boldsymbol{X}_t$ (Note $\boldsymbol{X}_t \neq \boldsymbol{X}_t'$) from the target domain $\boldsymbol{\mathcal{D}}_t$. Mathematically, this can be formulated as the optimization problem of minimizing the loss $\mathcal{L}$ of $\boldsymbol{\theta}$ over target data $\boldsymbol{X}_t$:

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{X}_t), \tag{1}$$

where $\mathcal{L}$ is the loss function (i.e., cross-entropy).

4 Methodology

Provided with labeled source data and k-shot target examples, we first present our meta adaptation framework for domain adaptive few-shot misinformation detection. To improve the adaptation performance, we introduce a second-order meta learning algorithm MetaAdapt based on learnable learning rate and task similarity. An illustration of MetaAdapt is provided in Figure 2. Upon deployment, the adapted models achieve considerable improvements thanks to the adaptive optimization and similarity-guided meta adaptation.

4.1 Few-Shot Meta Adaptation

Given model $\boldsymbol{f}$ with initial parameters $\boldsymbol{\theta}$, source dataset $\boldsymbol{X}_s$, few-shot target data $\boldsymbol{X}_t'$ and the number of tasks n in each iteration, we formulate the meta adaptation framework as a bi-level optimization problem and provide a mathematical formulation:

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum^{n} \mathcal{L}(\mathcal{A}lg(\boldsymbol{\theta}, \texttt{Sampler}(\boldsymbol{X}_s)), \boldsymbol{X}_t'), \tag{2}$$

in which Sampler stands for the source task sampler that draws source tasks of a fixed size from $\boldsymbol{X}_s$, $\mathcal{A}lg$ represents the optimization algorithm using first-order gradient descent, i.e.:

$$\mathcal{A}lg(\boldsymbol{\theta}, \boldsymbol{X}) = \boldsymbol{\phi} = \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{X}), \tag{3}$$

with α representing the task learning rate and $\boldsymbol{\phi}$ representing the learnt parameter set over $\boldsymbol{X}$.

In Equation (2), we are interested in learning an optimal parameter set $\boldsymbol{\theta}$ that minimizes the meta loss on the few-shot target set $\boldsymbol{X}_t'$, which can be denoted as the outer-level optimization of meta adaptation. The outer-level learning is achieved by deriving gradients w.r.t. $\boldsymbol{\theta}$ based on the meta loss using task-specific parameters (i.e., $\mathcal{A}lg(\boldsymbol{\theta}, \texttt{Sampler}(\boldsymbol{X}_s))$ or $\boldsymbol{\phi}$) and few-shot examples $\boldsymbol{X}_t'$. To obtain task-specific parameters, we sample a batch of source examples using `Sampler` and perform gradient descent steps on the original $\boldsymbol{\theta}$ (i.e, Equation (3)), which is known as the inner-level optimization. The inner-level optimization only requires first-order derivatives, however, to optimize the outer-level problem, it is necessary to differentiate through $\mathcal{A}lg$ (i.e., $\boldsymbol{\phi}$), which requires using second-order gradients (Finn et al., 2017).

We now take a closer look at how to compute the derivatives with chain rule in meta adaptation:

$$\frac{d\mathcal{L}(\mathcal{A}lg(\boldsymbol{\theta}, \boldsymbol{X}), \boldsymbol{X}_t')}{d\boldsymbol{\theta}} = \frac{d\mathcal{A}lg(\boldsymbol{\theta}, \boldsymbol{X})}{d\boldsymbol{\theta}} \nabla_{\boldsymbol{\phi}} \mathcal{L}(\mathcal{A}lg(\boldsymbol{\theta}, \boldsymbol{X}), \boldsymbol{X}_t'), \tag{4}$$

Note that $\mathcal{A}lg(\boldsymbol{\theta}, \boldsymbol{X})$ is equivalent to $\boldsymbol{\phi}$. The right-side component $\nabla_{\boldsymbol{\phi}} \mathcal{L}(\mathcal{A}lg(\boldsymbol{\theta}, \boldsymbol{X}), \boldsymbol{X}_t')$ refers to first-order derivatives by computing the meta loss using the few-shot examples $\boldsymbol{X}_t'$ and task-specific parameter set $\boldsymbol{\phi}$ ($\mathcal{L} \rightarrow \boldsymbol{\phi}$). This step can be computed with conventional gradient descent algorithms. The left-side component $\frac{d\mathcal{A}lg(\boldsymbol{\theta}, \boldsymbol{X})}{d\boldsymbol{\theta}}$ is a non-trivial step as it requires second-order derivatives (i.e., Hessian matrix) to track parameter-to-parameter changes through $\mathcal{A}lg(\boldsymbol{\theta}, \boldsymbol{X})$ to $\boldsymbol{\theta}$. In our implementation, we compute the meta gradients w.r.t. $\boldsymbol{\theta}$ using the meta evaluation loss similar to model agnostic meta learning (Finn et al., 2017). We also adopt adaptive learning rate α and β and cosine annealing to improve the convergence of source and meta tasks (Antoniou et al., 2018).

4.2 The Proposed MetaAdapt

While meta adaptation leverage source tasks to improve the target domain performance, it learns homogeneously from all source tasks without considering the informativeness of each individual task. To further exploit the source domain knowledge, we propose a task similarity-based MetaAdapt for domain adaptive few-shot misinformation detection. We first estimate the task-specific parameters with adaptive learning rates, followed by rescaling the meta loss using the task similarity scores. The proposed method selectively learns from source tasks, and thus, further improves the adaptation performance. We present the training details of MetaAdapt in Algorithm 1.

Algorithm 1: MetaAdapt Algorithm

```
Input Parameter set θ, source data X_s, k-shot data
    X'_t, number of iterations N, number of tasks n;
for iter ∈ {1, 2, ..., N} do
    for i ∈ {1, ..., n} do
        Sample source task from X_s;
        Update task parameter set φ_i with
          Equation (3);
        Compute meta loss and meta gradients using
          φ_i, X'_t as in Equation (4);
        Calculate similarity score s_i with
          Equation (5);
    end
    Normalize similarity scores s with Equation (6);
    Update original parameter θ with Equation (7);
end
```

We initialize the model and denote the parameters with $\boldsymbol{\theta}$. For source task i, the original parameters are updated with first-order derivatives as we sample source tasks from $\boldsymbol{X}_s$. Specifically in each step, the task gradients can be computed with $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \texttt{Sampler}(\boldsymbol{X}_s))$, as in Equation (3). After a few gradient descent steps, the parameters converge locally and we denote the updated parameters with $\boldsymbol{\phi}_i$. As we update multiple times in each source task, we denote the task gradients with $\boldsymbol{\phi}_i - \boldsymbol{\theta}$ for simplicity. Subsequently, the meta loss $\mathcal{L}(\boldsymbol{\phi}_i, \boldsymbol{X}_t')$ is computed using $\boldsymbol{\phi}_i$ and the few-shot target examples $\boldsymbol{X}_t'$. To compute the meta gradients with backpropagation, we follow the chain rule and compute the derivatives w.r.t. the original parameter set $\boldsymbol{\theta}$. Similar to Equation (4), we use $\frac{d\boldsymbol{\phi}_i}{d\boldsymbol{\theta}} \nabla_{\boldsymbol{\phi}_i} \mathcal{L}(\boldsymbol{\phi}_i, \boldsymbol{X}_t')$ to denote the meta gradients.

To compute task similarity scores, we leverage task and meta gradients. The objective of computing gradient similarity is to selectively learn from the source tasks. If the task and meta gradients yield a high similarity score, the parameters are converging to the same direction in both inner- and outer-loop optimization. Thus, the source task optimization path is more 'helpful' to improve the meta task performance (i.e., target domain performance). Or if, on the contrary, then the source task may be less effective for improving the meta task performance. Based on this principle, we compute task similarity score s_i with cosine similarity:

$$s_i = \texttt{CosSim}(\boldsymbol{\phi}_i - \boldsymbol{\theta}, \frac{d\boldsymbol{\phi}_i}{d\boldsymbol{\theta}} \nabla_{\boldsymbol{\phi}_i} \mathcal{L}(\boldsymbol{\phi}_i, \boldsymbol{X}_t')). \tag{5}$$

In each iteration, we sample n source tasks and compute the similarity scores for each pair of task and meta gradients. Then, the similarity scores $[s_1, s_2, ..., s_n]$ are transformed to a probability distribution using tempered softmax:

$$\boldsymbol{s} = \texttt{softmax}([\frac{s_1}{\tau}, \frac{s_2}{\tau}, ..., \frac{s_n}{\tau}]), \quad (6)$$

where τ is the temperature hyperparameter to be selected empirically. Finally, we update the original parameters with rescaled meta gradients:

$$\boldsymbol{\theta} - \beta \sum_{i}^{n} s_i \cdot \frac{d\boldsymbol{\phi}_i}{d\boldsymbol{\theta}} \nabla_{\boldsymbol{\phi}_i} \mathcal{L}(\boldsymbol{\phi}_i, \boldsymbol{X}'_t). \quad (7)$$

In summary, MetaAdapt computes task and meta gradients using sampled source tasks and few-shot target examples. Then, task similarity is computed to find more 'informative' source tasks, followed by tempered rescaling of the meta gradients. Finally, the updated model parameters should exploit the source domain knowledge and demonstrate improved performance on the target data distribution. The overall framework of MetaAdapt is illustrated in Figure 2. Unlike previous works (Motiian et al., 2017; Tseng et al., 2019; Zhao et al., 2021; Yue et al., 2022), we discard domain-adversarial or feature regularization methods, instead, we propose to leverage meta adaptation to guide the knowledge transfer from the source to target domain. Additionally, similarity-based gradients rescaling is designed to exploit different source tasks to achieve fine-grained adaptation performance.

5 Experiments

5.1 Settings

Model: Similar to (Li et al., 2021; Yue et al., 2022), we select RoBERTa as the base model to encode input examples in MetaAdapt. RoBERTa is a transformer model pretrained on a variety of NLP tasks before the COVID pandemic (Liu et al., 2019).

Evaluation: To validate the proposed method, we follow (Kou et al., 2022a; Li et al., 2021; Yue et al., 2022) and split the datasets into training, validation and test sets. The few-shot target examples are selected as the first k examples in the validation set and the rest validation examples are used for validating the model. For evaluation metrics, we adopt balance accuracy (BA), accuracy (Acc.) and F1 score (F1) to evaluate the performance. See evaluation details in Appendix A.

Datasets and Baselines: To examine MetaAdapt performance, we adopt multiple source and target datasets. We follow (Yue et al., 2022) and adopt FEVER (FE) (Thorne et al., 2018), GettingReal (GR) (Risdal, 2016), GossipCop (GC) (Shu et al., 2020a), LIAR (LI) (Wang, 2017) and PHEME (PH) (Buntain and Golbeck, 2017) as the source datasets. For the target domain, we adopt CoAID (Cui and Lee, 2020), Constraint (Patwa et al., 2021) and ANTiVax (Hayawi et al., 2022). Our naïve baseline leverages few-shot target examples to fine-tune the source pretrained models. We also adopt state-of-the-art baselines from domain adaptation and few-shot learning methods for domain adaptive few-shot misinformation detection: CANMD, ACLR, ProtoNet and MAML (Finn et al., 2017; Snell et al., 2017; Lin et al., 2022; Yue et al., 2022). Additionally, we select two large language models LLaMA and Alpaca to evaluate few-shot in-context learning (ICL) and parameter-efficient fine-tuning (PEFT) performance in misinformation detection (Touvron et al., 2023; Taori et al., 2023).

Implementation: We follow the preprocessing pipeline as in (Yue et al., 2022). Specifically, we translate special symbols (e.g., emojis) back into English, tokenize hashtags, mentions and URLs and remove special characters from the input. We use the 10-shot setting (i.e., $k = 10$), the model is trained using AdamW optimizer with 0.01 weight decay and no warm-up, where we sample 3 source tasks and perform 3 updates in inner-loop optimization. Then, we compute the meta loss and evaluate task similarity scores before rescaling the meta gradients with temperature τ. All our main experiments are repeated 3 times, we select the best model with the validation set for final evaluation. Hyperparameter selection (e.g., inner and outer learning rates, temperature etc.) and implementation details are provided in Appendix A.

Domain	Dataset	BA ↑	Acc. ↑	F1 ↑
Source	FEVER	0.796	0.796	0.817
	GettingReal	0.846	0.959	0.978
	GossipCop	0.776	0.869	0.917
	LIAR	0.607	0.632	0.712
	PHEME	0.863	0.867	0.898
Target	CoAID	0.889	0.972	0.985
	Constraint	0.970	0.971	0.973
	ANTiVax	0.932	0.921	0.931

Table 1: Supervised experiment results. The upper and lower parts report source and target dataset performance.

5.2 Main Results

Supervised results: We first report *supervised* results on all datasets in Table 1. The upper and

Source	Target	CoAID (2020)			Constraint (2021)			ANTiVax (2022)		
	Metric	**BA ↑**	**Acc. ↑**	**F1 ↑**	**BA ↑**	**Acc. ↑**	**F1 ↑**	**BA ↑**	**Acc. ↑**	**F1 ↑**
FE	Naïve	0.636	0.928	0.962	0.501	0.524	0.687	0.559	0.627	0.741
	CANMD	0.626	0.918	0.956	0.684	0.683	0.686	0.650	0.679	0.749
	ACLR	0.721	0.935	0.965	0.648	0.651	0.697	0.739	0.758	0.805
	ProtoNet	0.751	0.869	0.925	0.784	0.788	0.812	0.748	0.716	0.718
	MAML	0.780	**0.939**	**0.967**	0.812	0.808	0.797	0.826	0.808	0.823
	Ours	**0.829**$_{\pm 0.020}$	0.875$_{\pm 0.049}$	0.927$_{\pm 0.031}$	**0.828**$_{\pm 0.001}$	**0.826**$_{\pm 0.001}$	**0.829**$_{\pm 0.004}$	**0.868**$_{\pm 0.025}$	**0.880**$_{\pm 0.036}$	**0.904**$_{\pm 0.037}$
GR	Naïve	0.574	0.920	0.958	0.500	0.503	0.670	0.558	0.627	0.741
	CANMD	0.669	0.935	0.965	0.744	0.742	0.737	0.582	0.632	0.729
	ACLR	0.693	0.928	0.961	0.683	0.689	0.736	0.660	0.695	0.766
	ProtoNet	0.720	0.639	0.757	0.672	0.664	0.608	0.736	0.756	0.804
	MAML	0.813	**0.937**	**0.965**	0.808	0.803	0.786	0.819	0.802	0.819
	Ours	**0.830**$_{\pm 0.062}$	0.928$_{\pm 0.004}$	0.960$_{\pm 0.003}$	**0.819**$_{\pm 0.012}$	**0.819**$_{\pm 0.010}$	**0.823**$_{\pm 0.006}$	**0.886**$_{\pm 0.035}$	**0.882**$_{\pm 0.042}$	**0.902**$_{\pm 0.043}$
GC	Naïve	0.612	0.927	0.961	0.513	0.536	0.693	0.561	0.629	0.742
	CANMD	0.685	0.931	0.963	0.802	0.803	0.817	0.761	0.777	0.823
	ACLR	0.687	**0.933**	**0.964**	0.712	0.715	0.744	0.811	0.809	0.835
	ProtoNet	0.708	0.609	0.731	0.786	0.782	0.770	0.730	0.715	0.736
	MAML	0.816	0.926	0.959	0.813	0.809	0.801	0.826	0.810	0.826
	Ours	**0.824**$_{\pm 0.026}$	0.918$_{\pm 0.004}$	0.954$_{\pm 0.002}$	**0.826**$_{\pm 0.023}$	**0.826**$_{\pm 0.023}$	**0.833**$_{\pm 0.023}$	**0.896**$_{\pm 0.001}$	**0.907**$_{\pm 0.000}$	**0.930**$_{\pm 0.000}$
LI	Naïve	0.640	0.926	0.960	0.516	0.538	0.693	0.558	0.626	0.741
	CANMD	0.770	0.894	0.940	0.815	0.814	0.818	0.755	0.784	0.834
	ACLR	0.766	0.938	0.966	0.756	0.760	0.786	0.805	0.793	0.814
	ProtoNet	0.793	0.910	0.950	0.738	0.746	0.788	0.599	0.576	0.581
	MAML	0.813	**0.938**	**0.966**	0.813	0.809	0.800	0.824	0.807	0.824
	Ours	**0.815**$_{\pm 0.031}$	0.910$_{\pm 0.014}$	0.949$_{\pm 0.008}$	**0.820**$_{\pm 0.008}$	**0.820**$_{\pm 0.006}$	**0.828**$_{\pm 0.002}$	**0.873**$_{\pm 0.026}$	**0.883**$_{\pm 0.036}$	**0.906**$_{\pm 0.038}$
PH	Naïve	0.622	0.929	0.962	0.502	0.526	0.688	0.558	0.627	0.742
	CANMD	0.531	0.938	0.967	0.559	0.565	0.624	0.653	0.676	0.740
	ACLR	0.709	0.939	0.967	0.716	0.719	0.746	0.733	0.754	0.804
	ProtoNet	0.721	0.780	0.867	0.693	0.686	0.644	0.628	0.635	0.687
	MAML	0.780	**0.939**	**0.967**	0.816	0.812	0.802	0.819	0.805	0.823
	Ours	**0.828**$_{\pm 0.028}$	0.909$_{\pm 0.009}$	0.949$_{\pm 0.005}$	**0.818**$_{\pm 0.012}$	**0.818**$_{\pm 0.012}$	**0.828**$_{\pm 0.013}$	**0.896**$_{\pm 0.016}$	**0.880**$_{\pm 0.032}$	**0.902**$_{\pm 0.030}$
Avg	Naïve	0.617	0.926	0.961	0.506	0.525	0.686	0.559	0.627	0.741
	CANMD	0.656	0.923	0.958	0.721	0.721	0.736	0.680	0.710	0.775
	ACLR	0.715	0.935	0.965	0.703	0.707	0.742	0.750	0.762	0.805
	ProtoNet	0.739	0.761	0.846	0.735	0.733	0.724	0.688	0.679	0.705
	MAML	0.800	**0.936**	**0.965**	0.813	0.808	0.797	0.823	0.806	0.823
	Ours	**0.825**$_{\pm 0.033}$	0.908$_{\pm 0.016}$	0.948$_{\pm 0.010}$	**0.822**$_{\pm 0.011}$	**0.822**$_{\pm 0.010}$	**0.828**$_{\pm 0.010}$	**0.884**$_{\pm 0.021}$	**0.886**$_{\pm 0.029}$	**0.909**$_{\pm 0.030}$

Table 2: 10-shot cross-domain experiment results, the best and second best results are in bold and underlined. FE, GR, GC, LI and PH represent the source datasets FEVER, GettingReal, GossipCop, LIAR and PHEME.

lower parts of the table report the source and target dataset performance respectively. We observe the following: (1) overall, the performance on statements and posts achieves better performance than news articles. An example can be found on GossipCop (News) with 0.776 BA, compared to 0.863 on PHEME (Social network posts); (2) for disproportionate label distributions (e.g., CoAID), the BA metric reduces drastically compared to other metrics, indicating the difficulty of training fair models on unfair distributions. For instance, BA is circa 10% lower than accuracy and F1 on CoAID; (3) the RoBERTa model only achieves an average BA of 0.778 on source datasets, which suggests that transferring knowledge from the source to the target datasets can be a challenging task.

Few-shot adaptation results: The few-shot cross-domain experiments (10-shot) on all source-target combinations are presented in Table 2. In the table, rows represent the source datasets while the columns represent the target datasets. We include the adaptation methods in each row, while the metrics are reported in the columns under target datasets. For MetaAdapt, we report the mean results in the table and provide the standard deviation values using the $\pm$ sign. For convenience, the best results are marked in bold and the second best results are underlined. We observe: (1) adapting to the COVID domain is a non-trivial task upon dissimilar source-target label distributions. In the example of GossipCop $\rightarrow$ CoAID, baseline methods show BA values to be slightly higher than 0.6 (despite high accuracy and F1), suggesting that the model predicts the majority class with a

Setting	Dataset	BA ↑	Acc. ↑	F1 ↑
	CoAID	0.500	0.906	0.951
LLaMA-ICL	Constraint	0.500	0.523	0.687
	ANTiVax	0.500	0.664	0.798
	CoAID	0.515	0.908	0.952
Alpaca-ICL	Constraint	0.537	0.559	0.704
	ANTiVax	0.528	0.681	0.806
	CoAID	0.749	0.874	0.928
LLaMA-FT	Constraint	0.724	0.721	0.718
	ANTiVax	0.742	0.756	0.811
	CoAID	0.766	0.818	0.892
Alpaca-FT	Constraint	0.688	0.686	0.689
	ANTiVax	0.767	0.779	0.828
	CoAID	0.825	0.908	0.948
MetaAdapt	Constraint	0.822	0.822	0.828
	ANTiVax	0.884	0.886	0.909

Table 3: Comparison to large language models.

Setting	Dataset	BA ↑	Acc. ↑	F1 ↑
	CoAID	0.551	0.868	0.925
0-shot	Constraint	0.567	0.585	0.705
	ANTiVax	0.528	0.590	0.689
	CoAID	0.594	0.450	0.588
1-shot	Constraint	0.664	0.662	0.656
	ANTiVax	0.627	0.616	0.645
	CoAID	0.728	0.904	0.949
5-shot	Constraint	0.799	0.796	0.792
	ANTiVax	0.700	0.721	0.776
	CoAID	0.825	0.908	0.948
10-shot	Constraint	0.822	0.822	0.828
	ANTiVax	0.884	0.886	0.909
	CoAID	0.876	0.938	0.964
15-shot	Constraint	0.844	0.844	0.870
	ANTiVax	0.865	0.850	0.872

Table 4: Sensitivity on the number of target examples.

much higher likelihood; (2) by learning domain-invariant feature representation, baseline methods like CANMD and ACLR can improve the adaptation results compared to the naïve baseline. For instance, CANMD achieves over 42.3% and 21.7% average improvements on BA for Constraint and ANTiVax; (3) using the meta adaptation approach, MetaAdapt significantly outperform all baselines in the BA metric. For instance, MetaAdapt outperforms the best-performing baseline MAML 7.4% in BA on ANTiVax, with similar trends to be found in accuracy and F1. Specifically for CoAID (with over 90% positive labels), improvements on BA demonstrates that MetaAdapt can learn fair features for improved detection results despite slightly worse accuracy and F1 results. In summary, the results in Table 2 show that MetaAdapt is particularly effective in adapting early misinformation detection systems using limited target examples. In contrast to the baseline models, MetaAdapt can achieve significant improvements on all metrics across source-target dataset combinations. In the case of large domain discrepancy, MetaAdapt demonstrates superior performance by exploiting the few-shot target examples with second-order dynamics and similarity-based adaptive learning.

Comparison to large language models: To further demonstrate the effectiveness of MetaAdapt in domain adaptive misinformation detection, we compare our MetaAdapt with state-of-the-art large language models (LLMs). In particular, we adopt LLaMA-7B and Alpaca-7B and perform both few-shot in-context learning (i.e., ICL) and parameter-efficient fine-tuning (i.e., FT) on target datasets, with results presented in Table 3. We notice: (1) despite the significant increase in model parameters (from $\sim$0.1B to 7B), LLMs can still fail to distinguish misinformation without further tuning. For instance, LLaMA consistently scores 0.5 in BA by predicting the majority class in in-context learning. (2) Fine-tuning LLMs can significantly improve the performance in misinformation detection. With PEFT tuning, Alpaca achieves 40.7% performance improvement in BA across target datasets conditioned only on the few-shot target examples. (3) With the proposed MetaAdapt, smaller language models like RoBERTa are capable of outperforming fine-tuned large language models. On average, MetaAdapt outperforms LLaMA-PEFT and Alpaca-PEFT by 14.3% and 14.1% in the BA metric on target datasets. The results suggest that MetaAdapt is both effective and efficient in early misinformation detection by combining out-of-domain knowledge and meta adaptation.

Robustness study: We also evaluate the sensitivity of MetaAdapt with respect to the number of few-shot examples. In particular, we select the number from 0 to 15 and use MetaAdapt to perform the adaptation. The results are averaged across the source datasets and reported in Table 4. Note that for 0-shot experiments, we train the models on source datasets and directly evaluate on target data. We observe the following: (1) as expected, the adaptation performance improves rapidly with increasing number of few-shot examples. (2) Surprisingly, we observe performance drops in accuracy and F1 on CoAID when increasing the few-shot number from 0 to 1. This suggests that the large domain discrepancy between both domains may reduce the effectiveness of MetaAdapt; (3) Overall,

Metric	CoAID (2020)			Constraint (2021)			ANTiVax (2022)		
	BA ↑	**Acc. ↑**	**F1 ↑**	**BA ↑**	**Acc. ↑**	**F1 ↑**	**BA ↑**	**Acc. ↑**	**F1 ↑**
MetaAdapt	0.825	0.908	0.948	0.822	0.822	0.828	0.884	0.886	0.909
w/o Task Similarity	0.811	0.909	0.948	0.807	0.805	0.806	0.862	0.853	0.880
w/o Adaptive LR	0.801	0.919	0.954	0.811	0.808	0.801	0.860	0.860	0.885
w/o 2nd-order Grads	0.789	0.928	0.960	0.800	0.794	0.774	0.843	0.843	0.871

Table 5: Ablation study results.

the magnitude of improvements grows rapidly at first and then slowly plateaus as we increase the few-shot number. The largest improvements can be found from 1-shot to 5-shot setting, where the BA results improve by 18.1% on average. In sum, we observe that even limited number of target examples can improve the target domain performance with MetaAdapt. We provide additional hyperparameters sensitivity analysis in Appendix B.

Ablation study: We evaluate the effectiveness of the proposed components in MetaAdapt. In particular, we remove the proposed similarity-based gradients rescaling (w/o Task Similarity), adaptive learning rate (w/o Adaptive LR) and second-order optimization method (w/o 2nd-order Grads) in order and observe the performance changes. Experiment results on target datasets are averaged across the source datasets and are reported in Table 5. For all components, we observe performance drops on BA when removed from MetaAdapt[2]. For example, the performance of MetaAdapt reduces by 2.0% and 2.3% in BA when we remove task similarity and adaptive LR consecutively. We further replace MetaAdapt with first-order approximation, which results in a performance drop of 3.9% in the BA metric. Overall, the results suggest that the proposed components are effective for domain adaptive few-shot misinformation detection.

6 Conclusion

In this paper, we explore meta learning in domain adaptive few-shot misinformation detection. We propose MetaAdapt, a meta learning-based approach for cross-domain few-shot adaptation. MetaAdapt is the first to leverage few-shot target examples for exploiting the source domain knowledge under the guidance of limited target examples. Experiment results demonstrate the effectiveness of our method by achieving promising results on multiple challenging source-target dataset combinations, where MetaAdapt can significant outperform the state-of-the-art baselines and large language models by a significant margin.

[2]In the case of CoAID, Acc and F1 are less reliable evaluation metrics due to the imbalanced distribution of labels.

7 Limitations

Despite introducing meta learning for domain adaptive few-shot misinformation detection, we have not discussed the setting of cross-domain adaptation with multiple source datasets to further improve the performance for identifying early-stage misinformation. Due to the lack of early-stage misinformation data, we limit our choice of the target domain to COVID-19, which may hinder the generalization of the proposed method to other domains. Additionally, the proposed method does not leverage efficient approximation or first-order meta learning methods to reduce the computational costs in training MetaAdapt. As such, we plan to explore multi-source few-shot misinformation detection via efficient meta learning as future work.

Acknowledgments

This research is supported in part by the National Science Foundation under Grant No. IIS-2202481, CHE-2105005, IIS-2008228, CNS-1845639, CNS-1831669. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Antreas Antoniou, Harrison Edwards, and Amos Storkey. 2018. How to train your maml. In *International Conference on Learning Representations*.

Erik Brand, Kevin Roitero, Michael Soprano, and Gianluca Demartini. 2021. E-bart: Jointly predicting and explaining truthfulness. In *TTO*, pages 18–27.

Cody Buntain and Jennifer Golbeck. 2017. Automatically identifying fake news in popular twitter threads.

In *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, pages 208–215. IEEE.

Canyu Chen, Haoran Wang, Matthew Shapiro, Yunyu Xiao, Fei Wang, and Kai Shu. 2022. Combating health misinformation in social media: Characterization, detection, intervention, and open issues. *arXiv preprint arXiv:2211.05289*.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 492–502.

Sourya Dipta Das, Ayan Basak, and Saikat Dutta. 2021. A heuristic-driven ensemble framework for covid-19 fake news detection. In *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 164–176. Springer.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

Chengcheng Han, Zeqiu Fan, Dongxiang Zhang, Minghui Qiu, Ming Gao, and Aoying Zhou. 2021. Meta-learning adversarial domain adaptation network for few-shot text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1664–1673, Online. Association for Computational Linguistics.

Kadhim Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2022. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public health*, 203:23–30.

Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. Reinforcement learning-based counter-misinformation response generation: A case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, pages 2698–2709.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021a. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021b. Compare to the knowledge: Graph neural fake news detection with external knowledge. 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 754–763, Online. Association for Computational Linguistics.

Gongyao Jiang, Shuang Liu, Yu Zhao, Yueheng Sun, and Meishan Zhang. 2022. Fake news detection via knowledgeable prompt learning. *Information Processing & Management*, 59(5):103029.

Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. 2019. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4893–4902.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3432–3442, Minneapolis, Minnesota. Association for Computational Linguistics.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921.

Boshko Koloski, Timen Stepišnik Perdih, Marko Robnik-Šikonja, Senja Pollak, and Blaž Škrlj. 2022. Knowledge graph informed fake news classification via heterogeneous representation ensembles. *Neurocomputing*.

Ziyi Kou, Lanyu Shang, Yang Zhang, and Dong Wang. 2022a. Hc-covid: A hierarchical crowdsource knowledge graph approach to explainable covid-19 misinformation detection. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–25.

Ziyi Kou, Lanyu Shang, Yang Zhang, Christina Youn, and Dong Wang. 2021. Fakesens: A social sensing approach to covid-19 misinformation detection on social media. In *2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 140–147. IEEE.

Ziyi Kou, Lanyu Shang, Yang Zhang, Zhenrui Yue, Huimin Zeng, and Dong Wang. 2022b. Crowd, expert & ai: A human-ai interactive approach towards natural language explanation based covid-19 misinformation detection. In *Proc. Int. Joint Conf. Artif. Intell.(IJCAI)*, pages 5087–5093.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021. Towards few-shot fact-checking via perplexity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1971–1981, Online. Association for Computational Linguistics.

Tianyi Lei, Honghui Hu, Qiaoyang Luo, Dezhong Peng, and Xu Wang. 2022. Adaptive meta-learner via gradient similarity for few-shot text classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4873–4882, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shuang Li, Shiji Song, Gao Huang, Zhengming Ding, and Cheng Wu. 2018. Domain invariant and class discriminative feature learning for visual domain adaptation. *IEEE transactions on image processing*, 27(9):4260–4273.

Yichuan Li, Kyumin Lee, Nima Kordzadeh, Brenton Faber, Cameron Fiddes, Elaine Chen, and Kai Shu. 2021. Multi-source domain adaptation with weak supervision for early fake news detection. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 668–676. IEEE.

Hongzhan Lin, Jing Ma, Liangliang Chen, Zhiwei Yang, Mingfei Cheng, and Chen Guang. 2022. Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2543–2556, Seattle, United States. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.

Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. NLP-based feature extraction for the detection of COVID-19 misinformation videos on YouTube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. 2019. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.

Ahmadreza Mosallanezhad, Mansooreh Karami, Kai Shu, Michelle V Mancenido, and Huan Liu. 2022. Domain adaptive fake news detection via reinforcement learning. In *Proceedings of the ACM Web Conference 2022*, pages 3632–3640.

Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. 2017. Few-shot adversarial domain adaptation. *Advances in neural information processing systems*, 30.

Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. 2021. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1094–1103.

Qiong Nan, Danding Wang, Yongchun Zhu, Qiang Sheng, Yuhui Shi, Juan Cao, and Jintao Li. 2022. Improving fake news detection of influential domain via domain- and instance-level transfer. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2834–2848, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an infodemic: Covid-19 fake news dataset. In *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 21–29. Springer.

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. 2019. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32.

Megan Risdal. 2016. Getting real about fake news.

Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. 2020. Susceptibility to misinformation about covid-19 around the world. *Royal Society open science*, 7(10):201199.

Minho Ryu, Geonseok Lee, and Kichun Lee. 2022. Knowledge distillation for bert unsupervised domain adaptation. *Knowledge and Information Systems*, 64(11):3113–3128.

Fatima K Abu Salem, Roaa Al Feel, Shady Elbassuoni, Hiyam Ghannam, Mohamad Jaber, and May Farah. 2021. Meta-learning for fake news detection surrounding the syrian war. *Patterns*, 2(11):100369.

Nikita Mariam Santhosh, Jo Cheriyan, and Lekshmi S Nair. 2022. A multi-model intelligent approach for rumor detection in social networks. In *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, pages 1–5. IEEE.

Lanyu Shang, Ziyi Kou, Yang Zhang, Jin Chen, and Dong Wang. 2022a. A privacy-aware distributed knowledge graph approach to qois-driven covid-19 misinformation detection. In *2022 IEEE/ACM 30th International Symposium on Quality of Service (IWQoS)*, pages 1–10. IEEE.

Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2021. A multimodal misinformation detector for covid-19 short videos on tiktok. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 899–908. IEEE.

Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2022b. A duo-generative approach to explainable multimodal covid-19 misinformation detection. In *Proceedings of the ACM Web Conference 2022*, pages 3623–3631.

Lanyu Shang, Yang Zhang, Zhenrui Yue, YeonJung Choi, Huimin Zeng, and Dong Wang. 2022c. A knowledge-driven domain adaptive approach to early misinformation detection in an emergent health domain on social media. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 34–41. IEEE.

Amr Sharaf, Hany Hassan, and Hal Daumé III. 2020. Meta-learning for few-shot NMT adaptation. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 43–53, Online. Association for Computational Linguistics.

Kendrick Shen, Robbie M Jones, Ananya Kumar, Sang Michael Xie, Jeff Z HaoChen, Tengyu Ma, and Percy Liang. 2022. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 19847–19878. PMLR.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020a. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. 2020b. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 626–637.

Kai Shu, Ahmadreza Mosallanezhad, and Huan Liu. 2022. Cross-domain fake news detection on social media: A context-aware adversarial approach. In *Frontiers in Fake Media Generation and Detection*, pages 215–232. Springer.

Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2020c. *Disinformation, misinformation, and fake news in social media*. Springer.

Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 312–320.

Anthony Sicilia, Xingchen Zhao, and Seong Jae Hwang. 2021. Domain adversarial neural networks for domain generalization: When it works and how to improve. *arXiv preprint arXiv:2102.03924*.

Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 557–565.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Abhijit Suprem and Calton Pu. 2022. Midas: Multi-integrated domain adaptive supervision for fake news detection. *arXiv preprint arXiv:2205.09817*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. 2019. Cross-domain few-shot classification via learned feature-wise transformation. In *International Conference on Learning Representations*.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Nguyen Vo and Kyumin Lee. 2020. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7717–7731, Online. Association for Computational Linguistics.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.

Junfei Wu, Weizhi Xu, Qiang Liu, Shu Wu, and Liang Wang. 2022a. Adversarial contrastive learning for evidence-aware fake news detection with graph neural networks. *arXiv preprint arXiv:2210.05498*.

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.

Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022b. Cross-document misinformation detection based on event graph reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 543–558, Seattle, United States. Association for Computational Linguistics.

Zhenyu Wu, YaoXiang Wang, Jiacheng Ye, Jiangtao Feng, Jingjing Xu, Yu Qiao, and Zhiyong Wu. 2023. Openicl: An open-source framework for in-context learning. *arXiv preprint arXiv:2303.02913*.

Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, Xinjing Cheng, and Guoren Wang. 2022. Active learning for domain adaptation: An energy-based approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8708–8716.

Zhenrui Yue, Huimin Zeng, Ziyi Kou, Lanyu Shang, and Dong Wang. 2022. Contrastive domain adaptation for early misinformation detection: A case study on covid-19. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2423–2433.

Huimin Zeng, Zhenrui Yue, Ziyi Kou, Lanyu Shang, Yang Zhang, and Dong Wang. 2022. Unsupervised domain adaptation for covid-19 information service with contrastive adversarial domain mixup. pages 159–162.

Qiang Zhang, Hongbin Huang, Shangsong Liang, Zaiqiao Meng, and Emine Yilmaz. 2021. Learning to detect few-shot-few-clue misinformation. *arXiv preprint arXiv:2108.03805*.

Tong Zhang, Di Wang, Huanhuan Chen, Zhiwei Zeng, Wei Guo, Chunyan Miao, and Lizhen Cui. 2020. Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

An Zhao, Mingyu Ding, Zhiwu Lu, Tao Xiang, Yulei Niu, Jiechao Guan, and Ji-Rong Wen. 2021. Domain-adaptive few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1390–1399.

Pan Zhou, Yingtian Zou, Xiao-Tong Yuan, Jiashi Feng, Caiming Xiong, and Steven Hoi. 2021. Task similarity aware meta learning: Theory-inspired improvement on maml. In *Uncertainty in Artificial Intelligence*, pages 23–33. PMLR.

A Implementation

A.1 Datasets

We adopt FEVER (Thorne et al., 2018), GettingReal (Risdal, 2016), GossipCop (Shu et al., 2020a), LIAR (Wang, 2017) and PHEME (Buntain and Golbeck, 2017) as the source datasets. For the target domain, we adopt three COVID datasets: ANTiVax (2022) (Hayawi et al., 2022), CoAID (2020) (Cui and Lee, 2020) and Constraint (2021) (Patwa et al., 2021).

In the following, we present the details of our source and target datasets:

1. **FEVER** is a publicly available dataset for fact verification. FEVER consists of modified claims from Wikipedia without knowledge of the claims (Thorne et al., 2018).
2. **GettingReal** is a fake news dataset from Kaggle (Getting Real about Fake News). GettingReal contains text and metadata scraped from online resources (Risdal, 2016).
3. **GossipCop** provides a fake news data from news content, social network, and dynamic iteratcions. GossipCop is adopted from the FakeNewsNet dataset (Shu et al., 2020a).
4. **LIAR** is a publicly available dataset of fact verification. The provided statements are collected from politifact.com with analysis report and links to sources (Wang, 2017).
5. **PHEME** contains tweets rumours and non-rumours in certain breaking events (e.g., Germanwings crash). PHEME provides online interactions and structure of the tweets (Buntain and Golbeck, 2017).
6. **CoAID** is a dataset with COVID-19 misinformation. CoAID provides fake news on websites and social platforms as well as user iteractions under such sources (Cui and Lee, 2020).
7. **Constraint** is a shared taks on COVID-19 fake news detection. It contains over 10k annotated social media posts and articles of real and fake news on COVID-19 (Patwa et al., 2021).
8. **ANTiVax** is a novel dataset with over 15k COVID-19 vaccine-related tweets and annotations for vaccine misinformation detection (Hayawi et al., 2022).

Details of the above datasets are provided in Table 6, where **Neg.** and **Pos.** are the proportion of misinformation and valid information in the dataset (i.e., label distribution). **Len** represents the average token length of text and **Content** denotes the source type of the text (e.g., statement, news or social posts). Notice that CoAID is largely imbalanced with over 90% positive examples.

Datasets	Neg.	Pos.	Len	Content
FEVER	29.6%	70.4%	9.4	Statement
GettingReal	8.8%	91.2%	738.9	News
GossipCop	24.2%	75.8%	712.9	News
LIAR	44.2%	55.8%	20.2	Statement
PHEME	34.0%	66.0%	21.5	Social Network
CoAID	9.7%	90.3%	54.0	News / Statement
Constraint	47.7%	52.3%	32.7	Social Network
ANTiVax	38.3	61.7%	26.2	Social Network

Table 6: Details of the involved datasets.

A.2 Baseline Methods

As a naïve baseline, we pretrained the misinformation detection model on the source dataset and fine-tune with the few-shot examples, followed by evaluation on the test set. We additionally adopt the following state-of-the-art baselines for domain adaptive few-shot misinformation detection:

1. **Contrastive Adaptation Network for Misinformation Detection (CANMD)** proposes to use label correction in the pseudo-labeling process to generate labeled target examples. Then, contrastive learning is applied to learn domain-invariant and class-separating features. Therefore, we adapt CANMD by including the few-shot target examples in the training process as in our setting, we select the best results from the original CANMD and our adaptation (Yue et al., 2022).
2. **Adversarial Contrastive Learning for low-resource rumor detection (ACLR)** leverages language alignment and contrastive learning to improve corss-domain misinformation detection performance. ACLR also introduces adversarial augmentation to enhance the robustness of few-shot rumor detection. We replace the original graph convolution networks with our base transformer model for content-based misinformation detection (Lin et al., 2022).
3. **Prototypical Networks for Few-shot Learning (ProtoNet)** designs prototypical networks for few-shot classification problems. ProtoNet projects sample features to a metric space and perform inference by computing distances to prototypes of each class. We adapt ProtoNet to meta adaptation framework by adopting the same label space and the base transformer model as encoder for domain-adaptive few-shot misinformation detection (Snell et al., 2017).

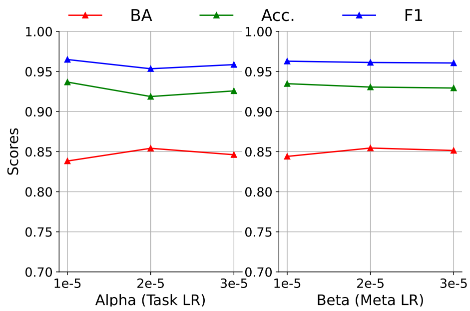

Figure 3: Sensitivity of the initial learning rates.

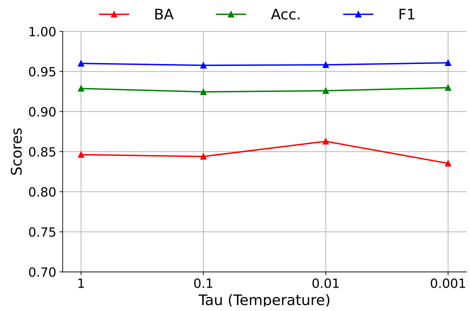

Figure 4: Sensitivity of temperature values.

4. **Model-Agnostic Meta-Learning (MAML)** is the first to leverage second-order derivatives for few-shot learning. MAML first update model parameters upon sampled tasks, followed by computing the meta loss and derive the second-order gradients w.r.t. the original parameters. Similarly, We adapt MAML to our meta adaptation framework with the homogeneous label space acorss tasks and the transformer encoder for misinformation detection (Finn et al., 2017).

A.3 Implementation Details

For our evaluation method, we follow the previous works (Kou et al., 2022a; Li et al., 2021; Yue et al., 2022) and split the datasets into training, validation, and test sets with the ratio of 7:2:1. If the dataset provides a default split, we directly use the provided split sets. The validation sets are used for label correction in (Yue et al., 2022) and constructing few-shot examples and saving the best model in training. For our few-shot adaptation setting, we select the first k examples (10-shot as default) from the original validation set and used the remaining examples for validation. We use accuracy and F1 score for evaluation. The balanced accuracy (BA) is additionally introduced to evaluate the adaptation performance in both classes equally, BA is defined as the mean of sensitivity and specificity.

For the naïve baseline, we train the base model on the source dataset using AdamW optimizer (learning rate 1e-5) without warm-up. Then, the model is fine-tuned on the few-shot examples under the same training condition. For other baseline methods, we use the original implementation (if provided) and follow the original hyperparameter configuration. Otherwise we reimplement the baseline methods and adopt the identical training pipeline (AdamW with 1e-5 learning rate and no warm-up). Few-shot learning baselines (i.e., ProtoNet & MAML) are adapted to the domain adaptive few-shot learning framework by using the few-shot target examples as query set. For the large language models, in-context learning is performed using the generative and perplexity-based approach based on (Wu et al., 2023), while fine-tuning is performed using low-rank adaptation as in (Hu et al., 2021a). In the MetaAdapt experiments, we adopt 3 as the number of tasks n and update the task specific model ϕ for 3 times in inner-level optimization, notice for each task we start from the original parameter set $\boldsymbol{\theta}$. To train MetaAdapt, we use a training batch size of 4 in each task and the learning rates (both α and β) are selected from $[1e-5, 2e-5, 3e-5]$. The temperature hyperparameter is selected from $[1, 0.1, 0.01]$ (See sensitivity analysis in Appendix B). We perform training with 500 or 1000 meta iterations and validate the model every 50 iterations, the best model is used for evaluation on the test sets.

B Additional Results

Sensitivity Analysis of Initial Learning Rates: We study the sensitivity of α and β on the CoAID dataset, the best results (averaged across source datasets) are presented in Figure 3. Overall, we observe that the performance of MetaAdapt is insensitive to the changes of both learning rates. Interestingly, we notice that accuracy values are negatively correlated with the BA scores, suggesting that accuracy may not be an ideal metric for data distributions with disproportionate label classes.
Sensitivity Analysis of Temperature Values: We also study the sensitivity of τ in gradient rescaling on CoAID, we present the best results (averaged across source datasets) in Figure 4. In short, the performance of MetaAdapt first increases and then starts to reduce, with the best results ocurring at $\tau = 0.01$, indicating the effectiveness of similarity-based gradients rescaling. Overall, the proposed MetaAdapt is robust to the hyperparameters and outperforms the baseline methods consistently.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7 Limitations

☑ A2. Did you discuss any potential risks of your work?
7 Limitations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

5 Experiments

☑ B1. Did you cite the creators of artifacts you used?
5 Experiments

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
5 Experiments & A.1 Datasets

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
5 Experiments & A.1 Datasets

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
5 Experiments & A.1 Datasets

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
A.1 Datasets

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
5 Experiments & A.1 Datasets

C ☑ Did you run computational experiments?

5 Experiments

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
5 Experiments

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
5 Experiments & A.3 Implementation Details

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5 Experiments

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5 Experiments & A.3 Implementation Details

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.