Unsupervised Domain Adaptation via Contrastive Adversarial Domain Mixup: A Case Study on COVID-19

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

Abstract—Training large deep learning (DL) models with high performance for natural language downstream tasks usually requires rich-labeled data. However, in a real-world application of COVID-19 information service (e.g., misinformation detection, question answering), a fundamental challenge is the lack of the labeled COVID data to enable supervised end-to-end training of the models for different downstream tasks, especially at the early stage of the pandemic. To address this challenge, we propose an unsupervised domain adaptation framework using contrastive learning and adversarial domain mixup to transfer the knowledge from an existing source data domain to the target COVID-19 data domain. In particular, to bridge the gap between the source domain and the target domain, our method reduces a radial basis function (RBF) based discrepancy between these two domains. Moreover, we leverage the power of domain adversarial examples to establish an intermediate domain mixup, where the latent representations of the input text from both domains could be mixed during the training process. In this paper, we focus on two prevailing downstream tasks in mining COVID-19 text data: COVID-19 misinformation detection and COVID-19 news question answering. Extensive domain adaptation experiments on multiple real-world datasets suggest that our method can effectively adapt misinformation detection and question answering systems to the unseen COVID-19 target domain with significant improvements compared to the state-of-the-art baselines.

Index Terms—Domain Adaptation, Contrastive Domain Mixup, Misinformation Detection, Question Answering

1 Introduction

Pre-trained language models [1], [2] have been proved to be an efficient method to improve the model's performance on many natural language processing (NLP) tasks on social media [3], [4], [5]. However, for downstream NLP tasks (e.g., misinformation detection and question answering) on a specific data domain, supervised training is usually required to fine-tune the pre-trained models on the target data domain to ensure the models' performance on such domain-specific tasks [6]. In this work, we focus on COVID-19 given its global impact of the ongoing pandemic and the "Infodemic"[1] it causes on social media [3]. Consider a real-world application of COVID-19 misinformation detection, if the language models trained on non-COVID datasets without any fine-tuning on COVID-19 specific data, these models might suffer from a severe issue of generalization and perform poorly on the COVID-19 datasets, due to the domain shift between the non-COVID training data distribution and the test COVID-19 data distribution.

Indeed, the ongoing pandemic of COVID-19 inspires a variety of studies [3], [7], [8] to develop NLP models to provide reliable COVID-19 information services across various social media platforms (e.g., Twitter, Facebook). However, the supervised learning approaches often require a large-scale training dataset while collecting annotations for COVID training data is extremely expensive and time consuming due to the cost and complexity in recruiting the qualified annotators and keep the annotations update to date to accommodate the dynamics of COVID-19 knowledge (e.g., different variants of the virus) [9]. Moreover, our unsupervised domain adaptation setting is motivated for a more general setting of any early-stage pandemic (not limited to COVID-19) where there is no ground-truth information about the novel disease at all, but the need for correct information is urgent. Therefore, it is critical to develop unsupervised domain adaptation frameworks to train COVID models so that knowledge from an existing

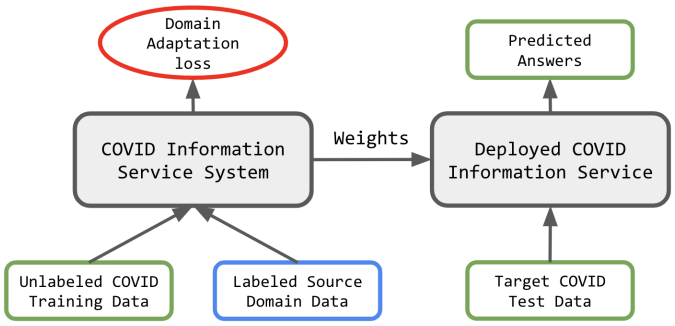


Figure 1: The Overview of Unsupervised Domain Adaptation. Labeled source domain data and unlabeled target domain data (COVID-19 training data) is available for domain adaptation training.

1. https://www.who.int/health-topics/infodemic#tab=tab_1

data domain could be adapted and transferred to the unseen COVID data domain without requiring any ground-truth training labels. The general unsupervised domain adaptation framework is shown in Figure 1. Under an unsupervised domain adaptation framework, labeled source domain data and unlabeled target domain data (COVID-19 training data) is available for domain adaptation training. After performing domain adaptation training, the adapted model is expected to produce high-quality predictions for the COVID-19 test data upon its deployment.

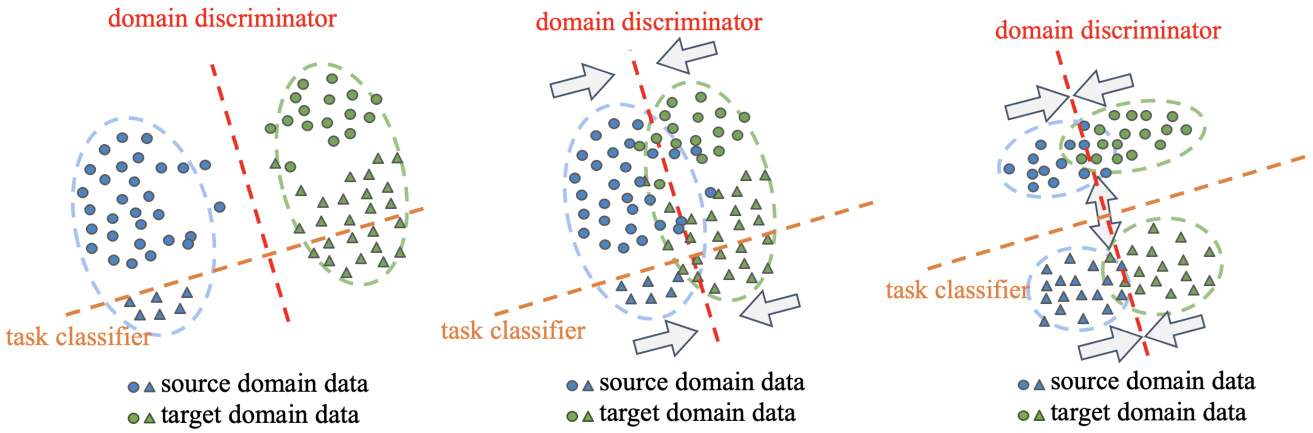
In this paper, we explore an unsupervised domain adaptation problem of COVID-19 information services on social media: *the COVID models trained on the label-rich source domain are adapted to an unlabeled target domain without requiring supervised training on the target domain.* To achieve this goal, we propose an unsupervised domain adaptation framework **C**ontrastive **A**dversarial **D**omain **M**ixup + (CADM+), which uses adversarial domain mixup and contrastive learning to bridge the gap between the source training domain and the target COVID domain. The overview of our framework is shown in Figure 2. Specifically, we firstly leverage pre-trained models to generate labeled target examples via pseudo labeling (Figure 2a). Next, we train a domain adversary to establish a learnable intermediate domain mixup to bridge the domain gap between source and target domains by perturbing latent representations of input texts from both domains towards each other (Figure 2b). Finally, we compute an RBF-based contrastive adaptation loss over the perturbed adversarial representations, and optimize it to encourage the model to learn class-aware features and further reduce the domain discrepancy (Figure 2c). Eventually, the COVID models could learn to project the target domain COVID-19 data into the learned smooth intermediate domain, and adapt knowledge from the source domain to make predictions for the data in the target domain.

To demonstrate the effectiveness of the proposed CADM+, we evaluate it in two real-world COVID-19 information services, namely COVID-19 misinformation detection and COVID-19 news question answering, which have been used widely and have significant impacts in the information space of pandemic [8], [9], [10], [11]. Regarding COVID-19 misinformation detection, it is shown in [7] that the widespread of COVID misinformation could pose a severe threat to the online ecosystem and the public health. For instance, in [7], the authors find that COVID-19 misinformation has a strong correlation between noncompliance of health guidance and reduced likelihood in receiving vaccines. In comparison, COVID-19 question answering (QA) systems automatically provide people with answers for their questions regarding certain COVID-19 texts, so that people do not need to read the entire document word-by-word to search for the answers [8]. Both services are in critical needs to public health and interest, but would originally require large amounts of labeled text data to enable end-to-end training of the corresponding misinformation detection models and the question answering models. For both COVID-19 information services, our experimental results suggest that our CADM+ effectively adapts pre-trained language models to the target COVID domain, and consistently outperforms state-of-the-art baselines on several real-world COVID-19 datasets (i.e., Constraint [11], ANTiVAx [10] and CoAID [8]).

A preliminary version of this work was presented in [12]. The current paper is a significant extension of the previous work in the following aspects. First, we extend our previous framework Contrastive Adversarial Domain Mixup (CADM) in [12] by explicitly exploring its deployment on a new COVID-19 information service (i.e., COVID-19 news question answering). In this paper, we refer the extended framework on both COVID-19 misinformation detection and COVID-19 news question answering as CADM+. In contrast, the conference paper only focuses on COVID-19 misinformation detection, which is a binary text classification problem. Second, to extend our CADM framework for the COVID-19 question answering (QA) problem (i.e., CADM+), we designed a new training pipeline and a new training loss. Under the new training pipeline, the training of the models (i.e., the CADM+ model and the domain discriminator) as well as training loss are defined over the text span of COVID data instead of focusing on classifying the `[CLS]` token in the misinformation detection task. As such, the QA model can efficiently learn COVID-19 data features from the unique data structure in the COVID-19 news QA application. Third, we added a new set of experiments to evaluate our proposed CADM+ framework, where one new linear domain mixup [13] is added for the misinformation detection application and several new baseline schemes [6], [14] are added for the COVID-19 news QA application. In particular, we firstly used a pre-trained question generation model to generate question-answer pairs for a set of true COVID-19 news. To this end, we created our new QA dataset for our social media based COVID-19 news QA application. Then, we implemented several state-of-the-art unsupervised domain adaptation baselines for question answering and compared our proposed framework against the baseline methods. Finally, we also extended the related work by reviewing the recent literature on unsupervised domain adaptation for question answering. We summarize the contributions of our work as follows [2]:

- We propose a novel unsupervised domain adaptation framework CADM+ for COVID-19 information services using contrastive learning and adversarial domain mixup.
- Our method learns a smoothed intermediate domain to transfer knowledge from the source domain to the target domain by perturbing the latent representations from both domains towards each other. Combined with contrastive adaptation loss, we bridge the gap between the source domain and the target COVID domain.
- To the best of our knowledge, our method is the first work that adopts latent adversarial examples to establish domain mixup and contrastive domain adaptation for adapting misinformation detection and question answering models to the unseen COVID domain.
- We demonstrate the effectiveness of our CADM+ on multiple real-world COVID-19 datasets, and our method outperforms the state-of-the-art baselines for

2. We adopt publicly available datasets in our experiments and will release the code upon publication of this work.



(a) Unadapted (b) Adversarial Domain Mixup (c) Contrastive Domain Adaptation

Figure 2: The Overview of Our Contrastive Adversarial Domain Mixup + (CADM+) for both Misinfomration Detection and Question Answering: firstly, in (a), a pre-trained models will generate labels for target domain examples via pseudo labeling, where the green triangles belong to one class (e.g., predicted true information in the misinformation detection task or predicted answer spans in given contexts in the question answering task) and the green dots belong to another class (e.g., predicted misinformation in the misinformation detection task or the predicted non-answer spans in the remaining contexts in the question answering task). In addition, a domain discriminator is trained. Then, in (b), the well-trained domain discriminator will establish an intermediate domain mixup by perturbing latent representations of input text from both domains towards each other. At the same time, in (c), we also compute a contrastive adaptation loss over the perturbed adversarial representations, and optimize it to further reduce the domain discrepancy and increase the models' performance. Note that (b) and (c) are executed alternately.

two critical COVID-19 information services considered.

2 Related Work

Misinformation Detection. Great efforts have been made to detect the misinformation from online platforms (e.g., social media). In content-based misinformation detection methods [15], [16], models are trained to extract linguistic features of input text to evaluate the credibility of information. In comparison, in [17], knowledge graphs are integrated into the misinformation detection framework to enhance the model's performance, since knowledge graphs could introduce additional information of the specific data domain for misinformation detection. However, these misinformation detection systems are built under a supervised learning setting [9], [18], but regarding COVID-19 misinformation detection, labeled COVID-19 misinformation data is not always accessible. Indeed, another thread of studies address the misinformation detection using unsupervised learning methods. For instance, in [19], a graph-based method is proposed to identify the seed set of fake and legitimate news and then perform progressive label spreading over the full dataset. In [20], [21], additional resources, such as user credibility or media credibility, are used to evaluate the trustworthiness of the news and build the misinformation detection systems. In [22], autoencoders are used to detect fake news by comparing reconstruction scores of fake news and true news. In [23], fake claims are detected by comparing the semantics between the claims and associated news sources. Nevertheless, the generalization and adaptability of such misinformation detection systems are not well studied. Therefore, we focus on domain adaptation of content-based language models for misinformation detection. Under *unsupervised* domain adaptation, the models are trained to adapted knowledge from a source labeled training dataset to the unknown COVID-19 dataset.

Question Answering. Question answering (QA) models predict answers conditioned on an input question and a context paragraph [1], [6]. In this paper, we formulate our COVID-19 QA problem as extractive QA on COVID-19 news, where the task is to extract answer spans from an unstructured COVID-19 news for a given question [1], [6]. We noted that existing works on COVID-19 question answering mainly focus on handling questions regarding long and difficult scientific articles [9]. However, we argue that for general public, the information from scientific research articles could be too difficult to understand, and a more popular and dominant resource of information during the pandemic is social media [8]. Therefore, we create a simple COVID-19 news QA dataset using verified COVID-19 news collected from social media to study our problem. Moreover, in [9], the COVID QA models are trained end-to-end to achieve high performance for the specific COVID-19 domain using labeled data, whereas in this work, we focus on the *unsupervised* domain adaptation problem in COVID-19 QA, where the training labels of COVID data are often inaccessible due to the cost of annotations and dynamics of the COVID-19 disease.

Domain Adaptation. Domain adaptation methods are primarily explored in computer vision tasks [24], and only limited domain adaptation methods are developed for misinformation detection and question answering. In [25], language models are post-trained in a domain-distinguishing task to improve the models' domain adaptation ability for misinformation detection. In [14], [26], domain adversarial

training is implemented for misinformation task and question answering, so that the models are trained to learn domain-invariant features. Utilizing contrastive methods, [6] propose to quantify and reduce the domain discrepancy using explicit distance measures (e.g., maximum mean discrepancy) to bridge the gap between source domain and target domain. However, the unsupervised domain adaptation methods have not been systematically studied in the COVID-19 domain. In this work, inspired by the idea of adversarial examples [27] and domain mixup [13], we propose to establish a smoothed intermediate training domain by perturbing the latent representations of the input from both source domain and target domain towards each other with a domain discriminator and perform contrastive training on the smoothed domain to transfer knowledge from the source training domain to the target COVID-19 domain.

3 Problem Statement

3.1 Setup

Data: We define two data domain distributions, namely the source domain data distribution $\mathcal{P}$ and the target domain data distribution $\mathcal{Q}$. Note that the data formats for the two COVID-19 information services considered in this work are different from each other. Regarding the COVID-19 misinformation detection task, we formulate it as a binary text classification, where each data point $(\boldsymbol{x}, y)$ contains an input segment of COVID-19 claim or news ($\boldsymbol{x}$) and a label $y \in \{0, 1\}$ ($y = 1$ for true information and $y = 0$ for false information). As for the COVID-19 news question answering service, each QA sample is a 3-tuple that consists of a question $\boldsymbol{x}_q$, a context $\boldsymbol{x}_c$ and an answer span $\boldsymbol{y}$. To differentiate the notations of the data sampled from the source distribution $\mathcal{P}$ and the target distribution $\mathcal{Q}$, we further introduce two definitions of the domain data:

- ***Source domain data***: We use the subscript s to denote the source domain data. In particular, for the COVID-19 misinformation detection task, the source domain data form a source domain dataset $\boldsymbol{\mathcal{X}}_s = \{(\boldsymbol{x}_s, y_s)|(\boldsymbol{x}_s, y_s) \sim \mathcal{P}\}$, and for the COVID-19 news question answering task, the source domain data is $\boldsymbol{\mathcal{X}}_s = \{(\boldsymbol{x}_{s,q}, \boldsymbol{x}_{s,c}, \boldsymbol{y}_s)|(\boldsymbol{x}_{s,q}, \boldsymbol{x}_{s,c}, \boldsymbol{y}_s) \sim \mathcal{P}\}$.
- ***Target domain data***: Similarly, we use the subscript t to denote the target domain data. That is, we have target domain datasets $\boldsymbol{\mathcal{X}}_t = \{\boldsymbol{x}_t|\boldsymbol{x}_t \sim \mathcal{Q}\}$, and $\boldsymbol{\mathcal{X}}_t = \{(\boldsymbol{x}_{t,q}, \boldsymbol{x}_{t,c})|(\boldsymbol{x}_{t,q}, \boldsymbol{x}_{t,c}) \sim \mathcal{Q}\}$ for the two COVID-19 information tasks respectively. Note, as discussed in Section 1, we focus on unsupervised domain adaptation and treat our target domain data as *unlabeled*. That is, during training, the ground truth labels of target domain data y_t or $\boldsymbol{y}_t$ are not used for both tasks.

Moreover, for the sake of simplicity, if not explicitly mentioned, we use $\boldsymbol{x}$ to denote the general input regardless of its domain, $\boldsymbol{x}_s$ to denote the source domain input and $\boldsymbol{x}_t$ for target COVID domain input in both misinformation detection model and the QA model.

Models: For misinformation detection, the model f takes an input text $\boldsymbol{x}$ (a COVID-19 claim or a piece of news) to predict whether the information contained in $\boldsymbol{x}$ is valid or not. In contrast, the QA model f takes in a question text $\boldsymbol{x}_q$ and a context $\boldsymbol{x}_c$, and is trained to extract an answer span from the context $\boldsymbol{x}_c$. The QA model predicts which token (out of all tokens) is the start token and which another token is the end token. Therefore, **the QA task is not a binary classification task**, which is clearly different from the misinformation detection task. Moreover, due to difference of the two tasks, our proposed solutions (i.e., the optimization objective) are also different in misinformation detection and QA, which we will elaborate in the next section.

3.2 Problem Formulation

Using the labels of the training data, finding the optimal model corresponding to a specific task is to minimize the cross-entropy loss $\mathcal{L}_{ce}$ over the training data:

$$\mathcal{L}_{ce} = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{X}}\Big[l(f(\boldsymbol{x}), y)\Big]. \tag{1}$$

Our goal is to adapt a classifier f trained on the source domain data distribution $\mathcal{P}$ to the target domain data distribution $\mathcal{Q}$. For a given target domain input $\boldsymbol{x}_t$, a well-adapted model aims at making predictions as correctly as possible. Mathematically, we formulate the problem of adapting the misinformation detection model and the QA model as follows:

- ***Misinformation Detection***:

$$\max_f \mathbb{E}_{(\boldsymbol{x}_t,y_t)\sim\mathcal{X}_{test}}\Big[\Phi(f(\boldsymbol{x_t}), y_t)\Big] \tag{2}$$

- ***Question Answering***:

$$\max_f \mathbb{E}_{(\boldsymbol{x}_{t,q},\boldsymbol{x}_{t,c},y_t)\sim\mathcal{X}_{test}}\Big[\Phi(f(\boldsymbol{x}_{t,q}, \boldsymbol{x}_{t,c}), \boldsymbol{y}_t)\Big] \tag{3}$$

Note that in Equation 2 and Equation 3, we use $\mathcal{X}_{test}$ to denote the target domain test data and Φ represents the performance metrics (e.g., accuracy or F1 score). Here, we only use the ground-truth labels of test data (i.e., y_t and $\boldsymbol{y}_t$) for evaluation. It is non-trivial to solve Equation 2 and Equation 3 in our unsupervised setting where the ground-truth labels of the target domain data are not available.

4 Solution

4.1 Domain Discriminator and Domain Adversarial Mixup

Domain Discriminator. The first step of our framework is to train a domain discriminator f_D to classify whether the input data belongs to the source or target domain. For a better understanding of our model, we visualize the structure of our domain discriminator in Figure 3. Firstly, a COVID-19 model is split into a BERT Encoder $\boldsymbol{f}_e$ and the task classifier f_y. Sharing the same feature space with f_y, the domain discriminator f_D is added after the BERT encoder.

However, we note that the input space of the domain discriminator of the misinformation detection application is different from the input space of the domain discriminator of the question answering application. For instance, the task classifier of the misinformation detection model takes the token `[CLS]` representation from the BERT encoder as input and returns a binary logit. In comparison, for the QA task, the task classifier of the QA model takes the representation of all question tokens and all context tokens, and outputs

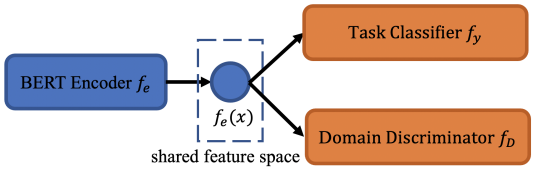


Figure 3: Split a COVID model into a BERT encoder f_e and the task classifier f_y. The binary domain discriminator f_D shares the same feature space with the task classifier, and also takes the same input as the task classifier.

two groups of logits, where the first group predicts the start position of the answer span and the second group predicts the end position. Therefore, unlike the domain discriminator in [14] that *always* takes the token `[CLS]` representation as input, the domain discriminator of our framework will take the same input of the task classifier, which is different for different tasks. Moreover, since we only consider two data domains in this paper (i.e., source data domain and target data domain), the domain discriminator f_D is also binary. Formally, the domain discriminators of the two tasks are defined as:

- **Misinformation Detection**:
$$\hat{y} = f_D(\boldsymbol{z}), \tag{4}$$
where $\boldsymbol{z}$ is the representation of token `[CLS]`.
- **Question Answering**:
$$\hat{y} = f_D(\boldsymbol{z}_{cq}), \tag{5}$$
where $\boldsymbol{z}_{cq}$ is the representation of question-context tokens.

Regarding the training of the domain discriminator, we highlight that although the ground truth labels of the target domain data are not used in our framework, we can still leverage the *domain labels* of the target domain data. In this work, the domain discriminator is a binary classifier, and we explicitly define the domain label y_D of the source domain data as $y_D = 0$ and the domain label of the target domain data as $y_D = 1$. Therefore, the training of the domain discriminator is formulated as:

- **Misinformation Detection**:
$$\min_{f_D} \mathbb{E}_{(\boldsymbol{x},y_D)\sim\mathcal{X}'}\Big[l(f_D(f_e(\boldsymbol{x})), y_D)\Big], \tag{6}$$
- **Question Answering**:
$$\min_{f_D} \mathbb{E}_{(\boldsymbol{x}_q,\boldsymbol{x}_c,y_D)\sim\mathcal{X}'}\Big[l(f_D(f_e(\boldsymbol{x}_q,\boldsymbol{x}_c)), y_D)\Big], \tag{7}$$

where $\mathcal{X}'$ represents the merged datasets of both source domain and target domain training data with domain labels.

Adversarial Domain Mixup. After the domain discriminator is trained, we show how to establish the adversarial domain mixup in the latent feature space of the model. Intuitively speaking, the poor generalization of the pre-trained model on target domain data is due to the large gap between the source domain and target domain. In return, bridging the gap between the source and target domains could contribute to the domain adaptation of the model. To achieve this goal, we propose to directly perturb the latent representations of the input data from both source domain and target domain towards the decision boundary of the domain discriminator as shown in Figure 2b. To this end, the perturbed representations (i.e., domain adversarial representations) from both domains could become closer to each other, indicating a reduced domain gap. Herein, the generated domain adversarial representations from both domains form a smoothed intermediate domain mixup in the latent feature space of the model.

Mathematically, the optimal perturbation $\boldsymbol{\delta}^*$ to perturb the latent representation $\boldsymbol{z}$ of a specific training sample $\boldsymbol{x}$ could be found by solving an optimization problem:

$$\begin{aligned}\mathcal{A}(f_e, f_D, \boldsymbol{x}, y_D, \epsilon) = &\max_{\boldsymbol{\delta}} \Big[l(f_D(\boldsymbol{z}+\boldsymbol{\delta}), y_D)\Big] \\ &s.t. \quad \|\boldsymbol{\delta}\| \le \epsilon, \quad \boldsymbol{z} = f_e(\boldsymbol{x}).\end{aligned} \tag{8}$$

Note that in the above equation, we introduce a hyper-parameter ϵ to bound the norm of the perturbation δ, so that the infinity solution could be avoided. Moreover, we highlight that δ is not a constant. Instead, for a specific training sample x, we generate its unique perturbation δ. For different training samples, the perturbations are also different. Eventually, after applying Equation 8 to all training samples in the merged training set $\mathcal{X}'$, we obtain the adversarial domain mixup $\mathcal{Z}'$:

$$\begin{aligned}\mathcal{Z}' &= \{\boldsymbol{z}'|\boldsymbol{z}' = \boldsymbol{z} + \mathcal{A}(f_e, f_D, \boldsymbol{x}, y_D, \epsilon), (\boldsymbol{x}, y_D) \in \mathcal{X}'\} \\ &:= \mathcal{Z}'_s \cup \mathcal{Z}'_t,\end{aligned} \tag{9}$$

where $\mathcal{Z}'_s$ are perturbed source features and $\mathcal{Z}'_t$ are perturbed target features. We use the projected gradient descent (PGD) to approximate the solution of Equation 8.

4.2 Contrastive Domain Adaptation

Next, we propose a contrastive adaptation loss over $\mathcal{Z}'$ to further adapt knowledge from the source data domain to the target data domain. By minimizing the proposed contrastive adaptation loss, the domain discrepancy between the source domain and the target domain will be reduced. Inspired by [6], our proposed contrastive adaptation loss is two-fold.

Firstly, we reduce the domain discrepancy among intra-class representations. That is, if input data of the representations is from the same class but from different data domains, we will reduce the discrepancy among these representations by minimizing an adaptation loss term. For instance, in the misinformation detection task, if a representation from the source data domain has a label of being true (or false) and a representation from the target data domain has a pseudo label of being true (or false), then these two representations are considered as intra-class representations and we reduce the domain discrepancy between them. However, in the question answering task, the label of input data is not true or false, but a span of answer. Therefore, in this case, we define the representation of all answer tokens as one class, and the representation of the combined question-context tokens as another class. We decrease the discrepancy among answer representations and among question-context representations, respectively. For instance, if a representation from the source data domain represents the span of answer and a representation from the target data domain also represents the span of pseudo-labeled answer, then these two representations are intra-class representations and we reduce the domain discrepancy between them.

The second level of our contrastive adaptation loss is defined for inter-class representations. As shown in Figure 2c, we enlarge the discrepancy of the representations from different classes. More specifically, in the misinformation detection task, the discrepancy between the representations of true information and false information will be enlarged, and in the QA task, the discrepancy between the answer representation and the question-context representation will be increased. As argued in [6], since the models are trained using labeled source domain data, they can effectively distinguish the representations of source domain data from different classes. However, in the target domain, without any adapted knowledge, the models tend to make mistakes in terms of recognizing the class of the representations. Therefore, maximizing the inter-class discrepancy could help the model to identify the class of the representations. In terms of COVID-19 misinformation, this helps the model to identify whether a representation of COVID-19 data is from the true class (i.e., true information) or the false class (i.e., false information). As for the COVID-19 news QA, the model learns to predict whether a representation represents the answer tokens or just question-context tokens.

In terms of computing our contrastive adaptation loss, we propose to measure the discrepancy among token classes using radial basis functions (RBF). In [28], RBF is proved to be an efficient tool to quantify uncertainty in deep neural networks. Recall pseudo labeling is used in our framework to classify the representations of target domain data into different classes, so that we can compute the intra-class loss and inter-class loss introduced above. Since our pseudo labeling process is designed to automatically filter out low-confident labels for the target domain data, using RBF to measure the discrepancy among token classes could efficiently improve the quality of the pseudo labels and ultimately contribute the domain adaptation training of the model. Formally, with the definition of the RBF kernel $k(z_1, z_2) = \exp[-\frac{\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|^2}{2\sigma^2}]$, we define the class-aware loss for the misinformation detection task and the question answering task as follows:

- **Misinformation Detection**:

$$\begin{aligned}
\mathcal{L}_{con}(\boldsymbol{\mathcal{Z}}') = & \\
& - \sum_{i=1}^{|\boldsymbol{\mathcal{Z}}'_s|} \sum_{j=1}^{|\boldsymbol{\mathcal{Z}}'_t|} \frac{\mathbb{1}(y_s^{(i)} = 0, \hat{y}_t^{(j)} = 0) k(\boldsymbol{z}_s^{(i)}, \boldsymbol{z}_t^{(j)})}{\sum_{l=1}^{|\boldsymbol{\mathcal{Z}}'_s|} \sum_{m=1}^{|\boldsymbol{\mathcal{Z}}'_t|} \mathbb{1}(y_s^{(l)} = 0, \hat{y}_t^{(m)} = 0)} \\
& - \sum_{i=1}^{|\boldsymbol{\mathcal{Z}}'_s|} \sum_{j=1}^{|\boldsymbol{\mathcal{Z}}'_t|} \frac{\mathbb{1}(y_s^{(i)} = 1, \hat{y}_t^{(j)} = 1) k(\boldsymbol{z}_s^{(i)}, \boldsymbol{z}_t^{(j)})}{\sum_{l=1}^{|\boldsymbol{\mathcal{Z}}'_s|} \sum_{m=1}^{|\boldsymbol{\mathcal{Z}}'_t|} \mathbb{1}(y_s^{(l)} = 1, \hat{y}_t^{(m)} = 1)} \\
& + \sum_{i=1}^{|\boldsymbol{\mathcal{Z}}'_s|} \sum_{j=1}^{|\boldsymbol{\mathcal{Z}}'_s|} \frac{\mathbb{1}(y_s^{(i)} = 1, y_s^{(j)} = 0) k(\boldsymbol{z}_s^{(i)}, \boldsymbol{z}_s^{(j)})}{\sum_{l=1}^{|\boldsymbol{\mathcal{Z}}'_s|} \sum_{m=1}^{|\boldsymbol{\mathcal{Z}}'_s|} \mathbb{1}(y_s^{(l)} = 1, y_s^{(m)} = 0)} \\
& + \sum_{i=1}^{|\boldsymbol{\mathcal{Z}}'_t|} \sum_{j=1}^{|\boldsymbol{\mathcal{Z}}'_t|} \frac{\mathbb{1}(\hat{y}_t^{(i)} = 1, \hat{y}_t^{(j)} = 0) k(\boldsymbol{z}_t^{(i)}, \boldsymbol{z}_t^{(j)})}{\sum_{l=1}^{|\boldsymbol{\mathcal{Z}}'_t|} \sum_{m=1}^{|\boldsymbol{\mathcal{Z}}'_t|} \mathbb{1}(\hat{y}_t^{(l)} = 1, \hat{y}_t^{(m)} = 0)},
\end{aligned} \tag{10}$$

where $\hat{y}_t$ is the pseudo label of the target domain samples and $\boldsymbol{z}$ denotes the representation of token `CLS`.

- **Question Answering**:

$$\begin{aligned}
\mathcal{L}_{con}(\boldsymbol{\mathcal{Z}}') = & -\frac{1}{|\boldsymbol{\mathcal{Z}}'|} \sum_{i=1}^{|\boldsymbol{\mathcal{Z}}'_s|} \sum_{j=1}^{|\boldsymbol{\mathcal{Z}}'_t|} k(\boldsymbol{z}_{s,a}^{(i)}, \hat{\boldsymbol{z}}_{t,a}^{(j)}) \\
& - \frac{1}{|\boldsymbol{\mathcal{Z}}'|} \sum_{i=1}^{|\boldsymbol{\mathcal{Z}}'_s|} \sum_{j=1}^{|\boldsymbol{\mathcal{Z}}'_t|} k(\boldsymbol{z}_{s,cq}^{(i)}, \hat{\boldsymbol{z}}_{t,cq}^{(j)}) \\
& + \frac{2}{|\boldsymbol{\mathcal{Z}}'_s|} \sum_{i=1}^{|\boldsymbol{\mathcal{Z}}'_s|} \sum_{j=1}^{|\boldsymbol{\mathcal{Z}}'_s|} k(\boldsymbol{z}_{s,a}^{(i)}, \boldsymbol{z}_{s,cq}^{(j)}) \\
& + \frac{2}{|\boldsymbol{\mathcal{Z}}'_t|} \sum_{i=1}^{|\boldsymbol{\mathcal{Z}}'_t|} \sum_{j=1}^{|\boldsymbol{\mathcal{Z}}'_t|} k(\hat{\boldsymbol{z}}_{t,a}^{(i)}, \hat{\boldsymbol{z}}_{t,cq}^{(j)}),
\end{aligned} \tag{11}$$

where $\hat{\boldsymbol{z}}_{t,a}$ and $\hat{\boldsymbol{z}}_{t,cq}$ are the mean vectors of representation of pseudo labeled answer and question-context tokens.

Overall Contrastive Adaptation Loss. Now, we merge the cross-entropy loss of the task classification problem and the above contrastive adaptation loss into a single optimization objective for the COVID model:

$$\mathcal{L}_{all} = \mathcal{L}_{ce}(\boldsymbol{\mathcal{X}}) + \lambda \mathcal{L}_{con}(\boldsymbol{\mathcal{Z}}'), \tag{12}$$

where $\mathcal{L}_{ce}$ is the cross-entropy loss over the training data with ground-truth label or the pseudo label, and λ is the hyperparameter to adjust the domain adatation strength. In our experiments, we sample mini-batches of source domain data and target domain to compute the overall loss instead of computing all combinations for $\mathcal{L}_{con}$. Moreover, we compute the RBF kernel with multiple bandwidths for $\mathcal{L}_{con}$, since multiple bandwidths of the RBF kernel encourage the model to learn a smoothed and generalized feature space [6]. Finally, we highlight that the intrinsic difference between the misinformation detection task and the QA task leads to the different designs of our solutions. Comparing Equation 10 and Equation 11 when computing the contrastive loss, the criterion used to determine intra- or inter-data points are drastically different in these two tasks.

4.3 Overall Framework

The overall framework is summarized in Algorithm 1. With the trained model $f = f_e \circ f_y$, we firstly compute the pseudo labels for the target domain data within a mini-batch (Line 7-Line 8). Line 9 - Line 16 show the process of establishing the adversarial domain mixup by performing PGD with f_D. Next, we compute our contrastive loss (Equation 12) over the generated adversarial domain mixup to update the model f (Line 17 - Line 18). Note that the quality of the adversarial domain mixup is highly related to the accuracy of the domain discriminator due to the fact that the adversarial domain mixup is established by the domain discriminator. Therefore, after we update the COVID-19 misinformation detection model or the COVID-19 news question answering model, the domain discriminator is also updated using training data with domain labels from source domain and target domain (Line 19 - Line 20).

Different from previous domain adaptation work in NLP tasks [14], [26], we leverage the power of adversarial examples to establish a smoothed intermediate domain mixup

Algorithm 1: Contrastive Adversarial Domain Mixup

Inputs Source data $\boldsymbol{\mathcal{X}}_s$, unlabeled target data $\boldsymbol{\mathcal{X}}_t$, pre-trained model $f = f_e \circ f_y$, pre-trained domain discriminator f_D ;
Hyperparameters: Number of iteration N, batch size B, confidence threshold τ, scaling factor λ adversarial radius ϵ, number of steps for PGD K and step size of PGD η;
Output: domain adapted model f ;
for *training iterations* **do**
  Load a mini-batch $\boldsymbol{M}_s$ from $\boldsymbol{\mathcal{X}}_s$;
  Load a mini-batch $\boldsymbol{M}_t$ from $\boldsymbol{\mathcal{X}}_t$;
  Compute pseudo labels for $\boldsymbol{M}_t$;
  Filter $\boldsymbol{M}_t$ with minimum confidence threshold τ;
  for $\boldsymbol{x}_i \in \boldsymbol{M}_s \cup \boldsymbol{M}_t$ **do**
    Compute $\boldsymbol{z}_i = f_e(\boldsymbol{x}_i)$;
    Initialize $\boldsymbol{z}'_i = \boldsymbol{z}_i$;
    for *k = 1, 2, ..., K* **do**
      $\mathcal{L}(\boldsymbol{z}'_i) = l(f_D(f_e(\boldsymbol{z}'_i)), y_D)$;
      $\boldsymbol{z}'_i = \prod_{\mathbb{B}(\boldsymbol{z}_i,\epsilon)}(\boldsymbol{z}'_i + \eta \mathsf{sign} \nabla_{\boldsymbol{z}'_i} \mathcal{L}(\boldsymbol{z}'_i))$, where $\prod$ is the projection operator ;
    end
  end
  Compute Equation 12 over the adversarial domain mixup and the training data $\boldsymbol{M}_s \cup \boldsymbol{M}_t$;
  Perform backpropagation and update f;
  Compute Equation 6 or Equation 7 over $\boldsymbol{M}_s \cup \boldsymbol{M}_t$;
  Perform backpropagation and update f_D;
end

instead of neutralizing the domain information of the representation via domain discriminator. Moreover, unlike [13], where linear interpolation is used to construct the domain mixup, our domain mixup is generated by the domain discriminator f_D and therefore is non-linear and learnable. Finally, our two-fold contrastive adaptation loss (i.e., inter-class loss term and the intra-class loss term) further transfers knowledge from the source domain to the target unseen domain. By minimizing the contrastive adaptation loss, our method minimizes the discrepancy between two domains via RBF distance and also encourages the model to learn class-separating features.

5 Evaluation

5.1 Experimental Design

Datasets: For the misinformation detection task, we use three source misinformation datasets (GossipCop [29], LIAR [15] and PHEME [30]) released *before* the COVID outbreak and two COVID misinformation datasets (Constraint [11] and ANTiVax [10]) collected *after* the outbreak as target datasets. As for the QA task, SearchQA [31], TriviaQA [32] and NewsQA [33] are selected as source datasets. Moreover, since there is no COVID-19 news QA dataset from current literature, we create our own COVID-19 news QA dataset. In particular, we pick the true news from CoAID [8] dataset as the context, and use a pre-trained T5 model [34] to generate question-answer pairs. In fact, CoAID is also a COVID-19 misinformation dataset, but we noticed that this dataset is extremely imbalanced (more than 90% are true information), which could cause a problem of labeling shifting [35] in our task. Therefore, we only use it to generate COVID-19 news question-answer pairs instead of using it for misinformation detection. The T5 model we used is pre-trained on SQuAD [36], so SQuAD is not used as source domain dataset in the QA experiments. In Table 1, we show two generated COVID-19 news question-answer pairs. The contexts are true COVID-19 news collected from social media platforms [8]. For the datasets that do not contain validation and test set, we split the data into training, validation, and test sets with the ratio of 7:1:2 as in [17], [26].

Table 1: Generated COVID-19 news question-answer pairs.

Context: Reported coronavirus disease COVID-19 cases likely represent only fraction of all sars-cov-2 virus that causes COVID-19 infections. This may be because unknown proportion of people that have mild or no symptoms do not seek medical care or do not get tested when they sought medical care. **Question**: What type of virus causes COVID-19 infections? **Answer**: Sars-cov-2 virus.
Context: COVID-Net provides national data on laboratory confirmed hospitalizations. On April 17, data were added on COVID-19-associated hospitalizations by age with race and ethnicity information and on selected underlying medical conditions, such as asthma cardiovascular disease. **Question**: What provides national data on laboratory confirmed hospitalizations? **Answer**: COVID-Net.

Table 2: Supervised training results.

Dataset	BA ↑	Acc. ↑	F1 ↑
CoAID	0.8892	0.9720	0.9846
Constraint	0.9350	0.9327	0.9323
ANTiVax	0.9303	0.9191	0.9291
CoAID News	-	0.5877	0.6442

Model: Following [26], the commonly used RoBERTa [2] was selected as the misinformation detection model, whereas for the QA task we use the BERT-QA model [1], [6], [14]. Moreover, we also provide supervised training results of the models to show the domain gap between the source domain and target domain, which could be regarded as performance upper bounds of all domain adaptation models including our CADM+.

Baselines: For COVID-19 misinformation detection and COVID-19 news question answering, we directly test the source domain pre-trained models on target domain test set and use the test results without any domain adaptation as the naive baseline. Moreover, further state-of-the-art domain adaptation methods are selected for comparison in both tasks. The baselines methods are selected based on the code availability and whether the algorithms' performance could be reproduced. In particular, for the misinformation detection task, we select 1) DAAT [25], where the misinformation detection model is post-trained to improve the domain-adversarial adaptation, 2) EADA [26], where energy-based domain adversarial training is performed using autoencoder, and 3) LDM, the linear domain mixup technique presented in [13]. In comparison, for the QA task, we select MRQA [14] and CAQA [6] as baseline algorithms for comparison. However, both methods are modified for a fair comparison: MRQA is modified to be unsupervised, and the question generation (QG) of CAQA is removed. Finally, in

both tasks, all compared methods including ours are trained using exactly the same labeled source domain datasets and the unlabeled target domain datasets for fair comparison. As for evaluation, the adapted models are tested on the target (COVID) datasets only.

Evaluation Metrics: Regarding the evaluation metrics, we focus on accuracy (or exact match for QA) and F1 score. However, we argue that in terms of COVID-19 misinformation detection, it is important to correctly identify both true information and false information even when the distribution of the dataset is imbalanced. Therefore, we also use balanced accuracy (BA) to evaluate the models' performance in the misinformation detection task. Formally, balanced accuracy (BA) is defined as the average value of sensitivity and specificity:

$$\mathrm{BA} = \frac{1}{2}(\mathrm{TPR} + \mathrm{TNR}) = \frac{1}{2}(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}), \quad (13)$$

where TPR represents sensitivity and TNR represents specificity. TP, TN are true positive and true negative, FP and FN refer to false positive and false negative, respectively.

5.2 Evaluation Results

COVID-19 Misinformation Detection. The first set of experiments are designed to evaluate the efficacy of our proposed CADM+ framework for the COVID-19 misinformation detection task. From Table 3, we observe that our method consistently outperforms all baseline algorithms in terms of adapting the models to the unseen COVID-19 domain. For instance, consider the setting where the model is adapted from source domain PHEME to target domain Constraint. This setting is the most challenging one: the model can only make random guess on the target Constraint dataset (BA=0.4889) without any domain adaptation. Under this setting, EADA failed to adapt the model by achieving a balanced accuracy of 0.4944. As for the other baseline DAAT, the adapted model only performs slightly better, namely BA=0.5227. In contrast, the performance of the model trained using our framework could be significantly improved (e.g., BA=0.6430). In addition to BA, similar trends could be observed on other metrics for this challenging setting as well. Regarding other adaptation settings with different source and target domain combinations, our adapted models could still achieve a better performance on all metrics. For instance, when adapting from LIAR to ANTiVAX, compared to non-adapted model, the increase of the model's performance on three metrics using our framework is 30.84%, 45.22% and 83.85%, which is significantly larger than the best baseline method EADA with 13.59%, 34.74% and 79.96%.

Moreover, we also observe that our CADM+ is more consistent than the other two baseline methods. For instance, regardless of the source domain dataset and target domain dataset, our method could consistently adapt the misinformation detection model to the target domain. However, on some source and target domain combinations, the baseline methods successfully adapt the model but on other combinations the baseline methods simply fail. For instance, compared to the results without any domain adaptation, the gap between source domain GossipCop and target domain Constraint could be efficiently reduced using our method, whereas the domain gap is increased when EADA is deployed. In terms of the concrete numerical results, the adapted model trained with our framework could achieve an increased balanced accuracy of 0.7787 from 0.5638 while EADA reduces the model's performance on balanced accuracy from 0.5638 to 0.5210.

COVID-19 News Question Answering. The second set of the experiments is conducted on the COVID-19 News QA task (Table 4). Since QA is not a binary classification problem (the number of overall classes is the length of the context), we only report exact match (E.M.) and F1 scores. We observe that our framework could better adapt the QA model from all three source datasets to the target COVID-19 domain than other two baselines. For instance, compared to the non-adapted model, the model's E.M. is increased by 22.16% when adapted from SearchQA to CoAID News using our framework, which is significantly greater than 2.05% of MRQA and 9.23% of CAQA. As for the other two source datasets, our method also outperforms the baseline methods on both metrics.

Finally, it is expected that the adapted models perform worse in both misinformation detection and question answering. This is because there is no overlap between the source datasets and the target COVID dataset. Moreover, as shown in Table 5, since the source domain is drastically different from the target COVID domain, and there is no ground-truth for target domain, adapting the models could be quite challenging.

5.3 Qualitative Results

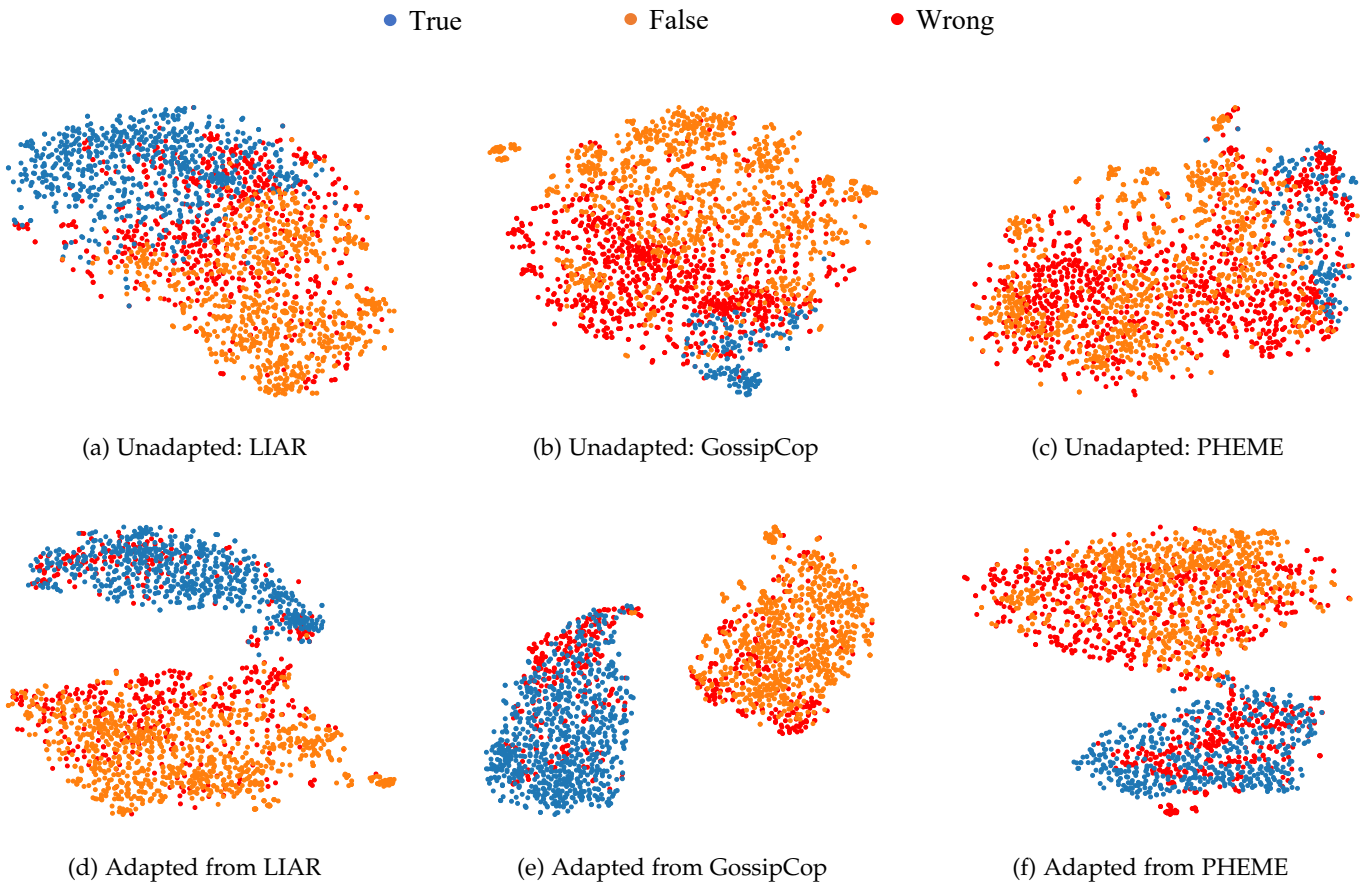
Next, we visualize the representations of `[CLS]` token in Figure 4 and Figure 5 for COVID misinformation detection application. From both figures (first rows), it is observed that the representations of true information and misinformation are clustered together before adaptation. That is, before adaptation, there exists no obvious gap between the true information and misinformation in the features space of the models. In other words, the classifier fails to correctly classify a COVID claim being true or false. In comparison, as shown in the second row of Figure 4 and Figure 5, with CADM+, the models learn to separate the features of true information and misinformation. However, we also observe some mistakes made by CADM+ for data samples near the decision boundaries of Figure 4 and Figure 5. We discuss the limitation of our scheme in the "Limitation and Future Work" section.

6 Limitations and Future Work

This paper focuses on the unsupervised domain adaptation problem for COVID-19 information services on social media data. In our experiments, we show the effectiveness of our proposed CADM+ framework in terms of adapting the misinformation detection model and the QA model from the source training domain to the unseen target COVID domain. Moreover, we highlight that there exists great potential to further augment our proposed CADM+ framework based on the evolving dynamics of the COVID pandemic. Firstly, with increasing medical research on COVID virus, more

Table 3: Results of domain adaptation for COVID-19 misinformation detection.

Target Dataset	Source Dataset	LIAR			GossipCop			PHEME		
	Metric	BA ↑	Acc. ↑	F1 ↑	BA ↑	Acc. ↑	F1 ↑	BA ↑	Acc. ↑	F1 ↑
Constraint	No Adaptation	0.7231	0.7322	0.7822	0.5638	0.5832	0.7110	0.4889	0.5047	0.6360
	DAAT	0.7606	0.7626	0.7795	0.7178	0.7276	0.7806	0.5227	0.5411	0.6763
	EADA	0.7776	0.7794	0.7950	0.5210	0.5430	0.6944	0.4944	0.4969	0.6391
	LDM	0.7995	0.8075	0.8406	0.5108	0.5336	0.6918	0.5029	0.5262	0.6884
	CADM+ (Ours)	**0.8288**	**0.8304**	**0.8420**	**0.7787**	**0.7780**	**0.7828**	**0.6430**	**0.6547**	**0.7301**
ANTiVax	No Adaptation	0.5444	0.4929	0.4162	0.5695	0.6501	0.7673	0.5294	0.6196	0.7531
	DAAT	0.6228	0.5778	0.5393	0.6692	0.7161	0.7918	0.5895	0.6498	0.7518
	EADA	0.6184	0.6642	0.7490	0.5509	0.6434	0.7709	0.5411	0.6328	0.7632
	LDM	0.5009	0.3973	0.0152	0.5000	0.6050	0.7539	0.5036	0.6076	0.7549
	CADM+ (Ours)	**0.7123**	**0.7158**	**0.7652**	**0.7522**	**0.7701**	**0.8152**	**0.6752**	**0.7323**	**0.8107**



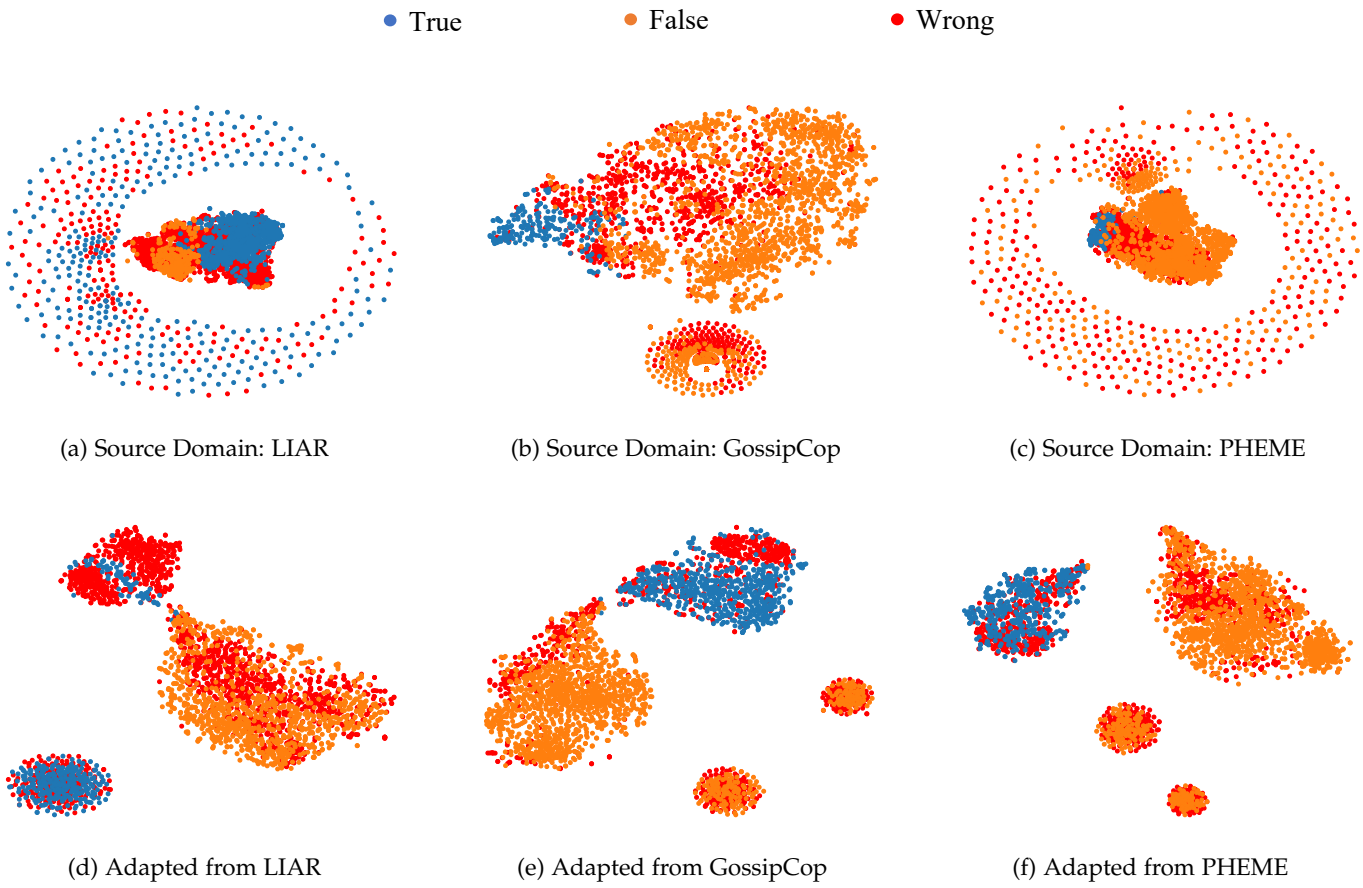
(a) Unadapted: LIAR (b) Unadapted: GossipCop (c) Unadapted: PHEME

(d) Adapted from LIAR (e) Adapted from GossipCop (f) Adapted from PHEME

Figure 4: **Constraint Adaptation**: adapted from source datasets (i.e., LIAR, GossipCop, PHEME).

labeled COVID data become available. Although enabling a pure supervised training of the COVID models could still be too expensive, it is feasible to enable semi-supervised training of the model using limited labeled COVID data. In terms of CADM+, one possible research direction is to add a new semi-supervision module to the current CADM+ framework to further improve its domain adaptation performance.

We further discuss some limitations of CADM+. Firstly, CADM+ mainly performs the adaptation w.r.t. the input data (latent representations). However, in addition to the domain shift of input data, another key factor that leads to a large domain discrepancy is the shift of labels, which is not considered in CADM+. With more labeled COVID training data, it is also possible to further enhance the CADM+ by adding a new label correction module, where the main challenge is the unknown distribution of target domain labels. As such, CADM+ could reduce the domain discrepancy with respect to the joint distribution of inputs and labels. Secondly, we observe another limitation of CADM+: degraded performance on the samples near the decision boundary. According to Figure 4 and Figure 5, it is observed that in general, the adapted models are still prone to making wrong predictions on the samples that are close to the decision boundary. For instance, in Figure 4d and Figure 5d, when the samples are close to the other class, then error occurs with a larger probability. To address such hard samples, uncertainty-based importance re-weighting techniques [37] could be adopted to improve the performance



(a) Source Domain: LIAR (b) Source Domain: GossipCop (c) Source Domain: PHEME

(d) Adapted from LIAR (e) Adapted from GossipCop (f) Adapted from PHEME

Figure 5: **ANTiVax Adaptation**: adapted from source datasets (i.e., LIAR, GossipCop, PHEME).

Table 4: Results of domain adaptation for COVID-19 news question answering.

Source Dataset	Target Dataset	CoAID News	
	Metric	E.M. ↑	F1 ↑
SearchQA	No Adaptation	0.3412	0.4239
	MRQA (supervised)	0.5995	0.6546
	MRQA (unsupervised)	0.3482	0.4291
	CAQA	0.3727	0.4771
	CADM+ (Ours)	**0.4168**	**0.4873**
TriviaQA	No Adaptation	0.3642	0.4504
	MRQA (supervised)	0.5922	0.6492
	MRQA (unsupervised)	0.3798	0.4623
	CAQA	0.3808	0.4605
	CADM+ (Ours)	**0.3913**	**0.4778**
NewsQA	No Adaptation	0.4724	0.5899
	MRQA (supervised)	0.6132	0.6641
	MRQA (unsupervised)	0.4899	0.5992
	CAQA	0.5095	0.6047
	CADM+ (Ours)	**0.5160**	**0.6106**

of CADM+.

Table 5: Detailed comparison between source domain and target domain.

Source Dataset	Domain	Training Label	Example
LIAR	Politics	Available	Newly elected republican senators sign pledge to eliminate food stamp program in 2015.
GossipCop	Gossip	Available	Cindy Crawford's daughter Kaia Gerber wears a wig after dining with Harry Styles.
PHEME	Rumors	Available	Charlie Hebdo became well known for publishing the Muhammed cartoons two years ago.
Target Dataset	**Domain**	**Training Label**	**Example**
Constraint	COVID	Unavailable	Heart conditions like myocarditis are associated with some cases of COVID19.
ANTiVax	COVID	Unavailable	The vaccine can cause infertility.

7 Discussion

The main challenge of our domain adaptation setting lies in the emerging nature of the target domain (e.g., COVID-19). The novel virus (i.e., COVID-19) itself is a new biological and medical concept, and there is no authoritative and ground-truth knowledge of the training samples in this emerging domain at the early stage of the pandemic. Moreover, unlike traditional domain adaptation settings in misinformation detection, where the source domain or target domain are daily-life related (such as political news, celebrity gossips), COVID-19 is relatively new and adapting

the knowledge from existing data to a new target domain could be more challenging. For instance, celebrities might get involved in political events, indicating potential overlaps between political news domain and the celebrity gossip domain. However, due to the novelty and emergency nature of COVID-19 as the target domain, the domain discrepancy between the source and target domain could be larger than normal, making the domain adaptation task in such settings more challenging and interesting.

8 Conclusion

In this paper, we present a novel unsupervised domain adaptation framework for COVID-19 information services on social media data. Our unsupervised framework is motivated by the fact that the ground-truth labels of the COVID-19 data are not always available but the need for high-quality information services is always persistent and urgent. In addition to COVID-19, our method has the potential to provide efficient solutions to many other information services (e.g., sentiment analysis, hate speech detection) on social media platforms, when the training labels of target domain data are missing. Our unsupervised domain adaptation is realized via a novel adversarial domain mixup and contrastive learning. Extensive experimental results on two real-world COVID-19 information services suggest that our method could successfully and efficiently adapt the models from source domain to the target domain without requiring labels of COVID-19 data for both tasks.

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

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[3] S. Malla and P. Alphonse, "Covid-19 outbreak: an ensemble pre-trained deep learning model for detecting informative tweets," *Applied Soft Computing*, vol. 107, p. 107495, 2021.

[4] A. Chiorrini, C. Diamantini, A. Mircoli, and D. Potena, "Emotion and sentiment analysis of tweets using bert." in *International Conference on Extending Database Technology/International Conference on Database Theory Workshops*, vol. 3, 2021.

[5] H. M. Zahera, I. A. Elgendy, R. Jalota, and M. A. Sherif, "Fine-tuned BERT model for multi-label tweets classification," in *Proceedings of the Twenty-Eighth Text REtrieval Conference*, 2019.

[6] Z. Yue, B. Kratzwald, and S. Feuerriegel, "Contrastive domain adaptation for question answering using limited text corpora," *arXiv preprint arXiv:2108.13854*, 2021.

[7] J. Roozenbeek, C. R. Schneider, S. Dryhurst, J. Kerr, A. L. Freeman, G. Recchia, A. M. Van Der Bles, and S. Van Der Linden, "Susceptibility to misinformation about covid-19 around the world," *Royal Society open science*, vol. 7, no. 10, p. 201199, 2020.

[8] L. Cui and D. Lee, "Coaid: Covid-19 healthcare misinformation dataset," *arXiv:2006.00885*, 2020.

[9] D. Oniani and Y. Wang, "A qualitative evaluation of language models on automatic question-answering for covid-19," in *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2020, pp. 1–9.

[10] K. Hayawi, S. Shahriar, M. A. Serhani, I. Taleb, and S. S. Mathew, "Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection," *Public health*, vol. 203, pp. 23–30, 2022.

[11] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, "Fighting an infodemic: Covid-19 fake news dataset," in *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Springer, 2021, pp. 21–29.

[12] H. Zeng, Z. Yue, Z. Kou, L. Shang, Y. Zhang, and D. Wang, "Unsupervised domain adaptation for covid-19 information service with contrastive adversarial domain mixup," *arXiv preprint arXiv:2210.03250*, 2022.

[13] M. Xu, J. Zhang, B. Ni, T. Li, C. Wang, Q. Tian, and W. Zhang, "Adversarial domain adaptation with domain mixup," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6502–6509.

[14] S. Lee, D. Kim, and J. Park, "Domain-agnostic question-answering with adversarial training," *arXiv preprint arXiv:1910.09342*, 2019.

[15] W. Y. Wang, "" liar, liar pants on fire": A new benchmark dataset for fake news detection," *arXiv preprint arXiv:1705.00648*, 2017.

[16] S. D. Das, A. Basak, and S. Dutta, "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*. Springer, 2021, pp. 164–176.

[17] Z. Kou, L. Shang, Y. Zhang, and D. Wang, "Hc-covid: A hierarchical crowdsource knowledge graph approach to explainable covid-19 misinformation detection," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. GROUP, pp. 1–25, 2022.

[18] Z. Kou, L. Shang, Y. Zhang, C. Youn, and D. Wang, "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)*. IEEE, 2021, pp. 140–147.

[19] S. C. R. Gangireddy, D. P, C. Long, and T. Chakraborty, "Unsupervised fake news detection: A graph-based approach," in *Proceedings of the 31st ACM conference on hypertext and social media*, 2020, pp. 75–83.

[20] S. Yang, K. Shu, S. Wang, R. Gu, F. Wu, and H. Liu, "Unsupervised fake news detection on social media: A generative approach," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 5644–5651.

[21] A. Dhiman and D. Toshniwal, "An unsupervised misinformation detection framework to analyze the users using covid-19 twitter data," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 679–688.

[22] D. Li, H. Guo, Z. Wang, and Z. Zheng, "Unsupervised fake news detection based on autoencoder," *IEEE access*, vol. 9, pp. 29 356–29 365, 2021.

[23] J. Gaglani, Y. Gandhi, S. Gogate, and A. Halbe, "Unsupervised whatsapp fake news detection using semantic search," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2020, pp. 285–289.

[24] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4893–4902.

[25] C. Du, H. Sun, J. Wang, Q. Qi, and J. Liao, "Adversarial and domain-aware bert for cross-domain sentiment analysis," in *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, 2020, pp. 4019–4028.

[26] H. Zou, J. Yang, and X. Wu, "Unsupervised energy-based adversarial domain adaptation for cross-domain text classification," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 1208–1218.

[27] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[28] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *International conference on machine learning*. PMLR, 2020, pp. 9690–9700.

[29] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big data*, vol. 8, no. 3, pp. 171–188, 2020.

[30] C. Buntain and J. Golbeck, "Automatically identifying fake news in popular twitter threads," in *2017 IEEE International Conference on Smart Cloud (SmartCloud)*. IEEE, 2017, pp. 208–215.

[31] M. Dunn, L. Sagun, M. Higgins, V. U. Guney, V. Cirik, and K. Cho, "Searchqa: A new q&a dataset augmented with context from a search engine," *arXiv preprint arXiv:1704.05179*, 2017.

[32] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," *arXiv preprint arXiv:1705.03551*, 2017.

[33] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman, "Newsqa: A machine comprehension dataset," *arXiv preprint arXiv:1611.09830*, 2016.

[34] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer." *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

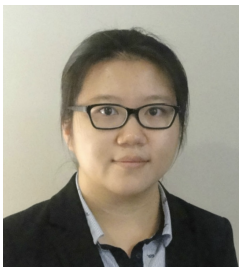
[35] K. Azizzadenesheli, A. Liu, F. Yang, and A. Anandkumar, "Regularized learning for domain adaptation under label shifts," *arXiv preprint arXiv:1903.09734*, 2019.

[36] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

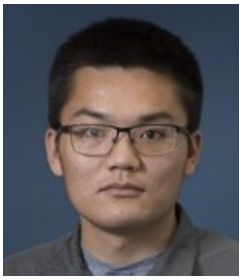
[37] Z. Han, Z. Liang, F. Yang, L. Liu, L. Li, Y. Bian, P. Zhao, B. Wu, C. Zhang, and J. Yao, "Umix: Improving importance weighting for subpopulation shift via uncertainty-aware mixup," *Advances in Neural Information Processing Systems*, vol. 35, pp. 37 704–37 718, 2022.

Huimin Zeng received the B.S. degree from Tongji University, Shanghai, China, in 2018, and the M.S. degree from Technical University of Munich, Munich, Germany, in 2021. He is currently pursuing the Ph.D. degree with the School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL, USA. His research interest lies in the intersection of social sensing and adversarial machine learning.

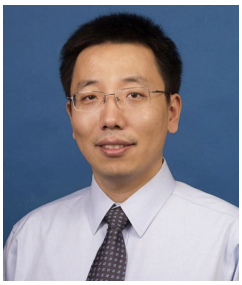
Zhenrui Yue received the B.S. degree and the M.S. degree from Technical University of Munich, Munich, Germany, in 2021. He is currently pursuing the Ph.D. degree with the School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL, USA. His research interest lies in language understanding and recommender system.



Lanyu Shang received the B.S. degree in applied mathematics from the University of California at Los Angeles, Los Angeles, CA, USA, in 2014, and the M.S. degree in data science from New York University, New York, NY, USA, in 2017. She is currently pursuing the Ph.D. degree with the School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL, USA. Her research interest primarily lies in online misinformation detection using social media data.



Yang Zhang received the Ph.D degree in computer science from the University of Notre Dame, Notre Dame, IN, USA. He received the B.S. degree from Wuhan University, Wuhan, China, in 2013, and the M.S. degree from Indiana University at Bloomington, Bloomington, IN, USA, in 2017. He is currently a post-doc researcher with the School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL, USA. His research interests include social sensing, deep learning, and human-centered artificial intelligence.



Dong Wang received the Ph.D. degree in computer science from the University of Illinois Urbana-Champaign (UIUC), Champaign, IL, USA, in 2012. He is currently an Associate Professor with the School of Information Sciences, UIUC. His research interests lie in the area of social sensing, computing and intelligence, human-centric AI, and smart city applications. Dr. Wang received the Best Paper Awards of AMC/IEEE International Conference on Advances in Social Networks Analysis and Mining in 2022 and the IEEE Real-Time and Embedded Technology and Applications Symposium in 2010, the Army Research Office Young Investigator Program Award in 2017, the Google Faculty Research Award in 2018, and the NSF CAREER Award in 2019. He is a member of the ACM and senior member of IEEE.