A Knowledge-driven Domain Adaptive Approach to Early Misinformation Detection in an Emergent Health Domain on Social Media

Lanyu Shang, Yang Zhang, Zhenrui Yue, YeonJung Choi, Huimin Zeng, Dong Wang
School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL, USA
{lshang3, yzhangnd, zhenrui3, yc55, huiminz3, dwang24}@illinois.edu

***Abstract*—This paper focuses on an important problem of early misinformation detection in an emergent health domain on social media. Current misinformation detection solutions often suffer from the lack of resources (e.g., labeled datasets, sufficient medical knowledge) in the emerging health domain to accurately identify online misinformation at an early stage. To address such a limitation, we develop a knowledge-driven domain adaptive approach that explores a good set of annotated data and reliable knowledge facts in a source domain (e.g., COVID-19) to learn the domain-invariant features that can be adapted to detect misinformation in the emergent target domain with little ground truth labels (e.g., Monkeypox). Two critical challenges exist in developing our solution: i) how to leverage the noisy knowledge facts in the source domain to obtain the medical knowledge related to the target domain? ii) How to adapt the domain discrepancy between the source and target domains to accurately assess the truthfulness of the social media posts in the target domain? To address the above challenges, we develop KAdapt, a knowledge-driven domain adaptive early misinformation detection framework that explicitly extracts relevant knowledge facts from the source domain and jointly learns the domain-invariant representation of the social media posts and their relevant knowledge facts to accurately identify misleading posts in the target domain. Evaluation results on five real-world datasets demonstrate that KAdapt significantly outperforms state-of-the-art baselines in terms of accurately detecting misleading Monkeypox posts on social media.**

I. Introduction

The rampant spread of misinformation on social media has become a serious societal issue and raised significant public concerns [1]. According to Pew Research Center, around 70% of U.S. adults describe online misinformation as a major threat to the country.[1] Social media misinformation encompasses a variety of domains, including healthcare, climate change, and politics [2]. Among them, health-related misinformation is a particularly important domain of misinformation that not only threatens the well-being of the general public but also reduces the trustworthiness of public authorities [3]. However, existing health misinformation detection solutions often suffer from undesirable detection performance on an emergent health domain (e.g., the recent outbreak of Monkeypox) due to the domain discrepancy from the original domain in which the detection model was trained and the lack of medical knowledge of the emergent health domain [2]. In particular, an emergent health domain refers to an emerging health event/topic that requires immediate attention (e.g., disease outbreaks, bioterrorist attacks) [4]. In this paper, we study the problem of domain adaptive health misinformation detection that aims at accurately identifying misleading social media posts in an emergent health domain.

Several recent efforts have been made to address the health misinformation detection problem on social media [5], [6]. Examples of these solutions include the content-based methods (e.g., semantic features), context-based models (e.g., propagation patterns), and hybrid schemes (e.g., content-comment similarities) [2]. Current solutions usually require a non-trivial amount of well-annotated data to supervise the training of misinformation detection models. However, it is often impractical to obtain the timely labels of social media posts in the emergent domain due to the labor-intensive and time-consuming process of label annotation [7]. In addition, there also exist a few knowledge-driven health misinformation detection solutions that leverage the knowledge facts extracted from medical documents (e.g., medical literature, fact-checking articles) in the given health domain [5], [8]. However, such solutions cannot be directly applied to detect misinformation in emergent health domains where only a very limited amount of medical documents are available. Therefore, the detection of misinformation in an emergent health domain remains a challenging problem yet to be addressed.

Motivated by the above limitations, we develop a knowledge-driven domain adaptive approach to address the problem of misinformation detection in an emergent health domain. Our goal is to leverage the vast amount of data from a high-resource source domain (i.e., the health domain with sufficient annotated data and medical knowledge) to detect misinformation in a low-resource target domain (i.e., the emergent health domain with little annotated data and medical knowledge). We show an example of our problem in Figure 1. In particular, we explore the large amount of annotated misinformation data (Figure 1(a)) and the reliable knowledge facts (Figure 1(b)) in the source domain (e.g., COVID-19) to learn the domain-invariant features that can be adapted to detect misinformation in the target domain (e.g., Monkeypox in Figure 1(c)). However, two critical challenges exist in developing our solution.

[1] https://pewrsr.ch/3tkE8nv

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 © 2022 IEEE

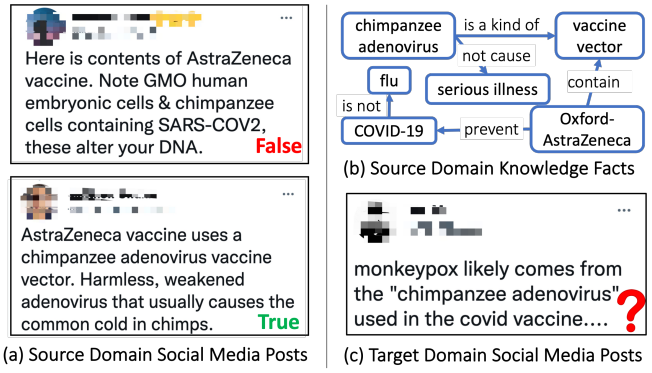


Figure 1: Knowledge-driven Domain Adaptive Health Misinformation Detection

Limited Knowledge in the Target Domain. The first challenge lies in the limited amount of health-related knowledge information in the target domain. The amount of available knowledge resources in the emergent target domain is often very limited at the early stage of an emergent disease (e.g., the outbreak of Monkeypox outside Africa). A possible solution to address the limited knowledge problem is to acquire reliable medical knowledge information from medical professionals. However, such a manual knowledge collection process is both time-consuming and labor-intensive [7]. In contrast, the medical knowledge from a relevant high-resource health domain often contains valuable information for detecting the misinformation in the target domain. For example, the medical knowledge fact "chimpanzee adenovirus" $\xrightarrow{\text{not cause}}$ "serious illness" (Figure 1(b)) from the source health domain (e.g., COVID-19) can be of great help for detecting misinformation in the post "Monkeypox likely comes from the 'chimpanzee adenovirus' used in covid vaccine" (Figure 1(c)). However, we also observe that the source domain often contains many noisy knowledge facts (e.g., "COVID-19" $\xrightarrow{\text{is not}}$ "flu" in Figure 1(b)) that are irrelevant to the posts in the target domain. Therefore, it remains a challenge to leverage the noisy knowledge facts in the source domain to facilitate the early detection of misinformation in the target domain.

Discrepancy Between Source and Target Domains. The second challenge lies in the discrepancy between the source and target domains in healthcare. Social media posts from different domains often present their unique data distributions (e.g., word frequency and vocabulary) and language patterns (e.g., semantic and syntax characteristics) [9]. For example, flu-related content appears more common in the social media discussion about COVID-19 due to the similar symptoms both diseases share. In contrast, homophobic claims are often observed among the Monkeypox-related social media posts because the initial cases during the recent Monkeypox outbreak are found in the LGBTQ+ community.[2] Therefore, a misinformation detection model that is trained in the source domain cannot be directly applied to identify misinformation in the target domain. There also exists a few domain adaptation based text classifiers to address the domain discrepancy issue [10], [11]. However, such solutions mainly focus on the textual features but ignore the correctness of health information embedded in the post, making them insufficient to identify the misleading information across different health domains [5]. Therefore, it remains a challenge to address the discrepancy between the source and target domains to accurately assess the truthfulness of social media posts in the target domain.

To address the above challenges, we develop KAdapt, a knowledge-driven domain adaptive early misinformation detection framework that explores the labeled posts and medical knowledge in a source domain to accurately identify health-related misinformation in an emergent target domain on social media. In particular, to address the first challenge, we design a post-driven knowledge extraction module to explicitly extract relevant knowledge facts from the noisy knowledge graph in the source domain. To address the second challenge, we develop a dual-adaptive representation learning model that aims at learning the domain-invariant representation of the social media posts and their relevant knowledge facts to jointly detect the misinformation in the emergent target domain. To the best of our knowledge, KAdapt is the first knowledge-driven domain adaptive solution for healthcare misinformation detection on social media. We evaluate KAdapt on a case study of domain adaptation between COVID-19 and Monkeypox on five different social media datasets. Evaluation results demonstrate that KAdapt achieves significant performance gains compared to state-of-the-art baselines by accurately detecting misleading Monkeypox posts on social media.

II. Related Work

A. Health Misinformation on Social Media

Health misinformation has become a severe issue on social media and has gained much attention in recent years [1]. Many efforts have been made in response to the prevalence of health misinformation on social media [5], [12], [13]. For example, Ghenai *et al.* developed a user-centric classification model that explores the linguistic and sentiment features for identifying social media users who are prone to propagate health misinformation [14]. Zhao *et al.* investigated the user behaviors in online health communities and proposed an Elaboration Likelihood Model (ELM) based framework to detect misleading health-related social media posts [15]. Weinzierl *et al.* designed a graph-based bootstrapping scheme that leverages a set of known health misconceptions to retrieve and detect health misinformation on social media [16]. However, existing solutions often assume there is an adequate amount of data that can be leveraged to train an effective health misinformation detection model. Thus, they are insufficient to detect misinformation in emerging health domains where a very limited amount of training data is available. In this paper, we develop a domain-aware health misinformation detection framework that investigates the domain discrepancy between the source and target health domains to accurately detect misleading social media posts in an emergent health domain.

[2]https://www.technologyreview.com/2022/06/17/1054408/homophobic-misinformation-spread-monkeypox-social-media/

B. Domain Adaptation

Domain adaptation is a transfer learning technique that aims at minimizing the impact of domain shift between the training data (i.e., source domain) and the testing data (i.e., target domain) [17]. Domain adaptation has attracted widespread interest in deep learning communities, such as computer vision and natural language processing [9]. More recently, a few domain adaptive misinformation detection solutions have been developed to address the domain discrepancy issue in misinformation detection [10], [18], [19]. For example, Li *et al.* developed a weakly supervised domain adaptation solution that leverages linguistic-based weak labels of news articles in the target domain for fake news detection [18]. Zhang *et al.* proposed a BERT-based domain adaptation neural network solution that maps the text representation of the social media posts in the source and target domains to the same feature space for classifying misleading posts [10]. While these solutions can learn the domain-invariant features from the input posts in the source and target domains to detect misinformation, they largely ignore the medical knowledge facts associated with the posts, which are fundamental for identifying misinformation on an emergent health domain [5]. In contrast, KAdapt designs a post-driven knowledge extraction strategy to explore relevant medical knowledge facts for detecting misinformation in an emergent health domain.

C. Health Knowledge Graph

There is a growing trend of utilizing knowledge graph to model the ontological relations between entities in the unstructured natural language data (e.g., medical literature, electronic health records) in health domains [20]. For example, Gong *et al.* designed a knowledge graph embedding model that explores the patients' medical records and drug information for safe medicine recommendations [21]. Groza *et al.* proposed a syntax-based medical misconception detection solution that leverages the ontological relation of medical concepts to identify medical misconceptions in social media posts [22]. Cui *et al.* incorporated a medical knowledge graph constructed from medical publications to detect health misinformation about known diseases (e.g., cancer, diabetes) [5]. However, the above solutions are inadequate to address the early misinformation detection problem in an emergent health domain due to the domain discrepancy between the source and target domains. To address such a limitation, we develop a domain-aware knowledge discriminator to effectively learn domain-invariant knowledge representation for the detection of emerging misinformation in the target health domain.

III. Problem Statement

In this section, we formally define the problem of adaptive health misinformation detection on social media. In particular, the adaptive health misinformation detection framework aims at adapting a misinformation classifier trained on the data from a *source domain* (e.g., COVID-19) to detect misinformation in the *target domain* (e.g., Monkeypox). Next, we define a few key concepts that will be used in the problem formulation.

Definition 1. Post (p): We define a post as a piece of text on social media (e.g., tweet) that is relevant to the domain of interest (e.g., COVID-19, Monkeypox). In particular, we define the set of M *source posts* (i.e., posts from the source domain) as $\mathcal{P}^s = \{p_1^s, p_2^s, \cdots, p_M^s\}$, and the set of N *target posts* (i.e., posts from the target domain) as $\mathcal{P}^t = \{p_1^t, p_2^t, \cdots, p_N^t\}$.

Definition 2. Source Article (c): A source article c is an article whose content is related to the source domain (i.e., COVID-19). In particular, we consider two types of articles in our study: 1) *news articles* from credible news publishers (e.g., Centers for Disease Control and Prevention (CDC), Mayo Clinic) and 2) *fact-checking articles* from mainstream fact-checking websites (e.g., FactCheck.org, politifact.org). In particular, we denote the set of L source articles as $\mathcal{C} = \{c_1, c_2, \cdots, c_L\}$.

Definition 3. Domain Label (z): We define a domain label $z \in \{0, 1\}$ of each post as the domain which the post belongs to. In particular, we define $z = 0$ for all the posts from the source domain and $z = 1$ for all the posts from the target domain. We denote $\mathcal{Z}$ as the domain labels for all the posts in the source and target domains.

Definition 4. Ground-truth Label (y): We consider the binary ground-truth label $y \in \{0, 1\}$ for each post. In particular, a post p is *misleading* (i.e., $y = 0$) if it contains false or unverified information which may contribute to both imminent and long-term harm to public health and safety [1]. Otherwise, the post is considered as *non-misleading* ($y = 1$). We also denote $\mathcal{Y}^s$ and $\mathcal{Y}^t$ as the ground-truth labels for the posts in the source and target domains, respectively.

Given the above definitions, we formulate the adaptive health misinformation detection problem as an adaptive binary classification problem that adapts a classifier trained on the source posts and source articles to classify each target post into two categories (i.e., misleading or non-misleading). For each $p_n^t \in \mathcal{P}^t$, our objective is

$$\arg\max_{\hat{y}_n} \Pr(\hat{y}_n = y_n | \mathcal{P}^s, \mathcal{P}^t, \mathcal{Y}^s, \mathcal{C}, \mathcal{Z}), \ \forall 1 \leq n \leq N \quad (1)$$

where y_n and $\hat{y}_n$ are the ground-truth and estimated label of the target post p_n^t, respectively.

IV. Solution

In this section, we present the KAdapt framework to address the early misinformation detection problem in an emergent health domain. The overall objective of KAdapt is to explore the widely available labeled posts and medical knowledge in a high-resource source domain to effectively detect misleading posts in an emergent low-resource target domain. Figure 2 shows an overview of the KAdapt framework. In particular, KAdapt consists of two main modules: 1) a *Post-driven Knowledge Extraction (PKE)* module that constructs a source knowledge graph and effectively extracts useful knowledge facts that are relevant to the content of the social media

posts, and 2) a *Dual-adaptive Representation Learning (DPL)* module that jointly learns the domain-invariant representations of social media posts and the associated knowledge facts from the source knowledge graph in PKE to accurately identify the misleading posts in the emergent health domain. We elaborate on each module in detail below.

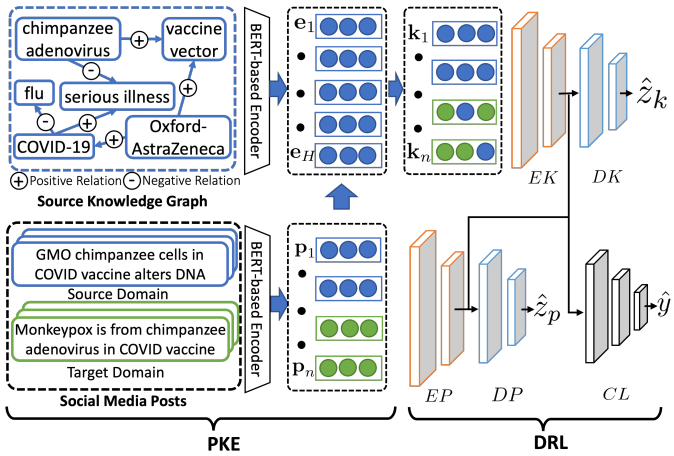


Figure 2: Overview of the KAdapt Framework

A. Post-driven Knowledge Extraction (PKE)

The post-driven knowledge extraction module aims at effectively extracting the medical knowledge information that is highly relevant to the content of the social media posts from both the source and target domains. Existing content-based domain adaptation misinformation detection solutions mainly focus on the post content but ignore the medical knowledge that is often crucial for identifying misleading information, especially in an emergent health domain [18]. To address such a limitation, KAdapt explicitly explores the medical knowledge from the source articles (i.e., articles in the source domain) to obtain the medical knowledge facts that can be leveraged to identify misinformation in the target domain. In particular, we observe that nouns or noun phrases in the source articles are often associated with important knowledge facts (i.e., meaningful entities and their relations) for detecting the misleading posts. For example, the knowledge fact "chimpanzee adenovirus" $\xrightarrow{\ominus}$ "serious illness" in Figure 2 is helpful for identifying the misleading post "Monkeypox is from chimpanzee adenovirus in COVID vaccine." Thus, we mainly focus on the noun or noun phrase entities and their relations in PKE to obtain the relevant knowledge facts. We first define the entity and relation below.

Definition 5. Entity (e): An entity e is defined as a noun (e.g., "vaccine") or noun phrase (e.g., "chimpanzee adenovirus") extracted from the source articles. Formally, we denote a set of H entities as $\mathcal{E} = \{e_1, e_2, \cdots, e_H\}$.

Definition 6. Relation (r): A relation r is defined as the semantic relation between a pair of relevant entities from the source articles. In particular, we consider two types of relations $\mathcal{R} = \{r^+, r^-\}$ in KAdapt. $r^+ \in \mathcal{R}$ refers to the *positive relation* between a pair of entities (e.g., the positive relation of "contain" between "Oxford-AstraZeneca" and "vaccine vector" in Figure 2). $r^- \in \mathcal{R}$ refers to the *negative relation* between a pair of entities (e.g., the negative relation of "not cause" between "chimpanzee adenovirus" and "serious illness" in Figure 2).

With the medical knowledge-related entities and relations defined above, we then construct the source knowledge graph to explore the relational information between different entities in the source articles. We formally define the source knowledge graph as follows.

Definition 7. Source Knowledge Graph ($\mathcal{G}$): The source knowledge graph is defined as directed graph $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, A\}$, where $\mathcal{E}$ and $\mathcal{R}$ represent the set of entities and their relations extracted from the source articles, respectively. $A = \{A^+, A^-\}$ represent the adjacent matrices that denote the relation between each pair of entities in $\mathcal{E}$. In particular, $A^+_{i,j} \in \{0, 1\}$ denotes whether there is a positive relation between entity e_i and e_j. Similarly, $A^-_{i,j} \in \{0, 1\}$ denotes whether there is a negative relation between entity e_i and e_j.

Given the source knowledge graph $\mathcal{G}$ defined above, our next goal is to learn the high-level representation (e.g., semantics, contexts) of the entities in $\mathcal{G}$ to aggregate the medical knowledge facts based on their relations in the source knowledge graph. In particular, we design a BERT-based entity encoder to encode the entities with different numbers of words to the latent vector representation of fixed length. Let $e_i = [u_1, u_2, \cdots, u_{n_i}]$ be an entity in $\mathcal{E}$, where u_k for $1 \leq k \leq n_i$ is the k^{th} word in entity e_i. We first obtain the pre-trained BERT word embedding [23] of each word u_k, denoted as $\mathbf{u}_k \in \mathbb{R}^d$, where d is the dimension of the word embedding. We then perform the mean-pooling and max-pooling on the extracted word embeddings and concatenate the pooled embeddings to obtain the entity embedding that aggregates the semantic features of each entity $e_i \in \mathcal{E}$. We also define the entity embedding of each entity $e_i \in \mathcal{E}$ as $\mathbf{e}_i \in \mathbb{R}^{2d}$, and the entity embedding matrix $E \in \mathbb{R}^{H \times 2d}$ as the matrix that contains the entity embeddings of all entities in $\mathcal{E}$.

Using the encoded entity embeddings, we develop a multi-relational graph convolutional network (MR-GCN) [24] to effectively learn the latent representation of each entity and aggregate the knowledge facts from the connected entities in the source knowledge graph $\mathcal{G}$. In particular, we define the multi-relational aggregation strategy in MR-GCN as:

$$\hat{\mathbf{e}}_i = \sigma\Big(\sum_{\mathbf{e}_j \in \mathcal{N}_i^+} \frac{1}{\omega_i^+} W^+ \mathbf{e}_j + \sum_{\mathbf{e}_j \in \mathcal{N}_i^-} \frac{1}{\omega_i^-} W^- \mathbf{e}_j + W \mathbf{e}_i \Big) \quad (2)$$

where $\hat{\mathbf{e}}_i$ is the learned representation of $e_i \in \mathcal{E}$ from MR-GCN. $\sigma(\cdot)$ is the non-linear ReLU activation function. $\mathbf{e}_i$ and $\mathbf{e}_j$ are the entity embeddings of the i^{th} and j^{th} entity in $\mathcal{E}$, respectively. $\mathcal{N}_i^+$ and $\mathcal{N}_i^-$ refer to the set of neighborhood entities of entity $e_i \in \mathcal{E}$ under the relation r^+ and r^-, respectively. ω_i^+ and ω_i^- are the normalization factors. W^+,

W^-, and W are the learnable weight parameters. The intuition of such an aggregation strategy is to learn the contextual information of each entity from its connected neighbors in the source knowledge graph.

While MR-GCN captures the knowledge facts in the source knowledge graph $\mathcal{G}$ based on the relations between different entities, it remains a challenge to identify the critical knowledge facts (i.e., entities and their relations) from $\mathcal{G}$ that can be applied to detect misinformation in the target domain. The reason is that the source knowledge graph $\mathcal{G}$ is constructed based on the articles from the source domain which often contains many knowledge facts that are irrelevant to the target domain. For example, the knowledge fact "COVID-19" $\xrightarrow{\oplus}$ "serious illness" in Figure 2 is not relevant to the post discussing "Monkeypox". To address this problem, we further design a post-driven knowledge extraction strategy to effectively obtain the important knowledge facts from $\mathcal{G}$ where the entities are relevant to the diversified content of social media posts from the source and target domains. For example, more knowledge facts related to the entity "chimpanzee adenovirus" can be retrieved based on the post "Monkeypox is from chimpanzee adenovirus in COVID vaccine." To this end, we explicitly measure the relevance between a given post and each entity in $\mathcal{E}$ to extract the key medical knowledge facts in $\mathcal{G}$. In particular, we encode each post $p \in \{\mathcal{P}^s, \mathcal{P}^t\}$ with the BERT-based encoder, denoted as $\mathbf{p} \in \mathbb{R}^{1\times 2d}$. Let $E \in \mathbb{R}^{H\times 2d}$ be the entity embedding matrix of all entities in $\mathcal{E}$, we update the adjacent matrices A^+ and A^- as follows.

$$\hat{A^+} = \text{softmax}\left((E\mathbf{p}^\top W_a^+) \odot A^+\right) \tag{3}$$

$$\hat{A^-} = \text{softmax}\left((E\mathbf{p}^\top W_a^-) \odot A^-\right) \tag{4}$$

where W_a^+ and W_a^- are the learnable weight parameters.

Finally, we obtain the post-driven medical knowledge fact representations for each post $p \in \{\mathcal{P}^s, \mathcal{P}^t\}$ under the positive and negative relations. In particular, we aggregate all the entity representations based on the relevance score measured in $\hat{A^+}$ and $\hat{A^-}$, followed by an average operation. We denote the post-driven knowledge fact representations under the positive and negative relations as $\mathbf{k}^+$ and $\mathbf{k}^-$, and concatenate them to obtain the final medical knowledge fact feature, denoted as $\mathbf{k}$.

B. Dual-adaptive Representation Learning (DRL)

Given the encoded posts and extracted knowledge fact representations from the PKE module, our next goal is to learn the domain-invariant representations of both the social media posts and the medical knowledge facts, which can accurately classify the misleading posts in the target domain without any supervision in the target domain. Current domain adaptation misinformation detection solutions mainly focus on the distribution difference of textual content in the post from the source and target domains but ignore the domain discrepancy of the associated knowledge facts from different domains which are the key information to examine the truthfulness of social media posts in healthcare domains [5]. To address such a limitation, we develop a dual-adaptive neural network that aims at capturing the critical information from the input posts and medical knowledge facts while minimizing divergence between the source and target domains. In particular, we first design two encoder networks with two dense layers each to learn the essential semantic features from the input posts and the medical knowledge fact representations. Formally, the encoded latent representations of the posts and medical knowledge facts are defined as $\mathbf{v}_p$ and $\mathbf{v}_k$, respectively:

$$\mathbf{v}_p = EP(\mathbf{p}) \text{ and } \mathbf{v}_k = EK(\mathbf{k}) \tag{5}$$

where EP and EK are the encoder networks for the posts and medical knowledge facts, respectively. $\mathbf{p}$ and $\mathbf{k}$ are the embeddings of the post and the medical knowledge fact, respectively.

In addition, we also design two discriminator networks with two dense layers each to accurately predict the domain label that indicate whether the encoded input post or the medical knowledge fact is from the source domain or the target domain. Formally, the estimated domain labels $\hat{z}_p$ and $\hat{z}_k$ of the encoded post $\mathbf{v}_p$ and medical knowledge fact $\mathbf{v}_k$ are defined as:

$$\hat{z}_p = DP(\mathbf{v}_p) \text{ and } \hat{z}_k = DK(\mathbf{v}_k) \tag{6}$$

where DP and DK are the discriminator networks for the encoded post $\mathbf{v}_p$ and medical knowledge fact $\mathbf{v}_k$, respectively.

With the encoder networks and discriminator networks, our next goal is to effectively learn the domain-invariant features that can capture the representative posts and medical knowledge fact features that are shared between the source and target domains. To this end, we adopt the adversarial loss function that aims at regulating the encoder networks to learn the non-discriminative features where the domain cannot be discriminated by the discriminator networks. Formally, let $\hat{z}_p$ and $\hat{z}_k$ be the domain label of the encoded post and its source knowledge features estimated by DP and DK, respectively. The cross-entropy-based adversarial loss function for DP and DK are defined as:

$$\mathcal{L}_P = \sum_{i=0}^{M+N} -z_i \log(\hat{z}_{p,i})_1 - (1-z_i)\log(1-(\hat{z}_{p,i})_0) \tag{7}$$

$$\mathcal{L}_K = \sum_{i=0}^{M+N} -z_i \log(\hat{z}_{k,i})_1 - (1-z_i)\log(1-(\hat{z}_{k,i})_0) \tag{8}$$

where z_i is the true domain label of the post p_i.

Finally, we develop a classification network to accurately examine the truthfulness of an input post by leveraging the encoded latent representations of the post $\mathbf{v}_p$ and the medical knowledge fact $\mathbf{v}_k$ from the encoder networks. In particular, we define a classification network CL as a stacked feed-forward neural network with a set of dense layers to predict the truthfulness of an input post. Formally, the classification network CL is defined as:

$$\hat{y}_i = CL([\mathbf{v}_{p,i}; \mathbf{v}_{k,i}]) \tag{9}$$

We optimize the classification network with the binary cross-entropy loss defined as:

$$\mathcal{L}_C = -\sum_{i=1}^{N}(1-y_i)\log(1-(\hat{y}_i)) - y_i\log(\hat{y}_i) \quad (10)$$

where y_i is the ground-truth label of the source post $p_i \in \mathcal{P}^s$. Recall that the adversarial loss functions $\mathcal{L}_P$ and $\mathcal{L}_K$ in Equations 7 and 8 are designed to ensure that the encoder networks can capture the domain-invariant features of the posts and medical knowledge facts from the source and target domains. $\mathcal{L}_C$ ensures the classification network can accurately identify a misleading social media post in the target domain by leveraging the features learned from the encoder networks under the supervision in the source domain.

Our overall learning objective is to jointly optimize encoder networks (EP, EK), discriminator networks (DP, DK), and classification network (CL) by minimizing the classification loss and maximizing the adversarial loss as follows.

$$\mathcal{L} = \mathcal{L}_C - \lambda\mathcal{L}_P - \beta\mathcal{L}_K \quad (11)$$

where λ and β are the hyperparameters that control the trade-off between the classification loss and adversarial loss. We adopt the Adaptive Moment Estimation optimizer [25] to solve the optimization problem in Equation 11 and accurately detect misinformation in different domains.

V. Evaluation

In this section, we evaluate the performance of the proposed KAdapt framework in detecting emerging health misinformation on social media. In particular, we consider COVID-19 and Monkeypox as the source domain and target domain in our study, respectively. Evaluation results show that KAdapt achieves substantial performance gains compared to state-of-the-art domain adaptation misinformation detection solutions.

A. Data

1) Source Articles: We consider two types of articles in the source domain (i.e., COVID-19) as our source articles: news articles and fact-checking articles. In particular, the news articles are collected from credible health news publishers (e.g., CDC, Mayo Clinic), and the fact-checking articles from the popular fact-checking websites (e.g., FactCheck.org, Politifact). We collect 259 articles in the source domain in total to construct the source knowledge graph in our work.

2) Source Posts: To collect the social media posts in the source domain, we adopt four public COVID-19 misinformation datasets, including Constraint [26], COVIDRumor [27], MMCoVaR [28], and ANTiVax [29]. We adopt the original ground-truth labels from each dataset and remove any duplicated posts. We summarize the source post datasets in Table I.

3) Target Posts: To evaluate the misinformation detection performance on the target posts, we collect a set of posts in the target domain (i.e., Monkeypox) from Twitter. Specifically, the target posts are collected by keyword search (e.g., "monkeypox", "monkey pox") using the official Twitter API. We obtain a total of 9165 posts from Twitter. We randomly select a subset of 500 posts as the test set to evaluate the performance of early misinformation detection. The remaining posts are used for the unsupervised adversarial training in KAdapt. In particular, we invite three independent professionals to manually annotate the truthfulness of each post in the test set and take the majority votes as the ground-truth labels. Finally, we obtain an annotated test set with 168 (33.6%) and 332 (66.4%) misleading and non-misleading posts, respectively.

Dataset	Number of Posts	Misleading	Non-misleading
Constraint	10,700	5,600	5,100
COVIDRumor	5,505	3,661	1,844
MMCoVaR	2,791	1,315	1,476
ANTiVax	12,326	4,156	8,170

Table I: Summary of Source Post Datasets

B. Baselines and Experiment Setup

1) Baselines:

- **BDANN** [10]: BDANN is a BERT-based domain adaptation neural network solution for multimodal fake news detection. In particular, we exclude the visual feature extractor in BDANN and leverage the BERT-based textual features of the source and target posts for detecting misleading posts.
- **MDA-WS** [18]: MDA-WS is a weakly supervised domain adaptation based fake news detection framework that leverages labeled source domain news articles and the word frequency based weak labels of target domain news articles to detect fake news in the target domain.
- **EANN** [30]: EANN is an event adversarial networks framework that learns transferable features from source news events for fake news detection on emerging news events. Specifically, we consider COVID-19 and Monkeypox as the source and emerging news events, respectively.
- **DETERRENT** [5]: DETERRENT is a graph attention neural network solution that utilizes the relational knowledge information in a biomedical knowledge base to detect misleading healthcare news.

2) Experiment Setup: To ensure a fair comparison, we keep the source and target posts to all compared methods the same in our evaluation. For DETERRENT, we use the same source articles as KAdapt to construct the source knowledge graph. In our experiments, we use all the source posts and unlabeled target posts for the unsupervised training of the encoder and discriminator networks. Additionally, we use the labeled source posts for the supervised training of the classification networks. The dimensions of the entity embeddings and post embeddings are 768. We set the total number of epochs as 40 with a batch size of 32. We adopt an initial learning rate of 0.0001. We run the experiments on Ubuntu 20.04 with four NVIDIA A40. We strictly follow the configurations of all baselines as documented in the original papers and carefully tune the hyperparameters to obtain the best results.

Method	Constraint				COVIDRumor				MMCoVaR				ANTiVax			
	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1
BDANN	0.564	0.434	0.592	0.501	0.646	0.627	0.781	0.697	0.588	0.683	0.707	0.695	0.629	0.647	0.676	0.661
MDA-WS	0.609	0.664	0.623	0.642	0.621	0.617	0.732	0.669	0.617	0.669	0.693	0.681	0.631	0.672	0.613	0.641
EANN	0.597	0.584	0.673	0.625	0.628	0.626	0.786	0.697	0.586	0.644	0.706	0.674	0.637	0.684	0.728	0.705
DETERRENT	0.628	0.663	0.635	0.649	0.654	0.662	0.825	0.735	0.636	0.679	0.758	0.716	0.642	0.690	0.711	0.701
KAdapt	**0.687**	**0.674**	**0.733**	**0.702**	**0.672**	**0.681**	**0.938**	**0.789**	**0.649**	**0.703**	**0.819**	**0.757**	**0.669**	**0.733**	**0.786**	**0.759**

Table II: Misinformation Detection Performance on Target Posts

Method	Constraint				COVIDRumor				MMCoVaR				ANTiVax			
	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1
KAdapt	**0.687**	**0.674**	**0.733**	**0.702**	**0.672**	**0.681**	**0.938**	**0.789**	**0.649**	**0.703**	**0.819**	**0.757**	**0.669**	**0.733**	**0.786**	**0.759**
KAdapt\K	0.592	0.589	0.669	0.626	0.631	0.603	0.736	0.663	0.611	0.624	0.698	0.659	0.639	0.673	0.718	0.695
KAdapt\P	0.621	0.657	0.672	0.664	0.652	0.658	0.746	0.699	0.622	0.665	0.738	0.701	0.641	0.678	0.704	0.691
KAdapt\D	0.619	0.639	0.677	0.657	0.648	0.652	0.737	0.692	0.623	0.659	0.721	0.689	0.647	0.688	0.721	0.704

Table III: Ablation Study Results

C. Detection Performance

In the first set of experiments, we evaluate the domain adaptation performance of all compared schemes on detecting misleading posts in the target domain (i.e., Monkeypox). We adopt the following evaluation metrics that are commonly used for the evaluation of classification performance: *Accuracy*, *Precision*, *Recall*, and *F1 Score*. The evaluation results are summarized in Table II. We observe that KAdapt consistently outperforms all the baselines on all source post datasets in terms of all evaluation metrics. For example, we observe that KAdapt outperforms the best-performing baseline (i.e., DETERRENT) on the COVIDRumor dataset by 5.4% in terms of the F1 score. We attribute such a performance improvement to the dual-adaptive representation learning design in KAdapt that learns not only the domain-invariant features from the social media posts in different domains but also the medical knowledge fact representations that can be applied to enhance the performance of misinformation detection in the target domain. In addition, the performance gains over the knowledge-independent domain adaptation baselines (e.g., BDANN, MDA-WS, EANN) also suggest the effectiveness of leveraging medical knowledge from the high-resource domain to complement the lack of medical knowledge during the early detection of misinformation in an emergent health domain.

D. Ablation Study

In the second set of experiments, we investigate the contribution of each key component in the KAdapt framework by conducting an ablation study. In particular, we consider the following three variants of KAdapt in our experiment.

- **KAdapt\K**: it removes the source knowledge graph and only learns domain-invariant features from the posts for misinformation detection.
- **KAdapt\P**: it excludes the post-driven knowledge extraction module and applies a mean-pooling layer to extract the knowledge features from the source knowledge graph.
- **KAdapt\D**: it removes the adversarial loss from the overall objective of KAdapt.

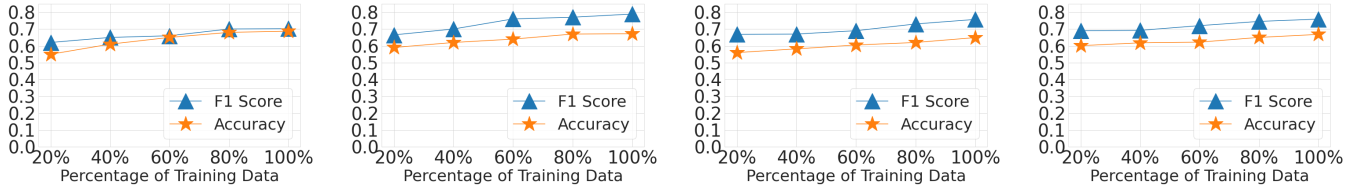
We summarize the evaluation results of the ablation study in Table III. We observe that KAdapt achieves its best performance when incorporating all key components in the framework. In particular, we note that the incorporation of the domain-invariant knowledge fact features from the source knowledge graph contributes most significantly to the overall performance of KAdapt, which can be attributed to the explicit exploration of useful medical knowledge information in KAdapt for the comprehensive assessment of the truthfulness of health-related social media posts.

E. Robustness Study

In the third set of experiments, we study the robustness of KAdapt against the amount of source posts in the training set. In particular, we vary the number of source posts from 20% to 100% of each training set. Then we evaluate the classification performance of KAdpat in terms of the accuracy and F1 score. We show the evaluation results in Figure 3. We observe that the overall performance of KAdapt improves as the number of training posts increases and gradually plateaus after the number of source posts reaches 60% or 80% of each dataset.

VI. Conclusion

In this paper, we study the early misinformation detection problem in an emergent health domain on social media. We present KAdapt, a knowledge-driven domain adaptive solution that explores the widely available annotated posts and medical knowledge in a high-resource source domain to accurately detect misinformation in the emergent target domain. We conduct a case study of the domain adaptation between COVID-19 and Monkeypox. Evaluation results demonstrate that KAdapt achieves substantial performance gains compared to state-of-the-art baselines in accurately detecting misleading social media posts in the target domain.



(a) Constraint (b) COVIDRumor (c) MMCoVaR (d) ANTiVax

Figure 3: Robustness Study - Number of Source Posts

ACKNOWLEDGEMENT

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] V. Suarez-Lledo, J. Alvarez-Galvez *et al.*, "Prevalence of health misinformation on social media: systematic review," *Journal of medical Internet research*, vol. 23, no. 1, p. e17187, 2021.

[2] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–40, 2020.

[3] M. B. Leonard, D. M. Pursley, L. A. Robinson, S. H. Abman, and J. M. Davis, "The importance of trustworthiness: lessons from the covid-19 pandemic," *Pediatric research*, vol. 91, no. 3, pp. 482–485, 2022.

[4] R. Haffajee, W. E. Parmet, and M. M. Mello, "What is a public health "emergency"?" *New England Journal of Medicine*, vol. 371, no. 11, pp. 986–988, 2014.

[5] L. Cui, H. Seo, M. Tabar, F. Ma, S. Wang, and D. Lee, "Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

[6] L. Shang, Z. Kou, Y. Zhang, and D. Wang, "A multimodal misinformation detector for covid-19 short videos on tiktok," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 899–908.

[7] Z. Kou, L. Shang, Y. Zhang, Z. Yue, H. Zeng, and D. Wang, "Crowd, expert & ai: A human-ai interactive approach towards natural language explanation based covid-19 misinformation detection," in *IJCAI*, 2022.

[8] 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.

[9] Z. Yue, H. Zeng, Z. Kou, L. Shang, and D. Wang, "Domain adaptation for question answering via question classification," *arXiv preprint arXiv:2209.04998*, 2022.

[10] T. Zhang, D. Wang, H. Chen, Z. Zeng, W. Guo, C. Miao, and L. Cui, "Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection," in *2020 international joint conference on neural networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[11] A. Rietzler, S. Stabinger, P. Opitz, and S. Engl, "Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification," *arXiv preprint arXiv:1908.11860*, 2019.

[12] 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.

[13] L. Shang, Z. Kou, Y. Zhang, and D. Wang, "A duo-generative approach to explainable multimodal covid-19 misinformation detection," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3623–3631.

[14] A. Ghenai and Y. Mejova, "Fake cures: user-centric modeling of health misinformation in social media," *Proceedings of the ACM on human-computer interaction*, vol. 2, no. CSCW, pp. 1–20, 2018.

[15] Y. Zhao, J. Da, and J. Yan, "Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches," *Information Processing & Management*, vol. 58, no. 1, p. 102390, 2021.

[16] M. A. Weinzierl and S. M. Harabagiu, "Automatic detection of covid-19 vaccine misinformation with graph link prediction," *Journal of biomedical informatics*, vol. 124, p. 103955, 2021.

[17] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 5, pp. 1–46, 2020.

[18] Y. Li, K. Lee, N. Kordzadeh, B. Faber, C. Fiddes, E. Chen, and K. Shu, "Multi-source domain adaptation with weak supervision for early fake news detection," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 668–676.

[19] Z. Yue, H. Zeng, Z. Kou, L. Shang, and D. Wang, "Contrastive domain adaptation for early misinformation detection: A case study on covid-19," *arXiv preprint arXiv:2208.09578*, 2022.

[20] L. Shang, Z. Kou, Y. Zhang, J. Chen, and D. Wang, "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)*. IEEE, 2022, pp. 1–10.

[21] F. Gong, M. Wang, H. Wang, S. Wang, and M. Liu, "Smr: medical knowledge graph embedding for safe medicine recommendation," *Big Data Research*, vol. 23, p. 100174, 2021.

[22] A. Groza, "Detecting fake news for the new coronavirus by reasoning on the covid-19 ontology," *arXiv preprint arXiv:2004.12330*, 2020.

[23] 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.

[24] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. v. d. Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *European semantic web conference*. Springer, 2018, pp. 593–607.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] 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.

[27] M. Cheng, S. Wang, X. Yan, T. Yang, W. Wang, Z. Huang, X. Xiao, S. Nazarian, and P. Bogdan, "A covid-19 rumor dataset," *Frontiers in Psychology*, vol. 12, 2021.

[28] M. Chen, X. Chu, and K. Subbalakshmi, "Mmcovar: multimodal covid-19 vaccine focused data repository for fake news detection and a baseline architecture for classification," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2021, pp. 31–38.

[29] 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.

[30] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "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*, 2018, pp. 849–857.