informs
http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 00, No. 0, Xxxx 0000, pp. 000–000
ISSN 0025-1909, EISSN 1526-5501

Submitted to *Management Science*

Early Detection of Misinformation for Infodemic Management: A Domain Adaptation Approach

Minjia Mao
Lerner College of Business and Economics, University of Delaware, Newark, DE, USA

Xiaohang Zhao*
School of Information Management & Engineering, Shanghai University of Finance and Economics, Shanghai, China

Xiao Fang*
Lerner College of Business and Economics, University of Delaware, Newark, DE, USA

*Corresponding author

Abstract. An infodemic refers to an enormous amount of true information and misinformation disseminated during a disease outbreak. Detecting misinformation at the early stage of an infodemic is key to reduce its harm to public health. An early stage infodemic is characterized by a large volume of unlabeled information concerning a disease. As a result, conventional misinformation detection methods are not suitable for this misinformation detection task because they rely on labeled information in the infodemic domain to train their models. To address this limitation, state-of-the-art methods learn their models using labeled information in other domains to detect misinformation in the infodemic domain. The efficacy of these methods depends on their ability to mitigate both covariate shift (i.e., differences in feature distributions) and concept shift (i.e., differences in labeling patterns) between the infodemic domain and the domains from which they leverage labeled information. However, these methods focus on mitigating covariate shift but overlook concept shift, rendering them less effective for the task. In response, we theoretically show the necessity of tackling both covariate and concept shifts as well as how to operationalize each of them. Built on the theoretical analysis, we develop a novel misinformation detection method that addresses both covariate and concept shifts. Using real-world datasets, we conduct extensive empirical evaluations to demonstrate the superior performance of our method over state-of-the-art misinformation detection methods as well as prevalent domain adaptation methods that can be tailored to solve the misinformation detection task.

Key words: misinformation detection, infodemic management, domain adaptation, covariate shift, concept shift, deep learning, contrastive learning, transfer learning, computational design science

1. Introduction

An infodemic refers to an overwhelming volume of true and false information spread during a disease outbreak (Van Der Linden 2022). Misinformation, or false information, in an infodemic misleads the public about the disease and causes significant harm to public health (Imhoff and Lamberty 2020, Freeman et al. 2022, Van Der Linden 2022). For example, the misinformation that the Ebola virus

is intentionally created by the Congolese government to eliminate people in the city of Beni has led to attacks on Ebola clinics by local residents, hindering timely treatments for Ebola.[1] More recently, during the outbreak of the coronavirus disease (COVID-19), a significant amount of misinformation has been diffused on social media. Examples include the misinformation that baking soda can cure coronavirus,[2] the false claim that injecting disinfectant can prevent the virus (Borah et al. 2022), and the incorrect attribution of the outbreak in Italy to Middle East illegal immigrants.[3] Such widespread of misinformation about COVID-19 has caused social panic, led people to dismiss health guidance, and weakened their confidence in vaccines, ultimately undermining pandemic response efforts (Bursztyn et al. 2020). Therefore, it is imperative to manage an infodemic to mitigate its adverse impact on public health (Freeman et al. 2022, Hwang and Lee 2025).

According to the World Health Organization (WHO), infodemic management is the use of risk- and evidence-based approaches to manage an infodemic and reduce its harm to public health.[4] The key to effective infodemic management is the detection of misinformation at the early stage of an infodemic (Buchanan 2020, Van Der Linden 2022). Early identification of misinformation discourages people from sharing it and prevents it from reaching a much larger population, thereby mitigating its potential harm to public health (Buchanan 2020). Moreover, the likelihood that a person trusts misinformation increases as the person is exposed longer to the misinformation (Zajonc 2001, Moravec et al. 2019). Early identification of misinformation reduces individuals' exposure time to misinformation, thereby lowering their likelihoods of trusting it. This, in turn, increases their chances of adhering to health guidance, ultimately benefiting public health as a whole.

An early stage infodemic features two salient characteristics. First, an emerging health event triggers a huge amount of true and false information spreading on various media platforms in a short time period. For example, during the early stage of the COVID-19 infodemic, a significant volume of true and false information concerning COVID-19 is diffused on social media (Yue et al. 2022), and the WHO declares a worldwide infodemic (Van Der Linden 2022). Second, it requires expert knowledge to distinguish between true information and misinformation in an infodemic. Moreover, during the early stage of a disease outbreak and its ensuing infodemic, even experts have no or limited knowledge about the disease, making it even more difficult to discern true information from misinformation. For example, during the early stage of the Ebola outbreak, scientists lack knowledge about the disease and its treatment (Adebimpe et al. 2015). Consequently, it is common

[1] See `https://news.un.org/en/story/2019/03/1034381` (last accessed on May 28, 2024)

[2] See `https://apnews.com/article/archive-fact-checking-8736262219` (last accessed on May 28, 2024)

[3] See `https://time.com/5789666/italy-coronavirus-far-right-salvini/` (last accessed on May 28, 2024)

[4] See `https://www.who.int/health-topics/infodemic#tab=tab_1` (last accessed on May 28, 2024)

that information disseminated at the early stage of an infodemic is unlabeled. As there is no labeled information (i.e., true or false), it is impossible to learn a misinformation detection model using conventional misinformation detection methods such as Abbasi et al. (2010), Nan et al. (2021), and Wei et al. (2022), which require labeled information in their training data.

A viable solution is to leverage labeled information in other domains to build a misinformation detection model for the infodemic domain. Indeed, popular fact-checking websites like PolitiFact[5] and Gossipcop[6] contain a wealth of verified true and false information in domains such as politics and entertainment. Moreover, misinformation across domains shares some common characteristics. For example, compared to true information, misinformation spreads faster, farther, and more widely (Oh et al. 2013, Vosoughi et al. 2018). As another example, fake news, an important type of misinformation, often features emotional headlines and demonstrates inconsistency between their headlines and contents (Siering et al. 2016). Because of these common characteristics shared across domains, it is possible to learn a misinformation detection model using labeled information in other domains and transfer it to the infodemic domain. However, information in each domain also has its unique characteristics. Understandably, the choice of words and the organization of words to convey sentiments and semantics differ across domains. For example, less than 20% of words used in politics news overlaps with those found in entertainment news (Shu et al. 2020). Because each domain possesses unique characteristics, a misinformation detection model learned with labeled information in other domains usually performs poorly in the infodemic domain.

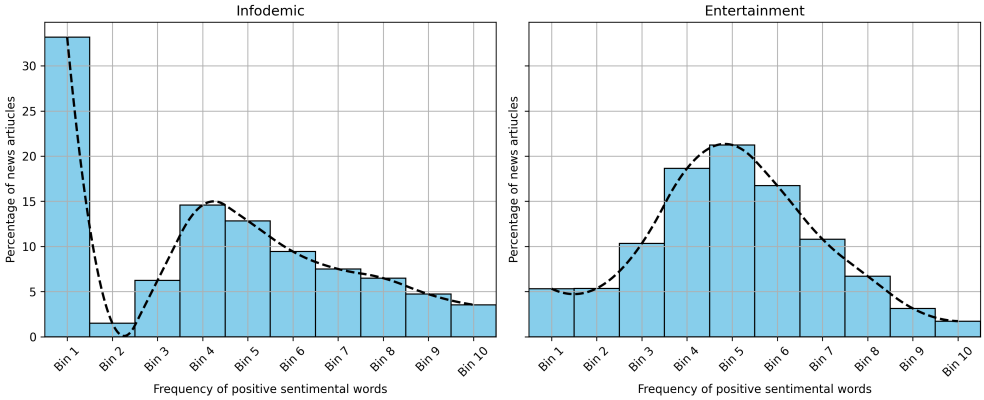
Domain adaptation is a widely employed technique to enhance the performance of a model trained in one domain but applied to another (Ben-David et al. 2010). It aims to mitigate disparities between data in different domains such that a model learned using data in one domain performs well in another domain (Ben-David et al. 2010). Concretely, data in a domain is represented as a joint distribution $p(\mathbf{x}, y), \mathbf{x} \in \mathcal{X}, y \in \mathcal{Y}$, where $\mathcal{X}$ denotes the feature space and $\mathcal{Y}$ is the label space (Liu et al. 2022). In the context of misinformation detection, $\mathbf{x}$ might be explicit features, such as the number of words in a piece of information, or the embedding of a piece of information (Zhou and Zafarani 2020). Label y denotes whether a piece of information is true or false. The joint distribution $p(\mathbf{x}, y)$ can be expressed as the product of $p(\mathbf{x})$ and $p(y|\mathbf{x})$, i.e., $p(\mathbf{x}, y) = p(\mathbf{x})p(y|\mathbf{x})$. Accordingly, to effectively alleviate disparities between data in different domains, it is essential to reduce both their differences in terms of $p(\mathbf{x})$ and their differences in terms of $p(y|\mathbf{x})$, where the former type of differences is referred to as *covariate shift* and the latter type of differences is known as *concept shift* (Liu et al. 2022). Figure 1 illustrates the presence of covariate shift and concept shift. For covariate shift, Figure

[5] See `https://www.politifact.com/` (last accessed on May 28, 2024)

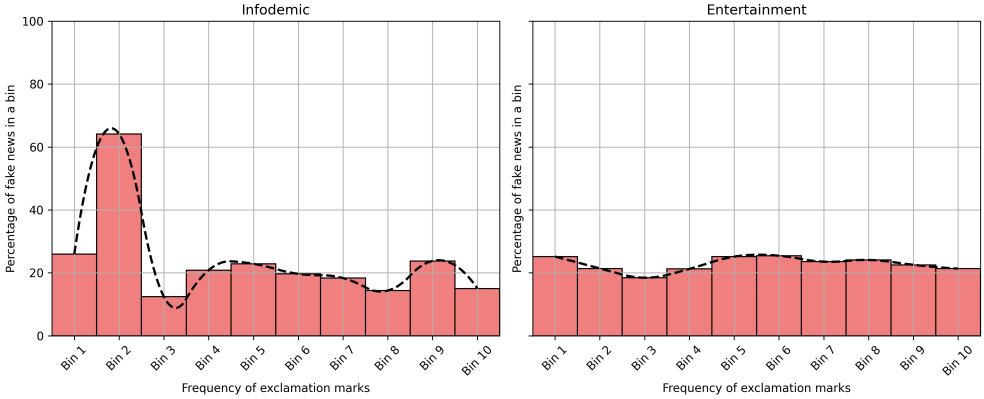
[6] See `https://en.wikipedia.org/wiki/Gossip_Cop` (last accessed on May 28, 2024)

1a shows that the distribution of a feature (i.e., the frequency of positive sentimental words in a news article) is skewed toward bin 1 in the infodemic domain, while it exhibits a bell-shaped curve in the entertainment domain. Figure 1b demonstrates concept shift by showing that the percentage of fake news conditioned on a feature (i.e., the frequency of exclamation marks in a news article) differs markedly between the two domains. Specifically, in the infodemic domain, the percentage of fake news spikes in bin 2, while it is stable across bins in the entertainment domain. A more detailed explanation of the figure is provided in Appendix A.

Figure 1 The Presence of Distribution Shifts across Domains.



(a) Covariate Shift between Infodemic and Entertainment domains.



(b) Concept Shift between Infodemic and Entertainment domains.

Note. Better view in color.

Existing domain adaptation methods focus on mitigating covariate shift between domains (Ganin et al. 2016, Long et al. 2017, Zhu et al. 2019, Peng et al. 2019). By applying these domain adaptation methods, Huang et al. (2021), Li et al. (2021), Yue et al. (2022) develop state-of-the-art misinformation detection methods, which learn a model from labeled information in one or more source domains to identify misinformation in a target domain with no labeled information.

Although these misinformation detection methods are applicable to our problem, their efficacy is limited because they fail to account for concept shift. In response to the research gaps, this study contributes to the literature with a novel misinformation detection method that tackles both covariate shift and concept shift. Our method design is grounded in our theoretical analysis, which shows the necessity of addressing both covariate shift and concept shift for effective domain adaptation as well as how to operationalize each of them. Guided by the theoretical analysis, we design two modules that respectively mitigate covariate shift and concept shift. Therefore, the primary methodological difference between our method and state-of-the-art domain adaptation-based misinformation detection methods (e.g., Yue et al. (2022)) lies in its module that mitigates concept shift. This module is directly informed by the theoretical analysis and provides a concrete way to operationalize and mitigate concept shift. Our method also differs from conventional misinformation detection methods such as Wei et al. (2022). Specifically, our method is a cross-domain method that learns a model from labeled information in other domains to classify unlabeled information in the infodemic domain. In contrast, conventional misinformation detection methods are single-domain methods that learn a model from labeled information in a domain (e.g., infodemic domain) to classify unlabeled information in the same domain.

2. Related Work

In this section, we review the Information Systems (IS) literature on misinformation, which not only situates our study in the IS literature but also reveals research gaps that motivate our study. We also survey existing methods for misinformation detection, which can be classified into three categories: single-domain, multi-domain, and cross-domain approaches. We then highlight the methodological novelty of our study.

2.1. IS Literature on Misinformation and Intervention Strategies

The IS literature on misinformation can be grouped into two themes: understanding why users believe and spread misinformation on social media platforms, and examining the effectiveness of intervention strategies mitigating such beliefs and behaviors. The belief of misinformation is driven by complex cognitive and contextual factors. Moravec et al. (2019) provide neurocognitive evidence that users tend to believe news headlines aligning with their political views, even when these headlines are labeled as misinformation. Their findings suggest a strong confirmation bias, in which ideological concordance takes precedence over analytical judgment. Pennycook and Rand (2021) challenge the prevailing notion that ideological motivation is the dominant driver of believing and sharing misinformation, instead suggesting that misinformation sharing is attributed to simple inattention. Adding another dimension, Mostagir and Siderius (2022) study social learning in the presence of misinformation, showing that Bayesian agents equipped with sophisticated reasoning

capabilities can be systematically wrong about the truth in environments with a large amount of misinformation.

Information content plays a key role in influencing the spread of misinformation. London et al. (2022) show that people share unverified messages, such as rumors, based on both content plausibility and non-content factors like vividness and sender credibility. Seol et al. (2024) examine the phenomenon of commingled partisan falsehoods — false statements embedded within otherwise accurate news articles. Through field and experimental studies, they show that such hybrid content significantly increases user sharing, especially when the content is distributed by reputable media firms. Spreading misinformation is also driven by broader social contextual factors. Oh et al. (2013) show that source ambiguity and increased public anxiety are key drivers of rumor propagation on social media during crisis events. Mostagir and Siderius (2023) use a network model to examine how structural inequality in access to accurate information affects vulnerability to misinformation. Their findings show that communities with limited access to knowledgeable agents are more susceptible to manipulation, and that moderate levels of inequality can, in some cases, result in worse misinformation outcomes than extreme inequality.

The effectiveness of intervention strategies has also been examined. One strategy is the use of fact-checking labels. Moravec et al. (2020) test both the intuitive (System 1) and deliberative (System 2) formats of such interventions and find that a combined approach is most effective in reducing the belief in misinformation. Bhattacherjee (2023) discovers the subtle effects of fact checking on a user's belief in misinformation, demonstrating that its effectiveness depends on factors such as the credibility of the fact checker and the strength of the user's attitude. Beyond tagging strategies, interventions that target users' cognitive processes have also been investigated. Moravec et al. (2022) demonstrate that prompting users to rate the truthfulness of news articles activates reflective thinking and reduces their belief in fake content. Hwang and Lee (2025) evaluate a platform-level policy adopted by Twitter, which nudges users searching for high-risk topics to credible sources. Their study shows that this policy significantly curbs the spread of misinformation by reducing both the original posting and subsequent sharing of false content. While the IS literature primarily investigates the misinformation problem from a behavioral perspective, few IS studies develop misinformation detection methods. In the following subsections, we review existing misinformation detection methods from both the IS and machine learning literature.

2.2. Single and Multi-domain Misinformation Detection Methods

Single-domain misinformation detection methods train a classification model using labeled information from a specific domain (e.g., true and fake news from the domain of politics), and then employ the model to classify unlabelled information in the same domain (e.g., unlabelled political

news). In this category, knowledge-based methods classify unlabelled information by comparing the content of the information (e.g., textual content of news) against known facts (Ciampaglia et al. 2015, Lee and Ram 2024). For example, Ciampaglia et al. (2015) retrieve pertinent facts from authoritative websites like Wikipedia, represent extracted facts as a knowledge graph, and employ the knowledge graph to automatically detect misinformation. Lee and Ram (2024) build a word network of a claim and its evidence, measure the level of consistency between them based on the network, and use the consistency score to predict the claim's veracity. Additionally, writing styles identified from the content of information, such as readability, sensationalism, informality, and subjectivity, have been utilized to detect misinformation (Zhou and Zafarani 2020). Recent methods consider the propagation of information on social media, in addition to the content of information (Papanastasiou 2020). Some methods employ propagation patterns to detect misinformation because misinformation spreads faster, farther, and more widely compared to true information (Vosoughi et al. 2018). Others harness the collective intelligence of crowds, such as likes, shares, and comments from social media users (Atanasov et al. 2017, Pennycook and Rand 2021). For example, Wei et al. (2022) employ both the content of news articles and social media users' comments on these news articles to detect fake news.

Multi-domain misinformation detection methods learn a classification model from labeled information in various domains (e.g., true and fake news in the politics, health, and sports domains), and then employ the model to classify unlabelled information in each of these domains (Nan et al. 2021, Zhu et al. 2023). The rationale underlying these methods is to leverage labeled information in multiple domains to enhance misinformation detection in each individual domain. However, because of variations in information content among different domains, a model learned with labeled information in these domains exhibits significant performance differences when applied to each of these domains (Zhu et al. 2023). That is, the effectiveness of the learned model in identifying misinformation in one domain could be significantly inferior to its effectiveness in detecting misinformation in another domain. To tackle this problem, several methods incorporate domain-specific characteristics into the model training. For example, Nan et al. (2021) propose to embed different domains as different vectors and integrate these vectors into the learning of a multi-domain misinformation detection model. Zhu et al. (2023) introduce a domain adapter for each specific domain, where each adapter is a collection of representative document embeddings from the corresponding domain. Representations of news articles in each domain are then adjusted by their corresponding domain adapter to train a multi-domain model.

2.3. Cross-domain Misinformation Detection Methods

Cross-domain misinformation detection methods leverage labeled information in one or more source domains (e.g., true and fake news in the politics and sports domains) to classify unlabelled information

in a target domain (e.g., unlabelled news in the health domain). Unlike multi-domain methods, which identify misinformation in various domains, cross-domain methods focus on detecting misinformation in one target domain. Depending on whether training data contain labeled or unlabeled information in a target domain, cross-domain methods can be further categorized into two subgroups.

One subgroup of cross-domain methods learns a model from training data consisting of labeled information in source domains as well as labeled information in a target domain to classify unlabeled information in the target domain (Mosallanezhad et al. 2022, Nan et al. 2022). These methods essentially utilize labeled source domain information as out-of-distribution data to improve the performance of misinformation detection in the target domain (Zhang et al. 2023). For example, Nan et al. (2022) pretrain a misinformation detection model using labeled information in source and target domains. Next, a language model is learned from labeled target domain information. Each piece of source domain information is then assigned a transferability score based on the degree to which the language model can predict its content. Finally, the pretrained model is fine-tuned using labeled source domain information weighted by their corresponding transferability scores as well as labeled target domain information.

The other subgroup learns a model from training data consisting of labeled information in source domains and unlabeled information in a target domain. Since there is no labeled target domain information, this subgroup of methods aims to transfer a model learned from labeled source domain information to the target domain. To this end, domain adaptation is a suitable and widely adopted technique. In the following, we review domain adaptation methods and their applications in cross-domain misinformation detection. In the literature of domain adaption, a domain consists of a feature space $\mathcal{X}$, a label space $\mathcal{Y}$, and a probability distribution $p(\mathbf{x}, y), \mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$ (Liu et al. 2022). As the probability distribution in a source domain usually differs from that in a target domain, a model learned from source domain data often performs poorly in the target domain (Ben-David et al. 2010). To tackle this issue, it is necessary to reduce the disparity between source and target domains in terms of $p(\mathbf{x}, y)$. Existing domain adaptation methods concentrate on mitigating the difference in feature distribution (i.e., $p(\mathbf{x})$) between source and target domains, known as covariate shift (Kouw and Loog 2018).[7] In this vein, Ganin et al. (2016) propose a domain adversarial network to learn domain-invariant features from data in source and target domains. Saito et al. (2018) devise

[7] In the field of data stream mining, concept shift or concept drift refers to the phenomenon that data samples with same features could have different labels at different timestamps (Roychowdhury et al. 2023). Given a model trained with previous data samples and a continuous influx of new data samples, the objective of data stream mining methods is to determine when and how to retrain the model (Vorburger and Bernstein 2006, Fang et al. 2013, Roychowdhury et al. 2023). These methods are not applicable to our problem because they require labeled data samples at all timestamps while instances of target domain information are unlabeled in our problem. Moreover, these methods aim to decide when and how to retrain a model in a dynamic environment, whereas the objective of our problem is to classify unlabeled target domain instances.

a mini-max mechanism to align feature distributions between source and target domains. In the maximum stage, two classifiers are learned from source samples such that the discrepancy between their classifications of target samples is maximized. In the minimum stage, features of target samples are adjusted to minimize such discrepancy. Rangwani et al. (2022) develop a domain adaption method that incorporates a smoothing mechanism into the method proposed by Ganin et al. (2016). Rostami and Galstyan (2023) construct a pseudo-dataset using a model learned from labeled source domain data and then mitigate feature distribution difference between the pseudo-dataset and target domain data.[8] While the methods reviewed above implicitly minimize covariate shift between source and target domains, there are methods explicitly measure and minimize covariate shift. For example, Long et al. (2015) utilize a metric known as the multi-kernel Maximum Mean Discrepancy (MMD) to quantify covariate shift, integrate this metric into a loss function, and minimize covariate shift by minimizing the loss function. In a follow-up study, Long et al. (2017) introduce another metric , known as the joint MMD, to measure covariate shift. In addition, Chen et al. (2022a) employ the nuclear-norm Wasserstein discrepancy to evaluate covariate shift between source and target domains. Recent studies have employed domain adaptation techniques to solve the cross-domain misinformation detection problem with unlabeled target domain information. For example, Li et al. (2021) extend the work of Ganin et al. (2016) to detect misinformation in a target domain by leveraging labeled information in multiple source domains whereas Ng et al. (2023) apply the domain adversarial network developed by (Ganin et al. 2016) to identify fake news and reviews. Huang et al. (2021) utilize the MMD metric to measure covariate shift between source domain information and target domain information, and mitigate covariate shift by minimizing the metric. Yue et al. (2022) develop a variant of the MMD metric to measure covariate shift between source and target domain information.

2.4. Key Novelty of Our Study

Our literature review suggests the following research gaps, as summarized in Table 1. First, existing single and multi-domain misinformation detection methods as well as cross-domain methods with labeled target domain information are not applicable to our problem. These methods require labeled target domain information (i.e., the infodemic domain in our study) in their training datasets, while our problem involves completely unlabeled information in the infodemic domain. Second, although existing cross-domain methods with unlabeled target domain information are applicable to our problem, they are less effective in solving the problem because they fail to address

[8] Although the title of the paper contains the phrase of concept shift, it actually mitigates covariate shift. This is evident in the objective function of the proposed algorithm, i.e., Equation (3) in (Rostami and Galstyan 2023). According to the equation, the algorithm aims to mitigate the difference between the feature distribution $p_T(\mathbf{X}_T)$ of target domain data and the feature distribution $\hat{p}_J(\mathbf{Z}_p)$ of the pseudo-dataset.

Table 1 Comparison Between Our Method and Existing Misinformation Detection Methods.

	Applicable to our problem	Addressing covariate shift	Addressing concept shift
Single- and multi-domain misinformation detection methods, e.g., Wei et al. (2022), Zhu et al. (2023)	No	No	No
Cross-domain misinformation detection methods with labeled target domain information, e.g., Mosallanezhad et al. (2022), Nan et al. (2022)	No	Yes	No
Cross-domain misinformation detection methods with unlabeled target domain information, e.g., Li et al. (2021), Yue et al. (2022)	Yes	Yes	No
Our method	Yes	Yes	Yes

concept shift. In fact, none of the existing misinformation detection methods is capable of tackling concept shift. However, to ensure the efficacy of a model learned using labeled source domain information in classifying unlabeled target domain information, it is essential to mitigate both covariate shift and concept shift between source and target domains (Liu et al. 2022). To this end, we propose a misinformation detection method that tackles both covariate shift and concept shift, in contrast to existing methods that focus on covariate shift solely. Therefore, the key methodological novelty of our method lies in its addressing of concept shift. To implement this methodological novelty, our study features two innovations. First, our theoretical analysis of domain adaptation not only underscores the importance of addressing both covariate shift and concept shift but also shows how to operationalize each of them. Second, informed by our theoretical analysis, we introduce a novel misinformation detection method, two modules of which are respectively designed to tackle covariate shift and concept shift. In Appendix B, we further demonstrate the practical advantages of our method by comparing it with existing misinformation detection methods along additional dimensions beyond domain adaptation.

3. Problem Formulation

We consider a misinformation detection problem at the early stage of an infodemic, when all instances of information in this domain are unlabeled. Concretely, let $\mathcal{D}_T$ be a dataset of N_T pieces of unlabeled information in the infodemic domain. The subscript T signifies target domain, which is the infodemic domain in this study. As an example, a piece of unlabeled information might encompass the textual content of a news article about COVID-19. Let $\mathbf{x}_i^T$ denote a set of features extracted from a piece of unlabeled information in $\mathcal{D}_T$, $i = 1, 2, \ldots, N_T$. We aim to predict the label (true or false) for each piece of information in $\mathcal{D}_T$.

To accomplish this objective, we are given k datasets of labeled information, i.e., $\mathcal{D}_{S_1}, \mathcal{D}_{S_2}, \ldots, \mathcal{D}_{S_k}$. The subscript S refers to source domain, which provides labeled information to facilitate the label prediction in the target domain. Concretely, $\mathcal{D}_{S_j}$ consists of N_{S_j} pieces of labeled information, $j = 1, 2, \ldots, k$. Considering an example of S_j being the politics domain, a piece of information in this domain could comprise the textual content of a political news article. Each piece of labeled information in $\mathcal{D}_{S_j}$ is represented as $(\mathbf{x}_i^{S_j}, y_i^{S_j})$, where $\mathbf{x}_i^{S_j}$ denotes features extracted from the information, $y_i^{S_j}$ is the label of the information, and $i = 1, 2, \ldots, N_{S_j}$. Label $y_i^{S_j} = 0$ indicates that the information is true and $y_i^{S_j} = 1$ shows that the information is false. We now formally define the problem of early detection of misinformation for infodemic management.

Definition 1 (Early Detection of Misinformation Problem (EDM)). Given a dataset $\mathcal{D}_T$ of N_T pieces of unlabeled information in the infodemic domain and k datasets of labeled information in various source domains, $\mathcal{D}_{S_1}, \mathcal{D}_{S_2}, \ldots, \mathcal{D}_{S_k}$, where $\mathcal{D}_{S_j}$ consists of N_{S_j} pieces of labeled information, $j = 1, 2, \ldots, k$, the objective of the problem is to learn a model from the data to classify each piece of information in $\mathcal{D}_T$ as true or false.

4. Method

In this section, we theoretically analyze the performance of a model learned using labeled source domain instances but applied to classify unlabeled target domain instances. Guided by the insights from our theoretical analysis, we then propose a novel method to solve the EDM problem.

4.1. Theoretical Analysis

We start with the setting of one target domain T and one source domain S. Let $\mathcal{D}_T$ be a dataset of N_T unlabelled instances in T and $\mathcal{D}_S$ be a dataset of N_S labeled instances in S. In the context of misinformation detection, an instance is a piece of information. We denote f_T as the true labeling function in the target domain that maps an instance with features $\mathbf{x}$ into the probability of the instance being 1, i.e., $f_T(\mathbf{x}) = p_T(y = 1|\mathbf{x})$. Similarly, f_S is the true labeling function in the source domain and $f_S(\mathbf{x}) = p_S(y = 1|\mathbf{x})$. Let h be a hypothesis learned from labeled instances in $\mathcal{D}_S$ and $h(\mathbf{x}) = p_h(y = 1|\mathbf{x})$. Following Ben-David et al. (2010), we define the source error $\epsilon_S(h)$ of hypothesis h as the expected difference between h and the true labeling function f_S in the source domain

$$\epsilon_S(h) = E_{\mathbf{x}\sim\mathbb{D}_S}[|h(\mathbf{x}) - f_S(\mathbf{x})|], \tag{1}$$

where $\mathbb{D}_S$ represents the feature distribution in the source domain. In a similar manner, we define the target error $\epsilon_T(h)$ of hypothesis h as the expected difference between h and the true labelling function f_T in the target domain

$$\epsilon_T(h) = E_{\mathbf{x}\sim\mathbb{D}_T}[|h(\mathbf{x}) - f_T(\mathbf{x})|], \tag{2}$$

where $\mathbb{D}_T$ denotes the feature distribution in the target domain. Recall that domain adaptation aims to learn h from labeled instances in $\mathcal{D}_S$ to classify each unlabelled instance in $\mathcal{D}_T$. Thus, the objective of domain adaptation is to minimize the target error $\epsilon_T(h)$. To achieve this objective, we show a bound of $\epsilon_T(h)$ in the following theorem.

Theorem 1 *For a source domain instance with features $\boldsymbol{x}_i^S$, let $c(\boldsymbol{x}_i^S)$ be the features of its nearest target domain instance, where $i = 1, 2, \ldots, N_S$ and the distance between a pair of instances are measured over their feature space. If f_T is L-Lipschitz continuous,*[9] *then for any $\eta \in (0, 1)$, with probability at least $(1-\eta)^2$,*

$$\epsilon_T(h) \leq \epsilon_S(h) + \hat{d}_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \frac{1}{N_S}\sum_{i=1}^{N_S} |f_S(\mathbf{x}_i^S) - f_T(c(\mathbf{x}_i^S))| + \frac{L}{N_S}\sum_{i=1}^{N_S} ||\mathbf{x}_i^S - c(\mathbf{x}_i^S)|| + C_1, \qquad (3)$$

where L is the Lipschitz constraint constant, C_1 is a constant, $\hat{d}_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ denotes the empirical $\mathcal{H}$-divergence between distributions $\mathbb{D}_S$ and $\mathbb{D}_T$, estimated using features of source domain instances in $\mathcal{D}_S$ and features of target domain instances in $\mathcal{D}_T$, and $||\boldsymbol{x}_i^S - c(\boldsymbol{x}_i^S)||$ represents the distance between $\boldsymbol{x}_i^S$ and $c(\boldsymbol{x}_i^S)$.

Proof. See Appendix D.

Theorem 1 establishes an upper bound of $\epsilon_T(h)$, which can be empirically computed using datasets $\mathcal{D}_S$ and $\mathcal{D}_T$ drawn from distributions $\mathbb{D}_S$ and $\mathbb{D}_T$, respectively. By Theorem 1, to minimize $\epsilon_T(h)$, we need to minimize the first three terms of its upper bound specified in Inequality (3). This is because the fourth term of the upper bound is inherently minimized, given that $c(\mathbf{x}_i^S)$ is the closest to $\mathbf{x}_i^S$, and the last term C_1 is a constant.

The first term $\epsilon_S(h)$ can be minimized by learning a classification model from labeled source domain instances. The second term $\hat{d}_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ is the empirical estimation of the $\mathcal{H}$-divergence between the source domain feature distribution $\mathbb{D}_S$ and the target domain feature distribution $\mathbb{D}_T$. $\mathcal{H}$-divergence has been extensively utilized in domain adaptation studies to measure distances between distributions. Please refer to Appendix C for a description of $\mathcal{H}$-divergence and its empirical estimation $\hat{d}_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$. Fundamentally, minimizing the $\mathcal{H}$-divergence between source and target domain feature distributions is to minimize covariate shift between source and target domains. Therefore, existing domain adaptation methods for mitigating covariate shift, such as Ganin et al. (2016) and Zhu et al. (2019), can be applied to minimize $\hat{d}_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$.

For the third term $\frac{1}{N_S}\sum_{i=1}^{N_S} |f_S(\mathbf{x}_i^S) - f_T(c(\mathbf{x}_i^S))|$, recall that $f_S(\mathbf{x}_i^S) = p_S(y = 1|\mathbf{x}_i^S)$ and $f_T(c(\mathbf{x}_i^S)) = p_T(y = 1|c(\mathbf{x}_i^S))$. Hence, to minimize the third term, we need to minimize the summation of differences between $p_S(y = 1|\mathbf{x}_i^S)$ and $p_T(y = 1|c(\mathbf{x}_i^S))$, summed over all source

[9] The L-Lipschitz continuous assumption is commonly used in theoretical analyses of machine learning algorithms (e.g., Arjovsky et al. 2017, Asadi et al. 2018, Kim et al. 2021).

domain instances and their corresponding nearest target domain instances. Therefore, minimizing the third term is essentially to minimize concept shift between source and target domains.

Next, we consider the setting of one target domain T and multiple source domains $S_1, S_2, \ldots, S_k$. Similar to the notations used in the setting of one source domain, we denote $\mathcal{D}_T$ as a dataset of N_T unlabelled instances in T and $\mathcal{D}_{S_j}$ as a dataset of N_{S_j} labeled instances in S_j, $j = 1, 2, \ldots, k$. Let h be a hypothesis learned from labeled instances in $\mathcal{D}_{S_1}, \mathcal{D}_{S_2}, \ldots, \mathcal{D}_{S_k}$. We denote f_T and f_{S_j} as the true labeling functions in T and S_j, respectively, $j = 1, 2, \ldots, k$. For a hypothesis h, $\epsilon_T(h)$ and $\epsilon_{S_j}(h)$ represent its errors in T and S_j, respectively, $j = 1, 2, \ldots, k$. Built on Theorem 1, the following proposition gives an upper bound of $\epsilon_T(h)$ in the setting of multiple source domains.

Proposition 1 *For a source domain instance with features $\boldsymbol{x}_i^{S_j}$ in $\mathcal{D}_{S_j}$, let $c(\boldsymbol{x}_i^{S_j})$ be the features of its nearest instance in $\mathcal{D}_T$, where $j = 1, 2, \ldots, k$ and $i = 1, 2, \ldots, N_{S_j}$. If f_T is L-Lipschitz continuous, then for any $\eta \in (0, 1)$, with probability at least $(1-\eta)^{2k}$,*

$$\epsilon_T(h) \leq \frac{1}{k}\sum_{j=1}^{k}\{\epsilon_{S_j}(h) + \hat{d}_{\mathcal{H}}(\mathcal{D}_{S_j}, \mathcal{D}_T) + \frac{1}{N_{S_j}}\sum_{i=1}^{N_{S_j}} |f_{S_j}(\mathbf{x}_i^{S_j}) - f_T(c(\mathbf{x}_i^{S_j}))| + \frac{L}{N_{S_j}}\sum_{i=1}^{N_{S_j}} ||\mathbf{x}_i^{S_j} - c(\mathbf{x}_i^{S_j})|| + C_j\}, \tag{4}$$

where L is the Lipschitz constraint constant, C_j is a constant, and $\hat{d}_{\mathcal{H}}(\mathcal{D}_{S_j}, \mathcal{D}_T)$ denotes the empirical $\mathcal{H}$-divergence between the feature distribution in the source domain S_j and that in the target domain T.

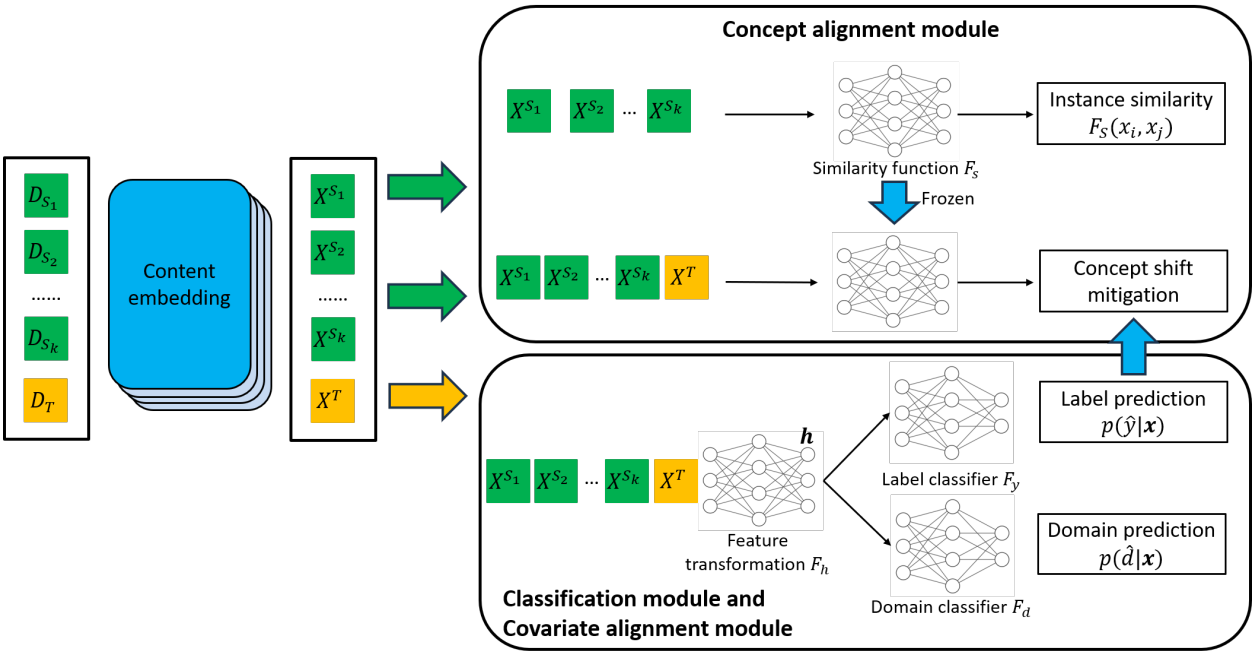
Proof. See Appendix D.

Proposition 1 offers useful insights for solving the EDM problem, which aims to learn a model from labeled information in multiple source domains to classify unlabeled information in the target infodemic domain with minimum error (i.e., $\epsilon_T(h)$). Informed by the proposition, we design a method that minimizes $\epsilon_T(h)$ through its three modules, each dedicated to reducing one of the first three terms in Inequality (4). Specifically, to reduce $\sum_{j=1}^{k} \epsilon_{S_j}(h)$, we implement a classification module, which is trained to accurately predict the label for each piece of source domain information. A covariate alignment module is developed to reduce $\sum_{j=1}^{k} \hat{d}_{\mathcal{H}}(\mathcal{D}_{S_j}, \mathcal{D}_T)$, thus diminishing covariate shift between source domains and the target domain. Finally, we propose a concept alignment module to mitigate $\sum_{j=1}^{k}\sum_{i=1}^{N_{S_j}} |f_{S_j}(\mathbf{x}_i^{S_j}) - f_T(c(\mathbf{x}_i^{S_j}))|$, thereby reducing concept shift between source domains and the target domain. The concept alignment module constitutes the main methodological novelty of our method, and we designate our method as Domain Adaptation with Concept Alignment (DACA).

4.2. Method Overview

Grounded in our theoretical analysis, the DACA method adeptly alleviates disparities between source and target domains through its covariate alignment and concept alignment modules. As a result, the model learned from labeled source domain information by its classification module can effectively

Figure 2 Overall Architecture of Domain Adaptation with Concept Alignment (DACA) Method



Note. Better view in color.

classify unlabeled infodemic domain information. Figure 2 illustrates the overall architecture of the DACA method. As shown, the inputs to the method consist of k datasets of labeled information in various source domains, $\mathcal{D}_{S_1}, \mathcal{D}_{S_2}, \ldots, \mathcal{D}_{S_k}$, and a dataset $\mathcal{D}_T$ of unlabeled information in the target infodemic domain. A content embedding function based on the method proposed by Zhu et al. (2023) represents each piece of input information as a vector. Specifically, it extracts and represents three types of features from an instance of input information (e.g., a news article): semantic, style, and emotion features. More concretely, the semantic feature vector is extracted by first encoding the input information as a token embedding sequence via the pretrained RoBERTa model (Liu et al. 2019) and then aggregating the sequence into a single vector using the TextCNN model (Kim 2014). Style features, such as the readability score, are derived from the input information and then converted into the style feature vector via a multi-layer perceptron transformation (Zhu et al. 2023). Similarly, emotion features, such as the counts of emoticons, pronouns, and uppercase letters, are computed from the input information and transformed into the emotion feature vector. Finally, the three vectors are aggregated into a single vector representing the input information through an attention mechanism (Zhu et al. 2023). Let $X^{S_j} = \{\mathbf{x}_i^{S_j} | i = 1, 2, \ldots, N_{S_j}\}$ and $X^T = \{\mathbf{x}_i^T | i = 1, 2, \ldots, N_T\}$ denote the set of embedding vectors extracted from input information in source domain S_j and target domain T, respectively, $j = 1, 2, \ldots, k$. These features serve as the inputs to the three modules of the DACA method.

The design of these modules is guided by the theoretical results in Theorem 1 and Proposition 1. Specifically, to minimize the target error, it is necessary to reduce the first three terms in Inequality (4). Accordingly, each module is designed to reduce one of these three terms. The classification module aims to reduce the term $\sum_{j=1}^{k} \epsilon_{S_j}(h)$ in Inequality (4). To this end, it trains a label classifier to accurately predict the label of a piece of information as true (0) or false (1). Concretely, the label classifier takes the features $\mathbf{x}$ of each piece of information as input and predicts its probability of being 1, i.e., $p(\hat{y}=1|\mathbf{x})$. Given that only source domain information is labeled, the label classifier is trained using labelled source domain information. The covariate alignment module trains a domain classifier to predict whether a piece of information is from a source domain or the target domain. As the objective of this module is to mitigate the feature distribution discrepancy between source domains and the target domain (i.e., reducing the term $\sum_{j=1}^{k} \hat{d}_{\mathcal{H}}(\mathcal{D}_{S_j}, \mathcal{D}_T)$ in Inequality (4)), it learns representations of information instances such that a well-trained domain classifier cannot discern whether an information instance is from a source domain or the target domain (Ganin et al. 2016).

Finally, to mitigate the term $\sum_{j=1}^{k} \sum_{i=1}^{N_{S_j}} |f_{S_j}(\mathbf{x}_i^{S_j}) - f_T(c(\mathbf{x}_i^{S_j}))|$ in Inequality (4), we propose a concept alignment module. Proposition 1 sheds light on the design of this module. According to the proposition, for each piece of information in a source domain, we need to find its nearest target domain information. Therefore, it is imperative to have a function that measures the similarity between two pieces of information over their feature space. To this end, we introduce a submodule that defines and learns a similarity function using labeled source domain information. Utilizing the similarity function, a concept shift mitigation submodule identifies the nearest target domain information for each piece of source domain information.[10] By Proposition 1, this submodule then mitigates concept shift between source domains and the target domain by minimizing $\sum_{j=1}^{k} \sum_{i=1}^{N_{S_j}} |f_{S_j}(\mathbf{x}_i^{S_j}) - f_T(c(\mathbf{x}_i^{S_j}))|$, where $f_{S_j}(\mathbf{x}_i^{S_j})$ and $f_T(c(\mathbf{x}_i^{S_j}))$ are estimated using the label classifier in the classification module.

4.3. Classification Module and Covariate Alignment Module

As depicted in Figure 2, the classification module and the covariate alignment module share a feature transformation layer. This layer, denoted by F_h, takes each feature vector $\mathbf{x}$ in X^{S_1}, X^{S_2}, ..., X^{S_k}, X^T as input and outputs the transformed vector $\mathbf{h}$. Formally, we have

$$\mathbf{h} = F_h(\mathbf{x}) = \text{MLP}_1(\mathbf{x}; \theta_h), \tag{5}$$

where MLP_1 is a multi-layer perceptron (MLP) with trainable parameters θ_h.

[10] The time complexity of computing the similarity between each piece of source domain information and every piece of target domain information is $O(\sum_{j=1}^{k} N_{S_j} N_T)$, where the source domain $\mathcal{D}_{S_j}$ consists of N_{S_j} pieces of labeled information, $j = 1, 2, \ldots, k$, and the target domain $\mathcal{D}_T$ contains N_T pieces of unlabeled information.

The classification module features an MLP-based label classifier, denoted by F_y, which takes the transformed feature vector $\mathbf{h}$ as input and outputs the probability of being labeled as 1 (i.e., false) for the piece of information characterized by feature vector $\mathbf{x}$. Concretely, we have

$$p(\hat{y}=1|\mathbf{x}) = F_y(\mathbf{h}) = \text{Sigmoid}\big(\text{MLP}_2(\mathbf{h};\theta_y)\big), \tag{6}$$

where $\hat{y}$ is the predicted label for the piece of information, MLP_2 is an MLP with the ReLU activation function for its internal layers and trainable parameters θ_y, and Sigmoid() denotes the sigmoid function that converts the output of MLP_2 to a probability value. The classification module is trained by minimizing the total cross-entropy loss summed over all labeled source domain information:

$$\begin{aligned}\mathcal{L}_y(\theta_h,\theta_y) &= -\sum_{\mathbf{x}\in X^S}\Big[y\log\big(p(\hat{y}=1|\mathbf{x})\big)+(1-y)\log\big(1-p(\hat{y}=1|\mathbf{x})\big)\Big]\\ &= -\sum_{\mathbf{x}\in X^S}\Big[y\log F_y(F_h(\mathbf{x}))+(1-y)\log\big(1-F_y(F_h(\mathbf{x}))\big)\Big],\end{aligned} \tag{7}$$

where $X^S = \cup_{j=1}^k X^{S_j}$ and y is the true label of $\mathbf{x}$. Minimizing $\mathcal{L}_y$ trains the classification module to score high probabilities for true labels, thereby inducing a model with low source error (i.e., $\sum_{j=1}^k \epsilon_{S_j}(h)$ in Proposition 1).

The covariate alignment module aims to mitigate covariate shift between source domains and the target domain, i.e., reducing $\sum_{j=1}^k \hat{d}_{\mathcal{H}}(\mathcal{D}_{S_j},\mathcal{D}_T)$ in Proposition 1. To this end, we adapt the domain adversarial method developed by Ganin et al. (2016) to implement this module. Let d denote the true domain for a piece of information characterized by feature vector $\mathbf{x}$, where $d=0$ if it belongs to a source domain and $d=1$ if it comes from the target domain. The module employs a domain classifier, denoted by F_d, that takes the transformed feature vector $\mathbf{h}$ as input and outputs its probability of belonging to the target domain. Mathematically, we have

$$p(\hat{d}=1|\mathbf{x}) = F_d(\mathbf{h}) = \text{Sigmoid}\big(\text{MLP}_3(\mathbf{h};\theta_d)\big), \tag{8}$$

where $\hat{d}$ is the predicted domain (i.e., source or target) for $\mathbf{x}$, and MLP_3 is an MLP with the ReLU activation function for its internal layers and trainable parameters θ_d. On the one hand, F_d is trained to accurately predict the domain for each piece of information. On the other hand, to reduce covariate shift, the vector $\mathbf{h}$ produced by the feature transformation layer F_h should be domain-invariant, i.e., when presented to F_d, the classifier cannot tell whether it is from a source domain or the target domain (Li et al. 2021). With this reasoning in mind, the learning of domain-invariant features can be formulated as a min-max game defined as follows(Ganin et al. 2016, Li et al. 2021):

$$\begin{aligned}&\min_{\theta_h}\max_{\theta_d}\sum_{\mathbf{x}\in X}\Big[d\log\big(p(\hat{d}=1|\mathbf{x})\big)+(1-d)\log\big(1-p(\hat{d}=1|\mathbf{x})\big)\Big]\\ &=\min_{\theta_h}\max_{\theta_d}\sum_{\mathbf{x}\in X}\Big[d\log F_d(F_h(\mathbf{x}))+(1-d)\log\big(1-F_d(F_h(\mathbf{x}))\big)\Big],\end{aligned} \tag{9}$$

where $X = (\cup_{j=1}^{k} X^{S_j}) \cup X^T$ and d is the true domain (i.e., source or target) that $\mathbf{x}$ belongs to. The maximization objective in Equation 9 promotes a more accurate domain classifier, while the minimization objective strengthens the domain invariance of the transformed feature vector $\mathbf{h}$ by making hard even for the most accurate domain classifier to discern its domain identity. Following Ganin et al. (2016), the two adversarial objectives in Equation 9 can be unified using a gradient reversal layer $R(\mathbf{z})$ defined as:

$$R(\mathbf{z}) = \mathbf{z}; \frac{dR(\mathbf{z})}{d\mathbf{z}} = -\mathbf{I}, \tag{10}$$

where z denotes an input vector to the gradient reversal layer, and $\mathbf{I}$ is an identity matrix. According to Equation 10, the gradient reversal layer makes no change to its input but flips the sign of the gradient passed through it. Equipped with the gradient reversal layer, solving the min-max problem in Equation 9 is equivalent to minimize the following loss (Ganin et al. 2016):

$$\mathcal{L}_d(\theta_h, \theta_d) = -\sum_{\mathbf{x} \in X} \Big[d \log F_d(R(F_h(\mathbf{x}))) + (1-d) \log \left(1 - F_d(R(F_h(\mathbf{x})))\right) \Big], \tag{11}$$

where we replace $F_h(\mathbf{x})$ in Equation 9 with $R(F_h(\mathbf{x}))$. To minimize the loss defined in Equation 11, parameters θ_d of the domain classifier F_d are updated with gradient descent, leading to a more accurate domain classifier. In contrast, because of the gradient reversal layer, parameters θ_h of the feature transformation layer F_h are adjusted with gradient upscent, resulting in more domain-invariant $\mathbf{h}$ (Ganin et al. 2016).

4.4. Concept Alignment Module

The key novelty of our method is the proposed concept alignment module, which mitigates concept shift between each source domain and the target domain. By Proposition 1, to mitigate concept shift, we need to learn a function that measures the similarity between two pieces of information. To this end, we propose a similarity function submodule. Specifically, for two pieces of information respectively characterized by feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, we define the similarity between them as

$$F_s(\mathbf{x}_i, \mathbf{x}_j) = \frac{\text{MLP}_4(\mathbf{x}_i; \theta_s) \cdot \text{MLP}_4(\mathbf{x}_j; \theta_s)}{||\text{MLP}_4(\mathbf{x}_i; \theta_s)||\,||\text{MLP}_4(\mathbf{x}_j; \theta_s)||}, \tag{12}$$

where F_s denotes the similarity function, MLP_4 is an MLP with learnable parameters θ_s that transforms a feature vector (e.g., $\mathbf{x}_i$) into a different representation space, notation $\cdot$ denotes the dot product of two vectors, and $||\ ||$ represents the L^2 norm of a vector. The objective of MLP_4 is to transform feature vectors into a representation space where information instances with the same label exhibit greater similarity than instances with different labels. This transformation facilitates concept alignment because differently labeled information instances (i.e., true information instances vs. false information instances) are better separated in the new representation space compared to the original one.

To learn the parameters θ_s of MLP_4, we employ a contrastive learning technique that maximizes the similarities between a focal information instance and its positive peers while minimizing the similarities with its negative peers. Peers can be defined in two ways (Le-Khac et al. 2020): the supervised way, which uses label information to identify positive and negative peers for a focal instance (Khosla et al. 2020), and the unsupervised way, which relies on heuristics to define peer relationships (Chen et al. 2020). Given that information instances in source domains have labels, we opt for the supervised way. Specifically, for each piece of source domain information characterized by feature vector $\mathbf{x}$, we randomly select another piece of source domain information, bearing the same label, as its positive peer (Gunel et al. 2021). We denote the feature vector of the positive peer as $\mathbf{x}^+$. In addition, we randomly select m other pieces of source domain information, each having the opposite label of the focal information instance characterized by $\mathbf{x}$, to serve as its negative peers (Chen et al. 2022b). We denote the feature vectors of these negative peers as $\mathbf{x}_l^-$, $l = 1, 2, \ldots, m$. The objective is to maximize the similarity $F_s(\mathbf{x}, \mathbf{x}^+)$ between $\mathbf{x}$ and $\mathbf{x}^+$, while minimizing the similarity $F_s(\mathbf{x}, \mathbf{x}_l^-)$ between $\mathbf{x}$ and $\mathbf{x}_l^-$, $l = 1, 2, \ldots, m$. Accordingly, the parameters θ_s are learned by minimizing the InfoNCE loss commonly used by contrastive learning methods (Oord et al. 2018, Chen et al. 2020):

$$\mathcal{L}_s(\theta_s) = -\sum_{\mathbf{x} \in X^S} \log \frac{\exp(F_s(\mathbf{x}, \mathbf{x}^+)/\tau)}{\exp(F_s(\mathbf{x}, \mathbf{x}^+)/\tau) + \Sigma_{l=1}^m \exp(F_s(\mathbf{x}, \mathbf{x}_l^-)/\tau)}, \tag{13}$$

where $X^S = \cup_{j=1}^k X^{S_j}$, F_s is given in Equation 12, and $0 < \tau < 1$ is a hyperparameter.[11] By minimizing $\mathcal{L}_s$ w.r.t. θ_s, this submodule learns F_s that measures the similarity between two pieces of information in their transformed representation space such that identically labeled information instances are close to each other and distinctly labeled information instances are distant from each other.

Applying the similarity function F_s, for each piece of source domain information characterized by feature vector $\mathbf{x}_i^{S_j}$, the concept shift mitigation submodule identifies its most similar target domain information characterized by feature vector $c(\mathbf{x}_i^{S_j})$, $j = 1, 2, \ldots, k$ and $i = 1, 2, \ldots, N_{S_j}$. It then mitigates concept shift between source domains and the target domain by minimizing $\sum_{j=1}^k \sum_{i=1}^{N_{S_j}} |f_{S_j}(\mathbf{x}_i^{S_j}) - f_T(c(\mathbf{x}_i^{S_j}))|$ in Proposition 1. To this end, $f_{S_j}(\mathbf{x}_i^{S_j})$ and $f_T(c(\mathbf{x}_i^{S_j}))$ are estimated using the label classifier described in Section 4.3. Specifically, we have

$$\hat{f}_{S_j}(\mathbf{x}_i^{S_j}) = p(\hat{y} = 1 | \mathbf{x}_i^{S_j}) = F_y(F_h(\mathbf{x}_i^{S_j})),$$

[11] Hyperparameter τ adjusts the emphasis on the two objectives (Frosst et al. 2019). Lower τ enables the learning of the similarity function to place more emphasis on maximizing $F_s(\mathbf{x}, \mathbf{x}^+)$ while higher τ shifts the learning to emphasize more on minimizing $F_s(\mathbf{x}, \mathbf{x}_l^-)$.

where $\hat{f}_{S_j}$ is the estimation of f_{S_j}, and F_h and F_y are defined by Equations 5 and 6, respectively. Similarly, $f_T(c(\mathbf{x}_i^{S_j}))$ is estimated by

$$\hat{f}_T(c(\mathbf{x}_i^{S_j})) = p(\hat{y}=1|c(\mathbf{x}_i^{S_j})) = F_y(F_h(c(\mathbf{x}_i^{S_j}))).^{12}$$

With the estimations of f_{S_j} and f_T, the concept shift mitigation submodule reduces concept shift by minimizing $\sum_{j=1}^{k}\sum_{i=1}^{N_{S_j}} |F_y(F_h(\mathbf{x}_i^{S_j})) - F_y(F_h(c(\mathbf{x}_i^{S_j})))|$. Accordingly, this submodule is trained by minimizing the following loss:

$$\mathcal{L}_c(\theta_h,\theta_y) = \sum_{\mathbf{x}\in X^S}\Big[\big(F_y(F_h(\mathbf{x})) - F_y(F_h(c(\mathbf{x})))\big)^2 - \big(F_y(F_h(\mathbf{x})) - F_y(F_h(d(\mathbf{x})))\big)^2\Big], \tag{14}$$

where $X^S = \cup_{j=1}^{k} X^{S_j}$. For each piece of source domain information with feature vector $\mathbf{x}\in X^S$, $c(\mathbf{x})$ and $d(\mathbf{x})$ respectively denote the feature vectors characterizing its most similar and most dissimilar target domain information, both of which are identified using the similarity function F_s. Minimizing the first term in $\mathcal{L}_c$ essentially minimizes $\sum_{j=1}^{k}\sum_{i=1}^{N_{S_j}} |F_y(F_h(\mathbf{x}_i^{S_j})) - F_y(F_h(c(\mathbf{x}_i^{S_j})))|$. Consequently, it enforces the desired property that similar source and target domain information have similar label predictions (Liu et al. 2022), thus reducing concept shift between source domains and the target domain. Minimizing the second term $-\big(F_y(F_h(\mathbf{x})) - F_y(F_h(d(\mathbf{x})))\big)^2$ maximizes the label prediction difference between $F_y(F_h(\mathbf{x}))$ and $F_y(F_h(d(\mathbf{x})))$. As a result, label predictions diverge for dissimilar source and target domain information, which further mitigates concept shift.

4.5. The DACA Method

The DACA method is trained to minimize the combined losses of its three modules:

$$\mathcal{L}_{\text{DACA}}(\theta_h,\theta_y,\theta_d,\theta_s) = \mathcal{L}_y(\theta_h,\theta_y) + \mathcal{L}_d(\theta_h,\theta_d) + \mathcal{L}_s(\theta_s) + \mathcal{L}_c(\theta_h,\theta_y) \tag{15}$$

where $\mathcal{L}_y$, $\mathcal{L}_d$, $\mathcal{L}_s$, and $\mathcal{L}_c$ are defined by Equations 7, 11, 13, and 14, respectively. The losses $\mathcal{L}_y$ and $\mathcal{L}_d$ are associated with the classification module and the covariate alignment module, respectively. The losses $\mathcal{L}_s$ and $\mathcal{L}_c$ correspond to the similarity function and concept shift mitigation submodules of the concept alignment module, respectively. There are two practical considerations when training DACA. First, the DACA method is trained in a two-stage manner. In the first stage (or the warmup stage), the method is trained by minimizing $\mathcal{L}_{\text{DACA}} - \mathcal{L}_c$. The rationale is that, to accurately estimate the concept alignment loss $\mathcal{L}_c$, DACA needs to learn a reliable similarity function first. In the second stage, DACA is trained by minimizing $\mathcal{L}_{\text{DACA}}$ as defined by Equation 15. Second, to speed up the training

[12] The label classifier is trained to minimize the error of classifying source domain information. Thus, it is reasonable to estimate f_{S_j} using the label classifier. Since target domain information is unlabeled, we proxy f_T using the label classifier, a strategy of which the effectiveness has been validated by previous studies (Long et al. 2013, 2018, Tachet des Combes et al. 2020).

of the DACA method, it is beneficial to mitigate concept shift between each pair of source domains. This is accomplished by minimizing a loss function for each pair of source domains. Concretely, for a pair of source domains S_j and S_n, the loss function has the same form as Equation 14 but with S_j treated as the source domain in the equation and S_n regarded as the target domain. Once trained, DACA predicts the probability that a piece of target domain information characterized by feature vector $\mathbf{x}_T$ is false as $F_y(F_h(\mathbf{x}_T))$, where F_h and F_y are defined by Equations 5 and 6, respectively.

5. Empirical Evaluation

5.1. Data

We evaluated the performance of our proposed DACA method using the publicly available datasets of English news, which have been widely employed to assess the performance of misinformation detection methods (Nan et al. 2021, Mosallanezhad et al. 2022, Zhu et al. 2023). One is the MM-COVID dataset, which contains 4,750 pieces of true news (i.e., true information) and 1,317 pieces of fake news (i.e., misinformation) on COVID-19 as well as user comments on these news (Li et al. 2020). Specifically, 8% of COVID news are accompanied by user comments. In our evaluation, we treated the COVID domain as the infodemic or target domain. For source domains, we utilized the FakeNewsNet dataset, which consists of true and fake news alongside their associated comments from the domains of entertainment and politics (Shu et al. 2020).[13] Specifically, there are 16,804 pieces of true news and 5,067 pieces of fake news from the entertainment domain, and 1,583 instances of true news and 1,287 instances of fake news from the politics domain. User comments accompany 27% of entertainment news and 59% of politics news. The average comment length is 8.2 words in the COVID domain, 7.7 words in the entertainment domain, and 8.2 words in the politics domain. Table 2 reports the summary statistics of the datasets used in our evaluation. Examples of true and fake news from the source and target domains are given in Figure 3.

Table 2 Summary Statistics of Evaluation Datasets.

Domain	Number of true news	Number of fake news
COVID (Infodemic / Target)	4,750	1,317
Entertainment (Source)	16,804	5,067
Politics (Source)	1,583	1,287

[13] Both datasets were constructed by extracting true and fake news articles from reputable fact-checking websites, such as PolitiFact (`https://www.politifact.com`), where domain experts and professional journalists provide verified assessments on the veracity of news articles (Shu et al. 2020, Li et al. 2020) Social media posts on X (formerly Twitter) containing links to these fact-checked news articles were collected and treated as user comments (Shu et al. 2020, Li et al. 2020). While most users comment only once on an article, multiple comments on an article from the same user are occasionally observed.

Figure 3 Examples of True and Fake News in Evaluation Datasets

COVID – True news

News content: Across San Francisco, trips to workplaces, parks, transit stations, and stores have collectively fallen to about 40% of normal levels since late February, as the region and then state enacted strict social distancing measures to halt the spread of the coronavirus. ...

User comment: wow, how did I miss this? Nice piece!

COVID – Fake news

News content: To prevent the spread of COVID-19, the Centers for Disease Control and Prevention (CDC) now recommends that Americans wear wearing cloth masks in public settings where other social distancing measures are difficult to maintain. ...

No user comments.

Entertainment – True news

News content: Update: Michael Phelps defended his Shark Week finale episode that left some viewers lukewarm on Sunday. ...

User comment: Seriously who cares how he did it. Doesn't change he's an Olympic gold medalist.

Politics – Fake news

News content: The state of Florida can now be added to the growing list of US states that have past bills to legalize the use marijuana. ...

User comment: This is not accurate.

5.2. Evaluation Procedure and Benchmark Methods

During the early stage of an infodemic, there is little understanding of the disease that causes the infodemic. Consequently, even experts encounter difficulty distinguishing true news from fake news in the infodemic domain, leaving all news in this domain unlabeled. Hence, this scenario entails each method (ours or a benchmark) utilizing labeled news in the entertainment and politics domains to predict the label for each piece of news in the COVID domain. Accordingly, the inputs to each method encompass labeled news in the entertainment and politics domains, along with user comments on these news, as well as *unlabeled* news in the COVID domain and their associated comments. As the infodemic progresses, experts gradually acquire more knowledge, leading to a small portion of the infodemic information being labeled. Consequently, we also consider scenarios of partially labeled infodemic information. Concretely, we compared our method with the benchmarks by increasing the number of labeled news articles in the COVID domain from 50 to 150, in increments of 50. For example, each method utilizes labeled news in the entertainment and politics domains, together with 50 labeled news articles in the COVID domain, to predict the label for each unlabeled news article in the COVID domain.

In each evaluation scenario, we conducted 25 experiments for every method (ours or a benchmark) and measured its performance using the metrics precision, recall, F1-score, and F2-score. To curb the spread of fake news and minimize their societal impact, it is crucial to identify as many instances of fake news as possible (Zhu et al. 2023). To this end, recall is an important metric as it measures

the effectiveness of a method in identifying fake news. In addition, F1-score evaluates a method's performance in both identifying fake news and avoiding predicting true news as fake. Similarly, F2-score quantifies the performance in both precision and recall, with more weight on recall. Concretely, let P be the number of fake news, TP be the number of fake news that are predicted as fake, and FP be the number of true news that are predicted as fake. For this case, precision is defined as $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, recall is defined as TP/P, F1-score is computed as $2\mathrm{TP}/(\mathrm{TP}+\mathrm{FP}+\mathrm{P})$, and F2-score, which weights recall higher than precision, is given by $5\mathrm{TP}/(\mathrm{TP}+\mathrm{FP}+4\mathrm{P})$.

As reviewed in Section 2 and summarized in Table 1, among all existing misinformation detection methods, only cross-domain misinformation detection methods with unlabeled target domain information can solve the EDM problem investigated in this paper. Therefore, we benchmarked our method against state-of-the-art methods in this category. Specifically, one benchmark is the MMD method, which minimizes covariate shift between source and target domain information, measured using the maximum mean discrepancy (MMD) metric (Huang et al. 2021). Another benchmark is the contrastive adaptation network for misinformation detection (CANMD) method, which assesses covariate shift using a variant of the MMD metric (Yue et al. 2022). In addition, we also compared our method against the domain adversarial neural network (DANN) method; it minimizes covariate shift by learning domain-invariant features from source and target domain information (Li et al. 2021). Moreover, general-purpose domain adaptation methods that were not originally designed for misinformation detection can also be adapted to solve our problem. As analyzed in Section 2.3, there are two categories of domain adaptation methods. One category explicitly measures and minimizes covariate shift while the other category implicitly minimizes covariate shift by learning domain-invariant features. Accordingly, we benchmarked against representative methods in each category. For the category that explicitly measures and minimizes covariate shift, we considered the multi-kernel MMD (MK-MMD) method (Long et al. 2015), the joint adaptation network (JAN) method (Long et al. 2017), and the discriminator-free adversarial learning network (DALN) method (Chen et al. 2022a) as our baselines. MK-MMD and JAN (Long et al. 2015, 2017) are commonly used domain adaption methods whereas DALN (Chen et al. 2022a) is a state-of-the-art method in this category. Representative methods in the other category include the maximum classifier discrepancy (MCD) method (Saito et al. 2018) and the smooth domain adversarial training (SDAT) method (Rangwani et al. 2022). MCD is widely used for domain adaptation and SDAT is a state-of-the-art implicit domain adaptation method. Table 3 lists all methods compared in our evaluation. One category of existing methods – cross-domain misinformation detection method with labeled target domain information, which is not applicable to the scenario of completely unlabeled target domain information, becomes applicable to the scenarios of partially labeled target domain

Table 3 Methods Compared in Our Evaluation.

Method	Notes
DACA	Our method
MMD	Cross-domain misinformation detection method based on the MMD metric (Huang et al. 2021)
CANMD	Cross-domain misinformation detection method based on a variant of the MMD metric (Yue et al. 2022)
DANN	Cross-domain misinformation detection method based on the learning of domain-invariant features (Li et al. 2021)
MK-MMD	Domain adaptation method based on the multi-kernel MMD metric (Long et al. 2015)
JAN	Domain adaptation method based on the joint MMD metric (Long et al. 2017)
DALN	Domain adaptation method using the Nuclear-norm Wasserstein Discrepancy (Chen et al. 2022a)
MCD	Domain adaptation method utilizing a mini-max mechanism to align feature distributions between source and target domains (Saito et al. 2018)
SDAT	Domain adaptation method employing a smoothing mechanism to learn domain-invariant features (Rangwani et al. 2022)

information. Therefore, we included REAL-FND (Mosallanezhad et al. 2022), a representative method in this category, as an additional benchmark in these scenarios.

In each evaluation scenario, all compared methods took identical inputs and employed the same content embedding method described in Section 4.2 to represent the textual contents of these inputs. We trained our DACA method using the Adam optimizer with a learning rate of 0.0001 (Kingma and Ba 2015). The dimensions of the vectors $\mathbf{x}$ and $\mathbf{h}$ were set to 320. The hyperparamters in Equation 13 of our method were set as follows: the number of instances of source domain information m was set to 3, and the hyperparameter τ was set to 0.5. Implementation details of the benchmark methods are given in Appendix E.

5.3. Evaluation Results

Table 4 presents the average precision, recall, F1-score, and F2-score for each method in the evaluation scenario of completely unlabeled COVID domain, along with standard deviations (in parentheses), across 25 experiments.[14] The t-test results show that our DACA method significantly outperforms each benchmark method in precision, recall, F1-score, and F2-score. In particular, our method achieves an average recall of 0.745. Such performance is attained without any labeled

[14] For each method (ours or benchmark), we conducted 25 runs, each with a different random seed. To ensure a fair comparison, all methods used the same seed in each run.

Table 4 Performance Comparison between Our Method and Benchmark Methods (Zero Labeled News Articles in the COVID Domain).

Method	Precision	Recall	F1-score	F2-score
DACA	**0.716** (0.030)	**0.745** (0.049)	**0.719** (0.034)	**0.739** (0.043)
MMD	0.676** (0.044)	0.704** (0.044)	0.678** (0.049)	0.701** (0.029)
CANMD	0.655** (0.028)	0.702** (0.026)	0.661** (0.030)	0.692** (0.027)
DANN	0.639** (0.035)	0.658** (0.043)	0.650** (0.030)	0.656** (0.040)
MK-MMD	0.682* (0.040)	0.715* (0.038)	0.687** (0.041)	0.708* (0.038)
JAN	0.673** (0.069)	0.710** (0.080)	0.686** (0.084)	0.702** (0.029)
DALN	0.714 (0.063)	0.669** (0.039)	0.660** (0.052)	0.676** (0.031)
MCD	0.629** (0.020)	0.671** (0.020)	0.608** (0.049)	0.663** (0.020)
SDAT	0.675** (0.015)	0.712** (0.002)	0.682** (0.015)	0.705** (0.002)

Note: Significance levels are denoted by * and ** for 0.05 and 0.01, respectively.

COVID news in the training data, highlighting the efficacy of our method in transferring a model learned from labeled news in the entertainment and politics domains to predict the labels of news in the COVID domain. Moreover, our method respectively outperforms three state-of-the-art misinformation detection methods—MMD, CANMD, and DANN—by 5.92%, 9.31%, and 12.06% in precision, by 5.82%, 6.13%, and 13.22% in recall, by 6.05%, 8.77%, and 10.62% in F1-score, and by 5.42%, 6.79%, and 12.65% in F2-score. Additionally, the performance advantages of our method over representative domain adaptation methods—MK-MMD, JAN, DALN, MCD, and SDAT—range from 0.28% to 13.83% in precision, from 4.20% to 11.36% in recall, from 4.66% to 18.26% in F1-score, and from 4.38% to 12.65% in F2-score. We also compared the performance of our method with the benchmarks in terms of specificity. Our method significantly outperforms all benchmarks, except DALN, which achieves slightly higher specificity. Since the key methodological difference between our method and the benchmark methods lies in the mitigation of concept shift by its concept alignment module, the performance improvements achieved by our method can be largely attributed to this module. Given the huge volume of information generated at the early stage of an infodemic, such performance advantages achieved by our method could result in substantially more instances of misinformation being identified by our method, in comparison to

the benchmarks. As a result, a greater volume of misinformation could be prevented from dissemination, thereby significantly benefiting public health and society at large (Buchanan 2020, Van Der Linden 2022).

Tables 5, 6, and 7 present the average precision, recall, F1-score, and F2-score (with standard deviations in parentheses) for each method across 25 experiments, when the target COVID domain contains 50, 100, and 150 labeled news articles, respectively. As reported, our DACA method significantly outperforms each benchmark method in recall, F1-score, and F2-score, across varying numbers of labeled news articles in the target domain. It also surpasses each benchmark in precision, though the improvements in precision are not statistically significant in some cases. In addition, as the number of labeled news articles in the target domain increases, misinformation detection becomes easier, leading to improved performance across all methods. For example, our method achieves a recall of 0.745, 0.871, 0.911, and 0.928, when the target domain contains 0, 50, 100, and 150 labeled news articles, respectively. Taken together, the evaluation results demonstrate the consistent effectiveness of our method in detecting misinformation as an infodemic progresses – from its very early stage, when all infodemic information is completely unlabeled, to later stages, when partial information is labeled.[15]

[15] To examine the boundary conditions under which the performance of our method and the benchmarks converges, we compared our method with the best-performing benchmark method (MK-MMD) under three scenarios of 10%, 25%, and 50% labeled news articles in the COVID domain. The comparison results show that the recalls of the two methods converge when 50% of the COVID news articles are labeled, while their F1-scores and F2-scores become similar when 25% of the COVID news articles are labeled. However, it is unlikely that 25% or more COVID news articles would be labeled at the early stage of an infodemic.

Table 5 **Performance Comparison between Our Method and Benchmark Methods (50 Labeled News Articles in the COVID Domain).**

Method	Precision	Recall	F1-score	F2-score
DACA	**0.819** (0.015)	**0.871** (0.015)	**0.838** (0.017)	**0.860** (0.013)
MMD	0.817 (0.018)	0.832** (0.016)	0.823* (0.014)	0.829** (0.014)
CANMD	0.794** (0.032)	0.820** (0.023)	0.803** (0.034)	0.815** (0.026)
DANN	0.812 (0.030)	0.803** (0.028)	0.803** (0.019)	0.804** (0.021)
MK-MMD	0.785** (0.027)	0.811** (0.019)	0.795** (0.023)	0.805** (0.020)
JAN	0.737** (0.037)	0.767** (0.019)	0.744** (0.032)	0.760** (0.021)
DALN	0.770** (0.044)	0.831* (0.031)	0.799* (0.032)	0.817** (0.035)
MCD	0.769** (0.011)	0.787** (0.012)	0.778** (0.011)	0.783** (0.010)
SDAT	0.742** (0.008)	0.766** (0.001)	0.754** (0.009)	0.761** (0.009)
REAL-FND	0.770** (0.052)	0.798** (0.027)	0.772** (0.046)	0.791** (0.028)

Note: Significance levels are denoted by * and ** for 0.05 and 0.01, respectively.

Table 6 **Performance Comparison between Our Method and Benchmark Methods (100 Labeled News Articles in the COVID Domain).**

Method	Precision	Recall	F1-score	F2-score
DACA	**0.865** (0.018)	**0.911** (0.015)	**0.883** (0.012)	**0.902** (0.016)
MMD	0.844** (0.021)	0.868** (0.016)	0.853** (0.012)	0.863** (0.012)
CANMD	0.843** (0.019)	0.847** (0.018)	0.844** (0.018)	0.846** (0.016)
DANN	0.859 (0.024)	0.859** (0.017)	0.857** (0.015)	0.858** (0.015)
MK-MMD	0.835** (0.017)	0.854** (0.014)	0.843** (0.011)	0.850** (0.012)
JAN	0.814** (0.029)	0.839** (0.019)	0.823** (0.022)	0.834** (0.018)
DALN	0.854 (0.016)	0.882** (0.012)	0.866* (0.011)	0.876** (0.009)
MCD	0.781** (0.024)	0.878** (0.018)	0.801** (0.032)	0.857** (0.019)
SDAT	0.811** (0.012)	0.852** (0.003)	0.831** (0.011)	0.844** (0.009)
REAL-FND	0.840** (0.028)	0.835** (0.023)	0.836** (0.019)	0.836** (0.020)

Note: Significance levels are denoted by * and ** for 0.05 and 0.01, respectively.

Table 7 **Performance Comparison between Our Method and Benchmark Methods (150 Labeled News Articles in the COVID Domain).**

Method	Precision	Recall	F1-score	F2-score
DACA	**0.888** (0.014)	**0.928** (0.013)	**0.904** (0.011)	**0.919** (0.014)
MMD	0.865* (0.016)	0.892** (0.009)	0.877** (0.009)	0.887** (0.006)
CANMD	0.861* (0.017)	0.870** (0.016)	0.865** (0.014)	0.868** (0.014)
DANN	**0.888** (0.014)	0.883** (0.017)	0.884* (0.011)	0.884** (0.014)
MK-MMD	0.853** (0.021)	0.878** (0.010)	0.863** (0.012)	0.873** (0.008)
JAN	0.837** (0.017)	0.860** (0.011)	0.847* (0.011)	0.855** (0.009)
DALN	0.872** (0.014)	0.888* (0.006)	0.870* (0.010)	0.899** (0.006)
MCD	0.785** (0.018)	0.880** (0.007)	0.830** (0.015)	0.859** (0.008)
SDAT	0.836** (0.006)	0.865** (0.005)	0.850** (0.003)	0.859** (0.006)
REAL-FND	0.868** (0.030)	0.870** (0.020)	0.869** (0.024)	0.870** (0.021)

Note: Significance levels are denoted by * and ** for 0.05 and 0.01, respectively.

As robustness checks, we conducted a series of supplementary experiments, which are detailed in Appendix F. Specifically, we (i) evaluated our method on two additional infodemic datasets (Appendices F.1 and F.2), (ii) altered the target domain from Infodemic to Finance (Appendix F.3), (iii) varied the forecasting objective from information veracity to diffusion virality (Appendix F.4), and (iv) compared the performance of our method with that of LLM-based approaches for misinformation detection (Appendix F.5). Across all evaluation settings, the empirical results consistently confirm the superior performance of our method relative to the benchmark methods. Furthermore, Appendix G presents sensitivity analyses: one examining the impact of incomplete source domain data on our method's performance (Appendix G.1), another assessing the effect of source-target domain closeness on performance (Appendix G.2), and a third evaluating the performance of our method under a chronological setting (Appendix G.3).

5.4. Performance Analysis

Having demonstrated the superior performance of our method over the benchmarks, it is intriguing to delve deeper and analyze the factors contributing to its performance advantages. To this end, we conducted ablation studies to investigate the contribution of each novel design artifact of our DACA method to its performance. In particular, we focused on the key novelty of our method – the concept alignment module. As elaborated in Section 4.4, it consists of two submodules: the similarity function submodule and the concept shift mitigation submodule. The former represents instances of information in a transformed space, whereas the latter computes distances between these instances in the transformed space and mitigates concept shift. To evaluate the overall contribution of the concept alignment module, we removed this module from the DACA method and designated the resulting method as W/o_SF_CSM(Without both the Similarity Function and the Concept Shift Mitigation submodules). The performance difference between DACA and W/o_SF_CSM reveals the overall contribution of the concept alignment module to the performance of our method. To further investigate the functioning mechanism of the concept alignment module, we dropped its similarity function submodule but kept its concept shift mitigation submodule. We named the resulting method W/o_SF(Without the Similarity Function submodule). The performance difference between DACA and W/o_SF uncovers the role of the similarity function submodule in concept alignment.

Table 8 presents the average performance of DACA, W/o_SF_CSM, and W/o_SF in the evaluation scenario of completely unlabeled COVID domain, along with their respective standard deviations (in parentheses), across 25 experiments. As reported, the performance of W/o_SF_CSM is significantly inferior to that of DACA, due to the removal of the concept alignment module. More specifically, W/o_SF_CSM trails DACA with precision decreased from 0.716 to 0.650, recall decreased from 0.745 to 0.658, F1-score decreased from 0.719 to 0.650, and F2-score decreased from

Table 8 Ablation Study of DACA.

Method	Precision	Recall	F1-score	F2-score
DACA	**0.716** (0.030)	**0.745** (0.049)	**0.719** (0.034)	**0.739** (0.043)
W/o_SF_CSM	0.650** (0.047)	0.658** (0.053)	0.650** (0.047)	0.656** (0.056)
W/o_SF	0.660** (0.049)	0.663** (0.047)	0.648** (0.053)	0.662** (0.050)

Note: Significance levels are denoted by * and ** for 0.05 and 0.01, respectively.

0.739 to 0.656. which collectively demonstrate the contribution of the concept alignment module to the performance of DACA. Furthermore, W/o_SF performs significantly worse than DACA but comparable to W/o_SF_CSM. It is noted that W/o_SF has the concept shift mitigation submodule while omitting the similarity function submodule. The comparison results thus show that, to realize the benefits of the concept shift mitigation submodule, it must be coupled with the similarity function submodule. This empirical finding is in line with the theoretical analysis in Section 4.1, which defines concept shift over source domain instances and their corresponding nearest target domain instances. Therefore, to mitigate concept shift, we need a suitable similarity function that measures similarities between instances in an appropriate representation space, which is realized by the similarity function submodule. We also evaluate the relative contributions of the covariate alignment and concept alignment modules by removing each module individually from DACA. The results show that the concept alignment module contributes more to the performance of DACA than the covariate alignment module. Overall, the ablation study demonstrates the significant contribution of the concept alignment module to the performance of our method. Moreover, the ablation study shows that it is appropriate to design this module with two submodules: similarity function and concept shift mitigation.

6. Conclusion

6.1. Summary and Contributions

To contain harmful effects of an infodemic on public health, it is crucial to detect misinformation at the early stage of the infodemic. An early stage infodemic is characterized by a large volume of unlabeled information spread across various media platforms. Consequently, conventional misinformation detection methods are not suitable for this misinformation detection task because they rely on labeled information in the infodemic domain to train their models. State-of-the-art misinformation detection methods learn their models with labeled information in other domains to detect misinformation in the infodemic domain, thereby applicable to the task. The efficacy of these methods depends on their ability to mitigate both covariate shift and concept shift between the

infodemic domain and the domains from which they leverage labeled information. These methods focus on mitigating covariate shift but overlook concept shift, making them less effective for the task. In response, we propose a novel misinformation detection method that addresses both covariate shift and concept shift. Through extensive empirical evaluations with widely used datasets, we demonstrate the superior performance of our method over state-of-the-art misinformation detection methods as well as prevalent domain adaptation methods that can be tailored to solve the misinformation detection task.

Our study makes the following contributions to the extant literature. First, our study belongs to the area of computational design science research in the IS field (Rai et al. 2017, Padmanabhan et al. 2022, Fang et al. 2025). This area of research develops computational algorithms and methods to solve business and societal problems and aims at making methodological contributions to the literature, e.g., Abbasi et al. (2010), Li et al. (2017), Zhao et al. (2023). In particular, the methodological contribution of our study lies in its addressing of concept shift, in addition to covariate shift. More specifically, we theoretically show the importance of addressing concept shift and how to operationalize it. Built on the theoretical analysis, we develop a novel concept alignment module to mitigate concept shift, as described in Section 4.4. Second, given its significant social and economic impact, misinformation detection and management has attracted attention from IS scholars, e.g., Moravec et al. (2019), Wei et al. (2022), Hwang and Lee (2025). Our study adds to this stream of IS research with a novel method that is effective in detecting misinformation at the early stage of an infodemic.

6.2. Implications for Infodemic Management and Future Work

Every epidemic is accompanied by an infodemic, a phenomenon known since the Middle Ages (Zarocostas 2020). The wide dissemination of misinformation during an infodemic misleads people to dismiss health guidance and pursue unscientific treatments, resulting in substantial harm to public health and significant social and economic consequences (Bursztyn et al. 2020, Romer and Jamieson 2020). Furthermore, the pervasive reach of the Internet and social media platforms accelerates the spread of misinformation and amplifies its harmful impacts on public health and society (Zarocostas 2020). Detecting misinformation at the early stage of an infodemic discourages people from believing and sharing it, thereby preventing it from going viral (Buchanan 2020). Hence, early detection of misinformation is crucial for managing an infodemic and mitigating its adverse effects. Accordingly, a direct implication of our study is to provide an effective early misinformation detection method for infodemic management. Our method effectively overcomes obstacles to early detection of misinformation. First, vast amount of information spread during the early stage of an infodemic makes manual identification of misinformation impractical. In response,

our deep learning-based method automatically learns to differentiate misinformation from true information. Second, there is no labeled information at the early stage of an infodemic, rendering traditional misinformation detection methods inapplicable. Accordingly, our method leverages labeled information in other domains to detect misinformation in the infodemic domain.

Our study empowers infodemic management in several ways. Operators of media platforms can employ our method to effectively detect misinformation in the infodemic domain that spreads on their platforms. Subsequently, they can flag and debunk identified misinformation, which helps curb the diffusion of misinformation and alleviate its negative impacts (Pennycook et al. 2020, Wei et al. 2022). Additionally, flagged misinformation is less likely to be clicked, thereby reducing revenues for those who monetize misinformation and diminishing their incentives to create more misinformation (Pennycook and Rand 2021). Moreover, platform operators can trace misinformation identified by our method back to its producers. They could then restrict those who frequently produce misinformation from publishing on their platforms. In this regard, platforms are encouraged to set up guidelines to regulate their content producers (Hartley and Vu 2020). Platform operators can also trace misinformation detected by our method to its spreaders. Understandably, blocking individuals who disseminate a large amount of misinformation is important for combating misinformation on their platforms.

Furthermore, our study sheds light on the value of cross-domain data sharing for infodemic management. At the early stage of an infodemic, even experts have difficulty in distinguishing between misinformation and true information. Moreover, recruiting experts to verify and label information is costly (Kim et al. 2018). Cross-domain data sharing enables us to utilize labeled information in other domains to detect misinformation in the infodemic domain. Hence, it is a viable and cost-effective approach to misinformation detection and infodemic management. To implement this approach, we need to address a challenge. Information from different domains exhibits different marginal and conditional distributions, known as covariate shift and concept shift, respectively. Thus, the challenge is how to mitigate covariate shift and concept shift between the infodemic domain and source domains that provide labeled information. Our proposed DACA method is effective in tackling this challenge. To facilitate cross-domain data sharing for infodemic management, media platforms are recommended to establish data exchange mechanisms to ensure that participating platforms are properly incentivized and labeled information is securely shared.

Our study has limitations and can be extended in several directions. First, our method employs information content to detect misinformation. Future work could extend our method by incorporating information propagation patterns. Specifically, for each piece of information, a diffusion graph can be constructed in which nodes represent users and edges represent the spread the information from one user to another. Next, a Graph Neural Network (GNN) can be employed

to learn graph-level representations of these diffusion graphs. Concretely, for each diffusion graph, a GNN computes node embeddings, which are then pooled (e.g., via mean pooling) to obtain the embedding of the graph. Finally, the resulting graph embeddings are concatenated with their corresponding information content embeddings and fed into our method. Moreover, while our method utilizes user comments by concatenating them with news content, advanced approaches could be developed to more effectively capture the signals in user comments for misinformation detection. Second, although our method is designed for completely unlabeled target domains, we have empirically demonstrated how its performance evolves as labels in a target domain gradually become available. Future research could investigate semi-supervised or active learning extensions of our method for partially labeled target domains. Third, it is interesting to incorporate our misinformation detection method into a fact-checking system. This direction involves aligning our domain adaptation framework with pipelines that retrieve and verify evidence in real time, potentially improving the interpretability and accountability of the misinformation detection process. Finally, our method can be generalized to solve domain adaptation problems in other contexts. Accordingly, broader evaluations across diverse problem contexts would further assess its generalizability and robustness.

7. Acknowledgments

The authors thank the department editor, the associate editor, and three anonymous reviewers for their guidance and constructive comments, which have greatly improved the paper. X. Zhao and X. Fang are corresponding authors. X. Fang is partially supported by the Lerner College of Business and Economics Small Research Grant. X. Zhao is partially supported by the National Natural Science Foundation of China (Grant No. 72401172).

References

Abbasi A, Zhang Z, Zimbra D, Chen H, Nunamaker Jr JF (2010) Detecting fake websites: The contribution of statistical learning theory. MIS Quarterly 435–461.

Adebimpe WO, Adeyemi DH, Faremi A, Ojo JO, Efuntoye AE (2015) The relevance of the social networking media in ebola virus disease prevention and control in southwestern nigeria. The Pan African Medical Journal 22(Suppl 1).

Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. Proceedings of the 34th International Conference on Machine Learning, 214–223.

Asadi K, Misra D, Littman M (2018) Lipschitz continuity in model-based reinforcement learning. Proceedings of the 35th International Conference on Machine Learning, 264–273 (PMLR).

Atanasov P, Rescober P, Stone E, Swift SA, Servan-Schreiber E, Tetlock P, Ungar L, Mellers B (2017) Distilling the wisdom of crowds: Prediction markets vs. prediction polls. Management Science 63(3):691–706.

Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW (2010) A theory of learning from different domains. Machine Learning 79:151–175.

Bhattacherjee A (2023) Can Fact-Checking Influence User Beliefs About Misinformation Claims: An Examination of Contingent Effects. MIS Quarterly 47(4):1679–1692.

Borah P, Austin E, Su Y (2022) Injecting disinfectants to kill the virus: Media literacy, information gathering sources, and the moderating role of political ideology on misperceptions about covid-19. Mass Communication and Society 1–27.

Buchanan T (2020) Why do people spread false information online? the effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation. Plos one 15(10):e0239666.

Bursztyn L, Rao A, Roth CP, Yanagizawa-Drott DH (2020) Misinformation during a pandemic. Technical report, National Bureau of Economic Research.

Chen L, Chen H, Wei Z, Jin X, Tan X, Jin Y, Chen E (2022a) Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7181–7190.

Chen Q, Zhang R, Zheng Y, Mao Y (2022b) Dual contrastive learning: Text classification via label-aware data augmentation. arXiv preprint arXiv:2201.08702 .

Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. International Conference on Machine Learning, 1597–1607 (PMLR).

Ciampaglia GL, Shiralkar P, Rocha LM, Bollen J, Menczer F, Flammini A (2015) Computational fact checking from knowledge networks. PloS one 10(6):e0128193.

Fang X, Hu PJ, Chau M, Chen H (2025) Computational design science: A critical information systems research area contributing to artificial intelligence and data science. Available at SSRN 5455094 .

Fang X, Sheng ORL, Goes P (2013) When is the right time to refresh knowledge discovered from data? Operations Research 61(1):32–44.

Freeman D, Waite F, Rosebrock L, Petit A, Causier C, East A, Jenner L, Teale AL, Carr L, Mulhall S, et al. (2022) Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in england. Psychological Medicine 52(2):251–263.

Frosst N, Papernot N, Hinton G (2019) Analyzing and improving representations with the soft nearest neighbor loss. Proceedings of the 36th International Conference on Machine Learning, 2012–2020 (PMLR).

Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, March M, Lempitsky V (2016) Domain-adversarial training of neural networks. Journal of Machine Learning Research 17(59):1–35.

Gunel B, Du J, Conneau A, Stoyanov V (2021) Supervised contrastive learning for pre-trained language model fine-tuning. 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021.

Hartley K, Vu MK (2020) Fighting fake news in the covid-19 era: policy insights from an equilibrium model. Policy Sciences 53(4):735–758.

Huang Y, Gao M, Wang J, Shu K (2021) Dafd: Domain adaptation framework for fake news detection. Proceedings of the 28th Neural Information Processing Conference, 305–316.

Hwang EH, Lee S (2025) A nudge to credible information as a countermeasure to misinformation: Evidence from twitter. Information Systems Research 36(1):621–636.

Imhoff R, Lamberty P (2020) A bioweapon or a hoax? the link between distinct conspiracy beliefs about the coronavirus disease (covid-19) outbreak and pandemic behavior. Social Psychological and Personality Science 11(8):1110–1118.

Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, Maschinot A, Liu C, Krishnan D (2020) Supervised contrastive learning. Advances in Neural Information Processing Systems 33:18661–18673.

Kim H, Papamakarios G, Mnih A (2021) The lipschitz constant of self-attention. Proceedings of the 38th International Conference on Machine Learning, 5562–5571 (PMLR).

Kim J, Tabibian B, Oh A, Schölkopf B, Gomez-Rodriguez M (2018) Leveraging the crowd to detect and reduce the spread of fake news and misinformation. Proceedings of the eleventh ACM International Conference on Web Search and Data Mining, 324–332.

Kim Y (2014) Convolutional neural networks for sentence classification. Moschitti A, Pang B, Daelemans W, eds., Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar,, 1746–1751.

Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations.

Kouw WM, Loog M (2018) An introduction to domain adaptation and transfer learning. arXiv preprint arXiv:1812.11806 .

Le-Khac PH, Healy G, Smeaton AF (2020) Contrastive representation learning: A framework and review. IEEE Access 8:193907–193934.

Lee K, Ram S (2024) Explainable Deep Learning for False Information Identification: An Argumentation Theory Approach. Information Systems Research 35(2):890–907.

Li Y, Jiang B, Shu K, Liu H (2020) Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. arXiv preprint arXiv:2011.04088 .

Li Y, Lee K, Kordzadeh N, Faber B, Fiddes C, Chen E, Shu K (2021) Multi-source domain adaptation with weak supervision for early fake news detection. Proceedings of the 2021 IEEE International Conference on Big Data (Big Data), 668–676 (IEEE).

Li Z, Fang X, Bai X, Sheng ORL (2017) Utility-based link recommendation for online social networks. Management Science 63(6):1938–1952.

Liu X, Yoo C, Xing F, Oh H, El Fakhri G, Kang JW, Woo J, et al. (2022) Deep unsupervised domain adaptation: A review of recent advances and perspectives. APSIPA Transactions on Signal and Information Processing 11(1).

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

London J, Li S, Sun H (2022) Seems Legit: An Investigation of the Assessing and Sharing of Unverifiable Messages on Online Social Networks. Information Systems Research 33(3):978–1001.

Long M, Cao Y, Wang J, Jordan M (2015) Learning transferable features with deep adaptation networks. Proceedings of the 32nd International Conference on Machine Learning, 97–105 (PMLR).

Long M, Cao Z, Wang J, Jordan MI (2018) Conditional adversarial domain adaptation. Proceedings of the 32nd International Conference on Neural Information Processing Systems, 1647–1657.

Long M, Wang J, Ding G, Sun J, Yu PS (2013) Transfer feature learning with joint distribution adaptation. Proceedings of the IEEE International Conference on Computer Vision, 2200–2207.

Long M, Zhu H, Wang J, Jordan MI (2017) Deep transfer learning with joint adaptation networks. Proceedings of the 34th International Conference on Machine Learning, 2208–2217 (PMLR).

Moravec PL, Kim A, Dennis AR (2020) Appealing to Sense and Sensibility: System 1 and System 2 Interventions for Fake News on Social Media. Information Systems Research 31(3):987–1006.
Moravec PL, Kim A, Dennis AR, Minas RK (2022) Do You Really Know if It's True? How Asking Users to Rate Stories Affects Belief in Fake News on Social Media. Information Systems Research 33(3):887–907.
Moravec PL, Minas RK, Dennis AR (2019) Fake news on social media: People believe what they want to believe when it makes no sense at all. MIS Quarterly 43(4).
Mosallanezhad A, Karami M, Shu K, Mancenido MV, Liu H (2022) Domain adaptive fake news detection via reinforcement learning. Proceedings of the ACM Web Conference 2022, 3632–3640.
Mostagir M, Siderius J (2022) Learning in a Post-Truth World. Management Science 68(4):2860–2868.
Mostagir M, Siderius J (2023) Social Inequality and the Spread of Misinformation. Management Science 69(2):968–995.
Nan Q, Cao J, Zhu Y, Wang Y, Li J (2021) Mdfend: Multi-domain fake news detection. Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 3343–3347.
Nan Q, Wang D, Zhu Y, Sheng Q, Shi Y, Cao J, Li J (2022) Improving fake news detection of influential domain via domain-and instance-level transfer. Proceedings of the 29th International Conference on Computational Linguistics, 2834–2848.
Ng KC, Ke PF, So MK, Tam KY (2023) Augmenting fake content detection in online platforms: A domain adaptive transfer learning via adversarial training approach. Production and Operations Management 32(7):2101–2122.
Oh O, Agrawal M, Rao HR (2013) Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. MIS Quarterly 407–426.
Oord Avd, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 .
Padmanabhan B, Fang X, Sahoo N, Burton-Jones A (2022) Machine learning in information systems research. MIS Quarterly 46(1):iii–xix.
Papanastasiou Y (2020) Fake news propagation and detection: A sequential model. Management Science 66(5):1826–1846.
Peng X, Bai Q, Xia X, Huang Z, Saenko K, Wang B (2019) Moment matching for multi-source domain adaptation. Proceedings of the IEEE/CVF International Conference on Computer Vision, 1406–1415.
Pennycook G, Bear A, Collins ET, Rand DG (2020) The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. Management Science 66(11):4944–4957.
Pennycook G, Rand DG (2021) The psychology of fake news. Trends in Cognitive Sciences 25(5):388–402.
Rai A, Burton-Jones A, Chen H, Gupta A, Hevner AR, Ketter W, Parsons J, Rao HR, Sarkar S, Yoo Y (2017) Editor's comments: Diversity of design science research. MIS Quarterly 41(1):iii–xviii.
Rangwani H, Aithal SK, Mishra M, Jain A, Radhakrishnan VB (2022) A closer look at smoothness in domain adversarial training. Proceedings of the 39th International Conference on Machine Learning, 18378–18399 (PMLR).
Romer D, Jamieson KH (2020) Conspiracy theories as barriers to controlling the spread of covid-19 in the us. Social Science & Medicine 263:113356.
Rostami M, Galstyan A (2023) Overcoming concept shift in domain-aware settings through consolidated internal distributions. Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 9623–9631.
Roychowdhury S, Kasa SR, Gupta K, Murthy PS, Chandra A (2023) Tackling concept shift in text classification using entailment-style modeling. NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models.
Saito K, Watanabe K, Ushiku Y, Harada T (2018) Maximum classifier discrepancy for unsupervised domain adaptation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3723–3732.
Seol S, Mejia J, Dennis AR (2024) Lying for Viewers: Commingled Partisan Falsehoods Increase Viewing and Sharing of News Media. MIS Quarterly 48(2):551–582.
Shu K, Mahudeswaran D, Wang S, Lee D, Liu H (2020) Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data 8(3):171–188.
Siering M, Koch JA, Deokar AV (2016) Detecting fraudulent behavior on crowdfunding platforms: The role of linguistic and content-based cues in static and dynamic contexts. Journal of Management Information Systems 33(2):421–455.
Tachet des Combes R, Zhao H, Wang YX, Gordon GJ (2020) Domain Adaptation with Conditional Distribution Matching and Generalized Label Shift. Advances in Neural Information Processing Systems, volume 33, 19276–19289.
Van Der Linden S (2022) Misinformation: susceptibility, spread, and interventions to immunize the public. Nature Medicine 28(3):460–467.
Vorburger P, Bernstein A (2006) Entropy-based concept shift detection. Proceedings of the Sixth International Conference on Data Mining, 1113–1118.
Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–1151.
Wei X, Zhang Z, Zhang M, Chen W, Zeng DD (2022) Combining crowd and machine intelligence to detect false news on social media. MIS Quarterly 46(2):977–1008.
Yue Z, Zeng H, Kou Z, Shang L, Wang D (2022) Contrastive domain adaptation for early misinformation detection: A case study on covid-19. Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2423–2433.
Zajonc RB (2001) Mere exposure: A gateway to the subliminal. Current Directions in Psychological Science 10(6):224–228.
Zarocostas J (2020) How to fight an infodemic. The lancet 395(10225):676.
Zhang YF, Wang J, Liang J, Zhang Z, Yu B, Wang L, Tao D, Xie X (2023) Domain-specific risk minimization for domain generalization. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 3409–3421.
Zhao X, Fang X, He J, Huang L (2023) Exploiting expert knowledge for assigning firms to industries: A novel deep learning method. MIS Quarterly 47(3):1147–1176.
Zhou X, Zafarani R (2020) A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys (CSUR) 53(5):1–40.
Zhu Y, Sheng Q, Cao J, Nan Q, Shu K, Wu M, Wang J, Zhuang F (2023) Memory-guided multi-view multi-domain fake news detection. IEEE Trans. Knowl. Data Eng. 35(7):7178–7191.

Zhu Y, Zhuang F, Wang D (2019) Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 5989–5996.

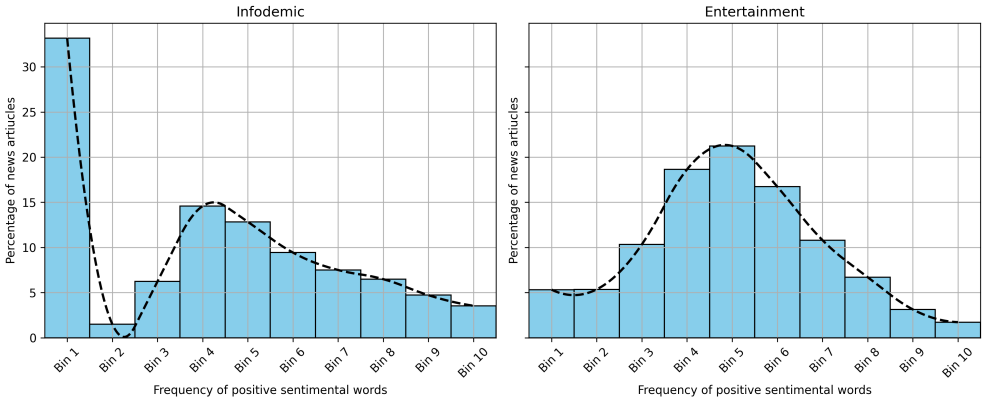
Appendix A Empirical Evidence on Data Distribution Shift

We empirically demonstrate the presence of covariate shift and concept shift between source and target domains. Let data in the source domain S follow the joint distribution $p_S(\mathbf{x}, y)$ for $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$, where $\mathcal{X}$ denotes the feature space and $\mathcal{Y}$ is the label space. Similarly, data in the target domain T are drawn from the joint distribution $p_T(\mathbf{x}, y)$. These joint distributions can be factorized as $p_S(\mathbf{x}, y) = p_S(\mathbf{x})p_S(y|\mathbf{x})$ and $p_T(\mathbf{x}, y) = p_T(\mathbf{x})p_T(y|\mathbf{x})$, respectively. To instantiate the source and target domains, we employ two real-world datasets. For the source domain (entertainment), we employed the FakeNewsNet dataset (Shu et al. 2020), which contains 16,804 true entertainment news pieces and 5,067 fake entertainment news pieces. For the target domain (infodemic), we utilized the MM-COVID dataset (Li et al. 2020), which consists of 4,750 true news pieces and 1,317 fake news pieces on COVID-19.

According to Liu et al. (2022), covariate shift refers to the difference between the feature distribution in the source domain (i.e., $p_S(\mathbf{x})$) and that in the target domain (i.e., $p_T(\mathbf{x})$). Previous studies have employed emotion features, such as the frequency of positive sentimental words and the frequency of exclamation marks in a news article, to distinguish between true and fake news (e.g., Zhang et al. 2021, Zhu et al. 2023). Accordingly, we demonstrated covariate shift between the infodemic and entertainment domains by showing their difference in the distribution of the frequency of positive sentimental words. Zhu et al. (2023) have calculated that frequency for each news article in both domains and normalized the frequencies. We further discretized the range of normalized frequencies into 10 equal-range bins, with larger bin ID indicating higher frequency of positive sentimental words. For each domain, we counted the number of new articles falling into each bin and plotted its proportion relative to the total number of news articles in that domain in Figure 1. As shown, the distribution of the frequency of positive sentimental words in the infodemic domain is skewed toward bin 1 (i.e., the lowest frequency among the 10 bins), while the distribution in the entertainment domain exhibits a bell-shaped curve. For example, according to Figure 1, news articles with the lowest frequency of positive sentimental words account for 33.18% of all infodemic news articles, while the corresponding proportion in the entertainment domain is only 5.37%.

Concept shift refers to the difference between $p_S(y|\mathbf{x})$ in the source domain and $p_T(y|\mathbf{x})$ in the target domain (Liu et al. 2022). Since both $p_S(y|\mathbf{x})$ and $p_T(y|\mathbf{x})$ condition on features $\mathbf{x}$, we employed the frequency of exclamation marks as a feature for fake news detection (Zhang et al. 2021, Zhu et al. 2023). Zhu et al. (2023) have calculated this frequency for each news article in both infodemic (target) and entertainment (source) domains and normalized the frequencies. We then discretized the range of normalized frequencies into 10 equal-range bins, with a larger bin ID indicating a higher frequency of exclamation marks. For each domain, we counted the number of *fake* new articles as well as the number of all news articles falling into each bin and plotted the percentage of *fake* new

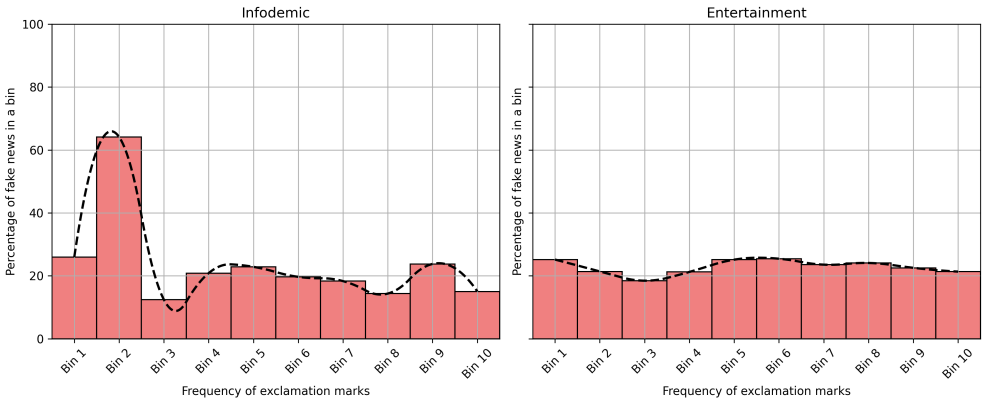
Figure 1 Covariate Shift between Infodemic and Entertainment Domains



Note. Better view in color.

articles in each bin in Figure 2. The figure illustrates concept shift between the two domains, as the label distribution conditioned on the emotion feature differs across domains. Specifically, the percentages of fake news surge in bin 2 for the infodemic domain, while they are stable across bins in the entertainment domain. The percentage of fake news in a bin generally indicates the likelihood that a news article is fake, given that the frequency of exclamation marks in the article falls into that bin. For example, when conditioned on the feature value falling into bin 2, the likelihood of an infodemic news article being fake is about 64%, while that for an entertainment news article being fake is around 21%.

Figure 2 Concept Shift between Infodemic and Entertainment Domains



Note. Better view in color.

Appendix B Method Comparison in More Dimensions

In Table 1, we compare our method against existing misinformation detection methods along five dimensions: (1) Domain, which indicates whether a method leverages single domain, multi-domain, or cross-domain data; (2) Methodological Contribution, which summarizes the core innovation of

a method; (3) Types of Input Data, which lists the types of data utilized for model training; (4) Training Temporality, which describes the timing of model training in terms of whether it requires labeled samples from the target domain; and (5) Inference Temporality, which specifies when the model is capable of making predictions, i.e., whether it can generate predictions immediately upon content release or must wait for additional signals like diffusion information or user engagement data. These dimensions were chosen to highlight the practical constraints of model deployment at different stages of an infodemic, particularly in the early phase, when time-sensitive detection using completely or partially unlabeled target domain data is critical (Li et al. 2021, Yue et al. 2022). The methods compared in the table are representative methods in each category of misinformation detection methods reviewed in Section 2. Table 1 highlights our method's advantages in real-world settings, particularly in terms of inference immediacy and minimal reliance on labeled target domain data.

Table 1: Comparison of Misinformation Detection Methods

Method	Domain	Methodological Contribution	Types of Input Data	Training Temporality	Inference Temporality
Wei et al. (2022)	Single	Leverage crowd intelligence via Bayesian result aggregation	News content (text, presence of images, videos or external links), user engagement data (human responses and reports)	Need sufficient time to label enough target domain samples	Need additional time for observing user engagement data
Sun et al. (2022)	Single	Counter noise and adversarial samples based on graph adversarial contrastive learning	News content (text), diffusion information (a tree connecting the original post with propagation events such as retweets and comments)	Need sufficient time to label enough target domain samples	Need additional time for collecting diffusion data
Zhang et al. (2024)	Single	Select relevant external knowledge subgraph with reinforcement learning, and perform modality integration with attentive hierarchical pooling	Multimodal news content (text and images), external knowledge	Need sufficient time to label enough target domain samples	Immediate prediction at the point of content release

Method	Domain	Methodological Contribution	Types of Input Data	Training Temporality	Inference Temporality
Zhu et al. (2023)	Multiple	Memory-guided multi-view and multi-domain integration based on a novel attention mechanism	News content (text, emotion features, style features)	Need sufficient time to label enough target domain samples	Immediate prediction at the point of content release
Lu et al. (2025)	Multiple	Employ domain disentanglement to extract domain invariant and specific representations from multimodal data	Multimodal news content (text and images)	Need sufficient time to label enough target domain samples	Immediate prediction at the point of content release
Mosallanezhad et al. (2022)	Cross	Introduce a reinforcement learning agent to generate domain invariant news representations	News content (text), user engagement data (comments, likes, or shares)	Need sufficient time to label enough target domain samples	Need additional time for observing user engagement data
Nan et al. (2022)	Cross	Transfer knowledge at both domain and instance levels using meta-learning and language perplexity	News content (text)	Need sufficient time to label enough target domain samples	Immediate prediction at the point of content release
Li et al. (2021)	Cross	Domain adaptation with weak-label heuristics to utilize unlabeled target data, only addressing covariate shift	News content (text)	No need to wait for target domain labels	Immediate prediction at the point of content release
Yue et al. (2022)	Cross	Contrastive domain adaptation with pseudo-labeling and label correction, only addressing covariate shift	News content (text)	No need to wait for target domain labels	Immediate prediction at the point of content release
Our Method	Cross	A novel method that mitigates both covariate shift and concept shift	News content (text, emotion, writing styles), user comments (optional)	No need to wait for target domain labels	Immediate prediction at the point of content release

Appendix C $\mathcal{H}$-divergence

$\mathcal{H}$-divergence has been widely utilized to measure distances between distributions over a feature space.

Definition A.1. ($\mathcal{H}$-divergence) (Ben-David et al. 2010). Given a feature space $\mathcal{X}$, let $\mathcal{H}$ denote a hypothesis space defined on $\mathcal{X}$ and $\mathbb{D}_S$ and $\mathbb{D}_T$ be two probability distributions over $\mathcal{X}$. For a hypothesis $h \in \mathcal{H}$, let $I(h)$ be the set such that $\mathbf{x} \in I(h)$ if and only if $h(\mathbf{x}) > \tau$, where τ is the classification threshold and an instance with features $\mathbf{x}$ is classified as 1 if $h(\mathbf{x}) > \tau$ and 0 otherwise. The $\mathcal{H}$-divergence between $\mathbb{D}_S$ and $\mathbb{D}_T$ is defined as

$$d_{\mathcal{H}}(\mathbb{D}_S, \mathbb{D}_T) = 2 \sup_{h \in \mathcal{H}} |Pr_{\mathbb{D}_S}[I(h)] - Pr_{\mathbb{D}_T}[I(h)]|, \tag{1}$$

where Pr denotes probability.[1]

The $\mathcal{H}$-divergence between two distributions can be empirically estimated using two sample datasets drawn from these distributions, respectively (Ben-David et al. 2010). Given a dataset of N_S source domain instances $\mathcal{D}_S$ and a dataset of N_T target domain instances $\mathcal{D}_T$, features of which are from distributions $\mathbb{D}_S$ and $\mathbb{D}_T$ respectively, the $\mathcal{H}$-divergence between $\mathbb{D}_S$ and $\mathbb{D}_T$ can be estimated using $\hat{d}_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ (Ben-David et al. 2010):

$$\hat{d}_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) = 2 \left(1 - \min_{h \in \mathcal{H}} \left[\frac{1}{N_S} \sum_{\mathbf{x}: h(\mathbf{x}) \leq \tau} I[\mathbf{x} \in \mathcal{D}_S] + \frac{1}{N_T} \sum_{\mathbf{x}: h(\mathbf{x}) > \tau} I[\mathbf{x} \in \mathcal{D}_T] \right] \right), \tag{2}$$

where indicator function $I[\mathbf{x} \in \mathcal{D}_S]$ is 1 if an instance with feature $\mathbf{x}$ belongs to $\mathcal{D}_S$ and $\mathcal{H}$ is a symmetric hypothesis space (i.e., for each hypothesis $h \in \mathcal{H}$, its inverse $1 - h$ is also in $\mathcal{H}$).

Appendix D Proof of Theorem and Proposition

Let $\epsilon_S(g_1, g_2) = E_{\mathbf{x} \sim \mathbb{D}_S}[|g_1(\mathbf{x}) - g_2(\mathbf{x})|]$ be the expected difference between g_1 and g_2 in the source domain, where g_1 and g_2 are two labeling functions. Similarly, $\epsilon_T(g_1, g_2) = E_{\mathbf{x} \sim \mathbb{D}_T}[|g_1(\mathbf{x}) - g_2(\mathbf{x})|]$ denotes the expected difference between g_1 and g_2 in the target domain. To prove Theorem 1, we first show an upper bound of $\epsilon_S(f_S, f_T)$:

Lemma 1 *Let f_S and f_T respectively denote the true source and target domain labeling functions, $\mathbb{D}_S$ be the source domain feature distribution, and $\mathcal{D}_S$ be a dataset with N_S samples drawn from $\mathbb{D}_S$, for any $\eta \in (0, 1)$, the following inequality holds with probability at least $1 - \eta$:*

$$\epsilon_S(f_S, f_T) = E_{\mathbf{x} \sim \mathbb{D}_S}[|f_S(\mathbf{x}) - f_T(\mathbf{x})|] \leq \frac{1}{N_S} \sum_{i=1}^{N_S} |f_S(\mathbf{x}_i^S) - f_T(\mathbf{x}_i^S)| + \sqrt{-\frac{\ln(\frac{\eta}{2})}{2N_S}}, \tag{3}$$

where $\mathbf{x}_i^S$ is the i-th sample drawn i.i.d. from $\mathbb{D}_S$.

[1] Ben-David et al. (2010) consider a hypothesis h that produces a discrete output, i.e., 0 or 1, while we consider a hypothesis h that predicts a continuous output between 0 and 1.

Proof. This proof relies on McDiarmid's inequality (Doob 1940). For notation convenience, define $\mathbf{z} = f_S(\mathbf{x}) - f_T(\mathbf{x})$ *where* $\mathbf{x} \sim \mathbb{D}_S$. *We then have* $\mathbf{z}_i = f_S(\mathbf{x}_i^S) - f_T(\mathbf{x}_i^S), i = 1, 2, \cdots, N_s$. *Further define* $g(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N_S}) = \frac{1}{N_S}\sum_{i=1}^{N_S} |\mathbf{z}_i|$. *Function* g *satisfies the property that*

$$|g(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N_S}) - g(\mathbf{z}_1, \ldots, \mathbf{z}_{k-1}, \mathbf{z}'_k, \mathbf{z}_{k+1}, \ldots, \mathbf{z}_{N_S})| \leq \frac{1}{N_S}, \tag{4}$$

which means that by changing any k*-th coordinate, the function will change at most* $\frac{1}{N_S}$. *This is because* $|\mathbf{z}_k| = |f_S(\mathbf{x}_k^S) - f_T(\mathbf{x}_k^S)| \in [0, 1]$.

Given that $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N_S}$ *are i.i.d. random samples, for any* $\epsilon > 0$, *by McDiarmid's inequality*

$$\begin{aligned} p(|g(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N_S}) - E[g(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N_S})]| \geq \epsilon) &\leq 2\exp(\frac{-2\epsilon^2}{\sum_{k=1}^{N_S}\frac{1}{N_S^2}}) \\ &= 2\exp(-2N_S\epsilon^2). \end{aligned} \tag{5}$$

Equivalently,

$$p(|g(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N_S}) - E[g(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N_S})]| \leq \epsilon) \geq 1 - 2\exp(-2N_S\epsilon^2). \tag{6}$$

By replacing ϵ *with* $\sqrt{-\frac{\ln(\frac{\eta}{2})}{2N_S}}$ *where* $\eta \in (0, 1)$, *we have*

$$p\left(|g(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N_S}) - E[g(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N_S})]| \leq \sqrt{-\frac{\ln(\frac{\eta}{2})}{2N_S}}\right) \geq 1 - \eta, \tag{7}$$

Note that $E[g(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N_S})] = E_{\mathbf{x}\sim\mathbb{D}_S}[|f_S(\mathbf{x}) - f_T(\mathbf{x})|]$ *by construction, while* $g(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N_S}) = \frac{1}{N_S}\sum_{i=1}^{N_S} |f_S(\mathbf{x}_i^S) - f_T(\mathbf{x}_i^S)|$ *by definition. Based on the inequality* $a - b \leq |b - a|$, *Inequality 7 means that Inequality 3 must hold with probability at least* $1 - \eta$, *which concludes our proof of Lemma 1.*

Next, we show, in Lemma 2, that the population $\mathcal{H}$-divergence is bounded by the empirical $\mathcal{H}$-divergence.

Lemma 2 *Let* $\mathcal{H}$ *be a hypothesis space with VC dimension* d. $\mathcal{D}_S$ *and* $\mathcal{D}_T$ *are datasets of instances drawn from distributions* $\mathbb{D}_S$ *and* $\mathbb{D}_T$, *respectively. Let* $d_{\mathcal{H}}(\mathbb{D}_S, \mathbb{D}_T)$ *denote the population* $\mathcal{H}$*-divergence between distributions* $\mathbb{D}_S$ *and* $\mathbb{D}_T$, *and* $\hat{d}_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ *be its empirical estimation from datasets* $\mathcal{D}_S$ *and* $\mathcal{D}_T$. *For any* $\eta \in (0, 1)$, *with probability at least* $1 - \eta$,

$$d_{\mathcal{H}}(\mathbb{D}_S, \mathbb{D}_T) \leq \hat{d}_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + 4\max\{\sqrt{\frac{d\ln(2N_S) + \ln(\frac{2}{\eta})}{N_S}}, \sqrt{\frac{d\ln(2N_T) + \ln(\frac{2}{\eta})}{N_T}}\}, \tag{8}$$

where N_S *is the number of instances in* $\mathcal{D}_S$ *and* N_T *is the number of instances in* $\mathcal{D}_T$.

Proof. According to Kifer et al. (2004), we have

$$P\left[\left|d_{\mathcal{H}}(\mathbb{D}_S, \mathbb{D}_T) - \hat{d}_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)\right| > \epsilon\right] \leq (2N_S)^d e^{-N_S\epsilon^2/16} + (2N_T)^d e^{-N_T\epsilon^2/16}. \tag{9}$$

By replacing $(2N_S)^d e^{-N_S\epsilon^2/16} + (2N_T)^d e^{-N_T\epsilon^2/16}$ *with* η, *we get*

$$\eta \leq 2\max\{(2N_S)^d e^{-N_S\epsilon^2/16}, (2N_T)^d e^{-N_T\epsilon^2/16}\}, \tag{10}$$

which means that $\eta \leq 2\left(2N_S\right)^d e^{-N_S\epsilon^2/16}$ *as well as* $\eta \leq 2\left(2N_T\right)^d e^{-N_S\epsilon^2/16}$. *In other words, we have* $\epsilon \leq 4\sqrt{(d\ln(2N_S)+\ln(2/\eta))/N_S}$ *as well as* $\epsilon \leq 4\sqrt{(d\ln(2N_T)+\ln(2/\eta))/N_T}$. *Because increasing* ϵ *will decrease the probability* $P\left[\left|d_{\mathcal{H}}\left(\mathbb{D}_S,\mathbb{D}_T\right)-\hat{d}_{\mathcal{H}}\left(\mathcal{D}_S,\mathcal{D}_T\right)\right|>\epsilon\right]$, *for*

$$\epsilon_{\max}=4\max\{\sqrt{\frac{d\ln(2N_S)+\ln(\frac{2}{\eta})}{N_S}},\sqrt{\frac{d\ln(2N_T)+\ln(\frac{2}{\eta})}{N_T}}\}, \tag{11}$$

we have

$$\begin{aligned}\eta &\geq P\left[\left|d_{\mathcal{H}}\left(\mathbb{D}_S,\mathbb{D}_T\right)-\hat{d}_{\mathcal{H}}\left(\mathcal{D}_S,\mathcal{D}_T\right)\right|>\epsilon\right]\\ &\geq P\left[\left|d_{\mathcal{H}}\left(\mathbb{D}_S,\mathbb{D}_T\right)-\hat{d}_{\mathcal{H}}\left(\mathcal{D}_S,\mathcal{D}_T\right)\right|>\epsilon_{\max}\right]\end{aligned} \tag{12}$$

which concludes the proof of Lemma 2. If $N_S=N_T$, *Lemma 2 will reduce to Lemma 1 in Ben-David et al. (2010).*

Definition A.2. (The ideal hypothesis). The ideal hypothesis h^* is defined as the hypothesis that minimizes the summation of the expected differences between $h\in\mathcal{H}$ and the target domain labeling function f_T in both source and target domains

$$h^*=\arg\min_{h\in\mathcal{H}}\epsilon_S(h,f_T)+\epsilon_T(h,f_T), \tag{13}$$

where $\epsilon_S(h,f_T)=E_{\mathbf{x}\sim\mathbb{D}_S}[|h(\mathbf{x})-f_T(\mathbf{x})|]$ and $\epsilon_T(h,f_T)=E_{\mathbf{x}\sim\mathbb{D}_T}[|h(\mathbf{x})-f_T(\mathbf{x})|]$.

The error incurred by the ideal hypothesis is $\lambda=\epsilon_S(h^*,f_T)+\epsilon_T(h^*,f_T)$. Now, we provide the proof of Theorem 1 utilizing the lemmas established above.

Theorem 1 *For a source domain instance with features* $\boldsymbol{x}_i^S$, *let* $c(\boldsymbol{x}_i^S)$ *be the features of its nearest target domain instance, where* $i=1,2,\ldots,N_S$ *and the distance between a pair of instances are measured over their feature space. If* f_T *is L-Lipschitz continuous, then for any* $\eta\in(0,1)$, *with probability at least* $(1-\eta)^2$,

$$\epsilon_T(h)\leq\epsilon_S(h)+\hat{d}_{\mathcal{H}}(\mathcal{D}_S,\mathcal{D}_T)+\frac{1}{N_S}\sum_{i=1}^{N_S}|f_S(\mathbf{x}_i^S)-f_T(c(\mathbf{x}_i^S))|+\frac{L}{N_S}\sum_{i=1}^{N_S}||\mathbf{x}_i^S-c(\mathbf{x}_i^S)||+C_1, \tag{14}$$

where L *is the Lipschitz constraint constant,* $\hat{d}_{\mathcal{H}}(\mathcal{D}_S,\mathcal{D}_T)$ *denotes the empirical* $\mathcal{H}$*-divergence between distributions* $\mathbb{D}_S$ *and* $\mathbb{D}_T$, $||\boldsymbol{x}_i^S-c(\boldsymbol{x}_i^S)||$ *represents the distance between* $\boldsymbol{x}_i^S$ *and* $c(\boldsymbol{x}_i^S)$, *and* $C_1=\lambda+\sqrt{-\frac{\ln(\frac{\eta}{2})}{2N_S}}+2\max\{\sqrt{\frac{d\ln(2N_S)+\ln(\frac{2}{\eta})}{N_S}},\sqrt{\frac{d\ln(2N_T)+\ln(\frac{2}{\eta})}{N_T}}\}$.

Proof. With probability at least $(1-\eta)^2$,

$$\begin{aligned}
\epsilon_T(h) &\le \epsilon_T(h^*) + \epsilon_S(h) + \epsilon_S(h^*) + \frac{1}{2} d_{\mathcal{H}}(\mathbb{D}_S, \mathbb{D}_T) \\
&\le \epsilon_T(h^*) + \epsilon_S(h) + \epsilon_S(h^*, f_T) + \epsilon_S(f_S, f_T) + \frac{1}{2} d_{\mathcal{H}}(\mathbb{D}_S, \mathbb{D}_T) \\
&= \epsilon_S(h) + \frac{1}{2} d_{\mathcal{H}}(\mathbb{D}_S, \mathbb{D}_T) + \epsilon_S(f_S, f_T) + \lambda \\
&\le \epsilon_S(h) + \frac{1}{2} \hat{d}_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \frac{1}{N_S} \sum_{i=1}^{N_S} |f_S(\mathbf{x}_i^S) - f_T(\mathbf{x}_i^S)| + C_1 \\
&= \epsilon_S(h) + \frac{1}{2} \hat{d}_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \frac{1}{N_S} \sum_{i=1}^{N_S} |f_S(\mathbf{x}_i^S) + f_T(c(\mathbf{x}_i^S)) - f_T(c(\mathbf{x}_i^S)) - f_T(\mathbf{x}_i^S)| + C_1 \\
&\le \epsilon_S(h) + \frac{1}{2} \hat{d}_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \frac{1}{N_S} \sum_{i=1}^{N_S} |f_S(\mathbf{x}_i^S) - f_T(c(\mathbf{x}_i^S))| + \frac{1}{N_S} \sum_{i=1}^{N_S} |f_T(c(\mathbf{x}_i^S)) - f_T(\mathbf{x}_i^S)| + C_1 \\
&\le \epsilon_S(h) + \frac{1}{2} \hat{d}_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \frac{1}{N_S} \sum_{i=1}^{N_S} |f_S(\mathbf{x}_i^S) - f_T(c(\mathbf{x}_i^S))| + \frac{L}{N_S} \sum_{i=1}^{N_S} ||\mathbf{x}_i^S - c(\mathbf{x}_i^S)|| + C_1.
\end{aligned} \tag{15}$$

The first step is based on the proof of Theorem 2 in Ben-David et al. (2010). The second step is followed by triangle inequality. The third step utilizes the definition $\lambda = \epsilon_S(h^*, f_T) + \epsilon_T(h^*, f_T)$ *and the definition* $\epsilon_T(h^*) = \epsilon_T(h^*, f_T)$. *The fourth step employs Lemma 1 and Lemma 2 to respectively expand the terms* $\epsilon_S(f_S, f_T)$ *and* $d_{\mathcal{H}}(\mathbb{D}_S, \mathbb{D}_T)$ *while collecting the constants into* C_1. *Because the inequalities given by Lemma 1 and Lemma 2 hold both with at least probability* $1-\eta$, *the inequality at the second step holds with at least probability* $(1-\eta)^2$ *under the independence assumption. The fifth step adds and subtracts* $f_T(c(\boldsymbol{x}_i^S))$. *The sixth step utilizes triangle inequality. The last step is from the assumption of L-Lipschitz continuous for* f_T.

By extending Theorem 1 to the case of multiple source domains, we have the following proposition.

Proposition 1 *For a source domain instance with features* $\boldsymbol{x}_i^{S_j}$ *in* $\mathcal{D}_{S_j}$, *let* $c(\boldsymbol{x}_i^{S_j})$ *be the features of its nearest instance in* $\mathcal{D}_T$, *where* $j = 1, 2, \ldots, k$ *and* $i = 1, 2, \ldots, N_{S_j}$. *If* f_T *is L-Lipschitz continuous, then for any* $\eta \in (0,1)$, *with probability at least* $(1-\eta)^{2k}$,

$$\epsilon_T(h) \le \frac{1}{k} \sum_{j=1}^{k} \{\epsilon_{S_j}(h) + \frac{1}{2} \hat{d}_{\mathcal{H}}(\mathcal{D}_{S_j}, \mathcal{D}_T) + \frac{1}{N_{S_j}} \sum_{i=1}^{N_{S_j}} |f_{S_j}(\mathbf{x}_i^{S_j}) - f_T(\mathbf{x}_i^{S_j,T})| + \frac{L}{N_{S_j}} \sum_{i=1}^{N_{S_j}} ||\mathbf{x}_i^{S_j} - \mathbf{x}_i^{S_j,T}|| + C_j\}, \tag{16}$$

where $C_j = \lambda_j + \sqrt{-\frac{\ln(\frac{\eta}{2})}{2N_{S_j}}} + 2\max\{\sqrt{\frac{d\ln(2N_{S_j})+\ln(\frac{2}{\eta})}{N_{S_j}}}, \sqrt{\frac{d\ln(2N_T)+\ln(\frac{2}{\eta})}{N_T}}\}$, λ_j *is the error incurred by the ideal hypothesis* $\lambda_j = \epsilon_{S_j}(h^*, f_T) + \epsilon_T(h^*, f_T)$, *and* L *is the constant for the Lipschitz condition.*

Proof. By Theorem 1, for any source domain $j = 1, 2, \cdots, k$, *the following inequality holds with probability at least* $(1-\eta)^2$ *for any* $\eta \in (0,1)$,

$$\epsilon_T(h) \le \epsilon_{S_j}(h) + \hat{d}_{\mathcal{H}}(\mathcal{D}_{S_j}, \mathcal{D}_T) + \frac{1}{N_{S_j}} \sum_{i=1}^{N_{S_j}} |f_{S_j}(\mathbf{x}_i^{S_j}) - f_T(c(\mathbf{x}_i^{S_j}))| + \frac{L}{N_{S_j}} \sum_{i=1}^{N_{S_j}} ||\mathbf{x}_i^{S_j} - c(\mathbf{x}_i^{S_j})|| + C_j. \tag{17}$$

Proposition 1 follows immediately by summing on each side of inequality (A17) over all source domains.

Appendix E Implementation Details of Benchmark Methods

MMD. Maximum Mean Discrepancy (MMD) is a metric to measure the discrepancy between distributions in a reproducing kernel Hilbert space (RKHS) (Borgwardt et al. 2006). Specifically, a kernel function k is associated with a feature map ϕ, i.e., $k(\mathbf{x}_S, \mathbf{x}_T) =< \phi(\mathbf{x}_S), \phi(\mathbf{x}_T) >$, where $\mathbf{x}_S$ and $\mathbf{x}_T$ are source and target distributions. Let $\mathcal{H}$ be an RKHS, and MMD is defined as the discrepancy between the mean embeddings of source and target distributions in $\mathcal{H}$:

$$\text{MMD} = d(p_S, p_T)^2 = ||\frac{1}{N_S}\phi(\mathbf{x}_S) - \frac{1}{N_T}\phi(\mathbf{x}_T)||^2_{\mathcal{H}}. \tag{18}$$

As reviewed in Section 2.3, MMD is an explicit metric that measures the discrepancy between source and target domains in terms of their feature distributions. In our implementation, we utilized the Gaussian kernel $k(\mathbf{x}_S, \mathbf{x}_T) = e^{||\mathbf{x}_S - \mathbf{x}_T||/\gamma}$ and set $\gamma = 0.5$.

CANMD. Contrastive Adaptation Network for Early Misinformation Detection (CANMD) is a domain adaptation method devised for misinformation detection (Yue et al. 2022). This method assigns pseudo labels to target domain instances. Let D_{ij} denote the multi-kernel MMD discrepancy between the covariate distribution of source domain instances with label i and that of target domain instances with pseudo label j, where $i, j \in \{0, 1\}$. To minimize the divergence between source and target domains, the loss is defined as

$$\mathcal{L} = D_{00} + D_{11} - \frac{1}{2}(D_{01} + D_{10}). \tag{19}$$

Following Yue et al. (2022), we implemented the multi-kernel MMD using 5 Gaussian kernels with $k(\mathbf{x}_S, \mathbf{x}_T) = e^{||\mathbf{x}_S - \mathbf{x}_T||/\gamma}$ and $\gamma = 2^k$, where the value of k for each Gaussian kernel was set to -3, -2, -1, 0, and 1, respectively. The Gaussian kernels were combined with equal weights.

DANN. The Domain Adversarial Neural Networks (DANN) method has two classifiers: a label classifier and a domain classifier (Ganin et al. 2016, Li et al. 2021). The former is learned to classify each piece of source or target domain information as true or false. The latter is used to learn domain-invariant features. The hyperparameter λ of DANN controls the relative weight between its domain and label classifiers. In our evaluation, we performed a grid search for the value of λ from 1 to 10, with an interval of 1. We found that $\lambda = 5$ yielded the best performance for DANN and chose $\lambda = 5$ accordingly.

MK-MMD. Multi-kernel MMD (MK-MMD) is an extension of single-kernel MMD (Long et al. 2015). Instead of using a single kernel function, MK-MMD employs a convex combination of multiple kernel functions. In our implementation, we used Gaussian kernels $k(\mathbf{x}_S, \mathbf{x}_T) = e^{||\mathbf{x}_S - \mathbf{x}_T||/\gamma}$ and $\gamma = 2^k$,

where the value of k for each Gaussian kernel was set to -3, -2, -1, 0, and 1, respectively. The Gaussian kernels were combined with equal weights.

JAN. Joint Adaptation Networks (JAN) measures the discrepancy between source and target domains at each layer of its classifier using MK-MMD (Long et al. 2017). It then combines these discrepancies and minimizes the aggregated discrepancy. The implementation of MK-MMD in JAN is the same as the implementation of MK-MMD described above.

DALN. Discriminator-free Adversarial Learning Network (DALN) utilizes the Nuclear-norm Wasserstein Discrepancy (NWD) to measure the covariate shift (Chen et al. 2022). To compute NWD, Chen et al. (2022) devise a domain classifier that predicts the probability of an instance belonging to the source domain. NWD is then defined as the difference between the average probability of source domain instances and that of target domain instances. The DALN method minimizes the covariate shift by minimizing NWD. In addition, the DALN method has a label classifier that predicts the label (true or false) for each instance. The loss of the method is a summation of NWD and the label classification loss.

MCD. The Maximum Classifier Discrepancy (MCD) method consists of three steps (Saito et al. 2018). In the first step, it trains two classifiers using labeled source domain instances. Next, it maximizes the discrepancy between two classifiers when applied to target domain instances. In the third step, it adjusts its feature extractor to minimize such discrepancy. MCD can be viewed as a generative adversarial network (GAN), with the feature extractor serving as the generator and the two classifiers functioning as the discriminator.

SDAT. Smooth Domain Adversarial Training (SDAT) is an implicit domain adaptation method based on DANN (Rangwani et al. 2022). Inspired by the observation that the label classifier of DANN often falls into sharp minima, Rangwani et al. (2022) devise a smooth training loss based on the Sharpness Aware Minimization (SAM) proposed by Foret et al. (2021). In our implementation, we set the neighborhood size ρ in SAM to 0.05 by following (Foret et al. 2021).

REAL-FND. REinforced Adaptive Learning Fake News Detection (REAL-FND) first pretrains a domain classifier and a label classifier (Mosallanezhad et al. 2022). Then, reinforcement learning is applied on content embeddings to maximize the confidence of the label classifier and adversary confidence of the domain classifier. Following Mosallanezhad et al. (2022), the reinforcement learning agent employs a simple feed-forward network with 2 layers. Each episode of the agent contains 20 actions to adjust the embedding.

Appendix F Robustness Analysis

F.1 The Second Infodemic Dataset: Detecting Misinformation in Chinese News

We conducted an additional empirical evaluation using a public dataset of Chinese news (Nan et al. 2021), which has been widely adopted in recent misinformation studies (e.g., Wu et al. 2023, Zhu et al.

2023, Peng et al. 2024). The dataset contains true and fake news articles in nine different domains. The fake news articles in the dataset were labeled by the Weibo Community Management Center, whereas the true news articles were verified by NewsVerify, a platform dedicated to evaluating the credibility of content on Weibo. We employed news articles in three of these domains: entertainment, politics, and health, consistent with the setting of the empirical study with English news datasets. Not all news articles in the health domain are related to infodemic. We specifically chose news articles from this domain that contain words associated with the COVID pandemic, such as COVID, COVID test, or pandemic. Table 2 lists the number of true and fake news articles in each domain. User comments accompany 53.22% of COVID news, 84.27% of politics news, and 94.64% of entertainment news. The average comment length is 16.4 words in the COVID domain, 14.1 words in the entertainment domain, and 18.8 words in the politics domain.

Table 2 Summary Statistics of Chinese News Dataset (Three Domains).

Domain	Number of true news	Number of fake news
COVID (Infodemic / Target)	201	124
Entertainment (Source)	1000	440
Politics (Source)	306	546

We followed the evaluation procedure described in Section 5.2, except that Chinese words in the dataset were embedded using the pretrained BERT-wwm model (Cui et al. 2021). Table 3 compares the performance between our method and each benchmark method in evaluation scenario 1. As reported, our method significantly outperforms each benchmark, improving precision by a range of 3.53% to 15.22%, recall by a range of 3.88% to 25.33%, F1-score by a range of 3.33% to 25.12%, and F2-score by a range of 3.87% to 22.50%.

F.2 The Third Infodemic Dataset: Detecting Misinformation in Social Media Posts

We employed the Constraint dataset (Patwa et al. 2021) as the third COVID infodemic dataset. This dataset comprises 10,700 social media posts related to COVID-19, with an average length of 27.05 English words per post. Among these posts, 5,100 posts are labeled as fake while 5,600 as true. The fake posts are collected from fact-checking websites whereas the true posts are gathered from verified Twitter accounts (Patwa et al. 2021). Considering that social media posts are typically much shorter than news articles, we selected two source domain datasets — PHEME (Zubiaga et al. 2016) and FEVER (Thorne et al. 2018) — whose document lengths are comparable to that of the Constraint dataset. The PHEME dataset is a collection of Twitter posts related to breaking news events, including 48,619 true and 13,824 fake posts, annotated by a team of journalists. The

Table 3 Performance Comparison between Our Method and Benchmark Methods using the Chinese News Dataset

Method	Precision	Recall	F1-score	F2-score
DACA	**0.822** (0.010)	**0.836** (0.004)	**0.824** (0.010)	**0.833** (0.017)
MMD	0.794** (0.001)	0.804** (0.006)	0.797** (0.001)	0.802** (0.013)
CANMD	0.788** (0.011)	0.792** (0.013)	0.789** (0.012)	0.791** (0.024)
DANN	0.716** (0.007)	0.713** (0.002)	0.709** (0.001)	0.713** (0.031)
MK-MMD	0.778** (0.010)	0.786** (0.006)	0.780** (0.009)	0.784** (0.017)
JAN	0.717** (0.023)	0.723** (0.014)	0.718** (0.023)	0.722** (0.034)
DALN	0.794** (0.007)	0.764** (0.013)	0.763** (0.003)	0.769** (0.034)
MCD	0.739** (0.044)	0.667** (0.074)	0.658** (0.082)	0.680** (0.056)
SDAT	0.771** (0.012)	0.777** (0.002)	0.772** (0.008)	0.776** (0.017)

Note: Significance levels are denoted by * and ** for 0.05 and 0.01, respectively.

average post length is 21.5 words. The FEVER dataset consists of 80,035 true claims and 29,775 false claims, all human-labeled, with an average claim length of 9.4 words.[2] We followed the experimental procedure described in Section 5.2 to evaluate the performance of our method and the benchmark methods. Table 4 presents the average precision, recall, F1-score, and F2-score for each method, along with standard deviations (in parentheses), across 25 experiments. As reported, our DACA method significantly outperforms each benchmark method across all evaluation metrics. For example, DACA achieves an average recall of 0.758, surpassing that of the best-performing benchmark method (MCD) by 6.61%.

F.3 Alternative Target Domain: Detecting Financial Misinformation

We consider the diffusion of true and false information in the finance domain and target the task of detecting financial misinformation, which is an emerging but increasingly important research area with substantial implications for market stability and investor protection (Clarke et al. 2021, Rangapur et al. 2023, Liu et al. 2025). By applying our proposed method to this domain, we aim to show that the concept alignment mechanism can effectively extract domain-invariant misinformation signals for other contexts as well.

[2] We used a preprocessed version of the PHEME dataset, which can be accessed at `https://www.kaggle.com/datasets/nicolemichelle/pheme-dataset-for-rumour-detection/data`.

Table 4 **Performance Comparison between Our Method and Benchmark Methods using the Constraint Dataset.**

Method	Precision	Recall	F1-score	F2-score
DACA	**0.768** (0.041)	**0.758** (0.042)	**0.757** (0.044)	**0.760** (0.042)
MMD	0.620** (0.042)	0.606** (0.029)	0.597** (0.036)	0.608** (0.030)
CANMD	0.710** (0.039)	0.700** (0.037)	0.698** (0.038)	0.703** (0.037)
DANN	0.712** (0.046)	0.704** (0.045)	0.702** (0.047)	0.705** (0.045)
MK-MMD	0.608** (0.041)	0.598** (0.038)	0.590** (0.041)	0.600** (0.038)
JAN	0.644** (0.047)	0.631** (0.038)	0.622** (0.040)	0.633** (0.063)
DALN	0.681** (0.041)	0.664** (0.040)	0.657** (0.044)	0.667** (0.040)
MCD	0.734* (0.049)	0.711** (0.061)	0.704** (0.078)	0.715** (0.058)
SDAT	0.663** (0.040)	0.649** (0.046)	0.640** (0.057)	0.652** (0.045)

Note: Significance levels are denoted by * and ** for 0.05 and 0.01, respectively.

In this evaluation, we utilized the Chinese news dataset (Nan et al. 2021), details of which are summarized in Appendix F.1. We employed the politics and entertainment domains in this dataset as the source domains, and the finance domain as the target domain. Table 5 lists the number of true and fake news in each domain. We followed the experimental procedure described in Section 5.2 to evaluate the performance of our method and the benchmark methods. Table 6 presents the average precision, recall, F1-score, and F2-score for each method, along with standard deviations (in parentheses), across 25 experiments. As reported, our DACA method significantly outperforms each benchmark method across all metrics. This additional evaluation demonstrates the effectiveness of our method in a context beyond COVID.

Table 5 **Summary Statistics of Chinese News Dataset (Financial Misinformation Detection).**

Domain	Number of true news	Number of fake news
Finance (Target)	959	362
Entertainment (Source)	1000	440
Politics (Source)	306	546

F.4 Beyond Binary Classification: Forecasting the Number of News Comments

We expand the scope of our study by extending our proposed method to a new task. Specifically, the new task is to predict the number of comments a piece of information (e.g., a news article) will

Table 6 Performance Comparison between Our Method and Benchmark Methods on Financial Misinformation Detection.

Method	Precision	Recall	F1-score	F2-score
DACA	**0.804** (0.034)	**0.823** (0.027)	**0.810** (0.030)	**0.819** (0.027)
MMD	0.770** (0.034)	0.786** (0.031)	0.775** (0.033)	0.783* (0.031)
CANMD	0.745** (0.038)	0.786** (0.035)	0.755** (0.039)	0.777** (0.036)
DANN	0.714** (0.045)	0.694** (0.065)	0.690** (0.057)	0.697** (0.058)
MK-MMD	0.762** (0.042)	0.782** (0.037)	0.769** (0.041)	0.778** (0.038)
JAN	0.743** (0.041)	0.776** (0.041)	0.752** (0.044)	0.769** (0.040)
DALN	0.774* (0.034)	0.761** (0.036)	0.762** (0.034)	0.763** (0.032)
MCD	0.740** (0.025)	0.786** (0.021)	0.737** (0.040)	0.777** (0.022)
SDAT	0.689** (0.044)	0.708** (0.041)	0.700** (0.059)	0.704** (0.041)

Note: Significance levels are denoted by * and ** for 0.05 and 0.01, respectively.

receive. Understandably, this new task is fundamentally different from the original misinformation detection task, as the former is a regression problem while the latter is a binary classification task. Moreover, the number of comments a piece of information receives is a natural proxy for its diffusion and virality, as a high (or low) comment count typically indicates broad (or limited) diffusion and strong (or weak) audience engagement. Solving both tasks in tandem enables more refined infodemic management. While the original task focuses on detecting misinformation, the new task assesses the potential impact of a piece of information by predicting its comment count. Solving these two tasks together helps platforms prioritize their resources toward combating misinformation with high virality potential, as such content more likely leads to severe consequences.

We now formally define the new task. Let $\mathcal{D}_{S_j}$ denote a dataset of N_{S_j} pieces of labeled information in the source domain S_j, $j = 1, 2, \ldots, k$. Each piece of labeled information in $\mathcal{D}_{S_j}$ is represented as $(\mathbf{x}_i^{S_j}, y_i^{S_j})$, where $\mathbf{x}_i^{S_j}$ denotes features extracted from the information, label $y_i^{S_j}$ represents the number of comments associated with the information, and $i = 1, 2, \ldots, N_{S_j}$. Let $\mathcal{D}_T$ denote a dataset of N_T pieces of unlabeled information in the infodemic domain T. Each piece of unlabeled information in $\mathcal{D}_T$ is represented as $\mathbf{x}_i^T$, i.e., features extracted from the information, $i = 1, 2, \ldots, N_T$. The new task is defined as follows:

Given a dataset $\mathcal{D}_T$ of N_T pieces of unlabeled information in the infodemic domain and k datasets of labeled information in various source domains, $\mathcal{D}_{S_1}, \mathcal{D}_{S_2}, \ldots, \mathcal{D}_{S_k}$, where $\mathcal{D}_{S_j}$ consists of N_{S_j} pieces

of labeled information, $j = 1, 2, \ldots, k$, the objective is to learn a model from the data that predicts the number of comments each piece of information in $\mathcal{D}_T$ will receive.

We modified our DACA method for the new task. First, we replaced the classification loss with the mean squared error loss to accommodate the training of a regression task. Second, to learn the similarity function, we calculated the median number of comments received among all labeled source domain information and divided them into two groups: one containing information receiving above-median comment counts and the other with below-median counts. Pairs from the same group were labeled as positive, while pairs from different groups were labeled as negative. For each benchmark method, we also replaced the classification loss with the mean square error loss.[3] We measured the performance of each method on the new task using standard metrics for regression models (Bandari et al. 2012): the root mean square error (RMSE) and the mean absolute error (MAE). Let y_i^T denote the actual number of comments received by a piece of infodemic information, and let $\hat{y}_i^T$ be the corresponding predicted number, where $i = 1, 2, \ldots, N_T$. RMSE is computed as

$$\text{RMSE} = \sqrt{\sum_{i=1}^{N_T} \frac{(y_i^T - \hat{y}_i^T)^2}{N_T}}, \tag{20}$$

and MAE is formulated as

$$\text{MAE} = \sum_{i=1}^{N_T} \frac{|y_i^T - \hat{y}_i^T|}{N_T}. \tag{21}$$

Lower RMSE and MAE indicate better predictive performance.

We employed the Chinese news dataset (Nan et al. 2021), details of which are summarized in Appendix F.1. We utilized the politics and entertainment domains in this dataset as the source domains, and the COVID domain as the infodemic domain. In the politics domain, the average and maximum number of comments received per piece of information are 96.31 and 1,554, respectively; in the entertainment domain, 170.84 and 1,761; and in the COVID domain, 10.01 and 473. We followed the experimental procedure described in Section 5.2 to evaluate the performance of our method and the benchmark methods. Table 7 presents the average RMSE and MAE for each method, along with standard deviations (in parentheses), across 25 experiments. As reported in the table, our method achieves an average RMSE of 18.548 and an average MAE of 7.766, outperforming all benchmark methods. The evaluation results demonstrate the flexibility of our method in addressing infodemic management tasks beyond the binary classification setting.

[3] We excluded MCD and SDAT from the benchmark list because adapting these methods to a regression setting is nontrivial. Specifically, MCD relies on a distance measure that is inherently defined within a classification framework, while SDAT incorporates a smoothing loss that lacks a direct counterpart for regression tasks.

Table 7 Performance Comparison between Our Method and Benchmark Methods on Comment Count Prediction.

Method	RMSE↓	MAE↓
DACA	**18.548** (4.678)	**7.766** (1.294)
MMD	22.075** (4.917)	9.840** (2.744)
CANMD	21.168* (4.397)	8.699* (1.660)
DANN	21.066* (3.771)	10.181** (2.010)
MKMMD	28.296** (8.011)	13.488** (7.123)
JAN	29.899** (6.311)	9.097** (0.709)
DALN	21.152* (4.694)	9.639** (2.461)

Note: Significance levels are denoted by * and ** for 0.05 and 0.01, respectively.

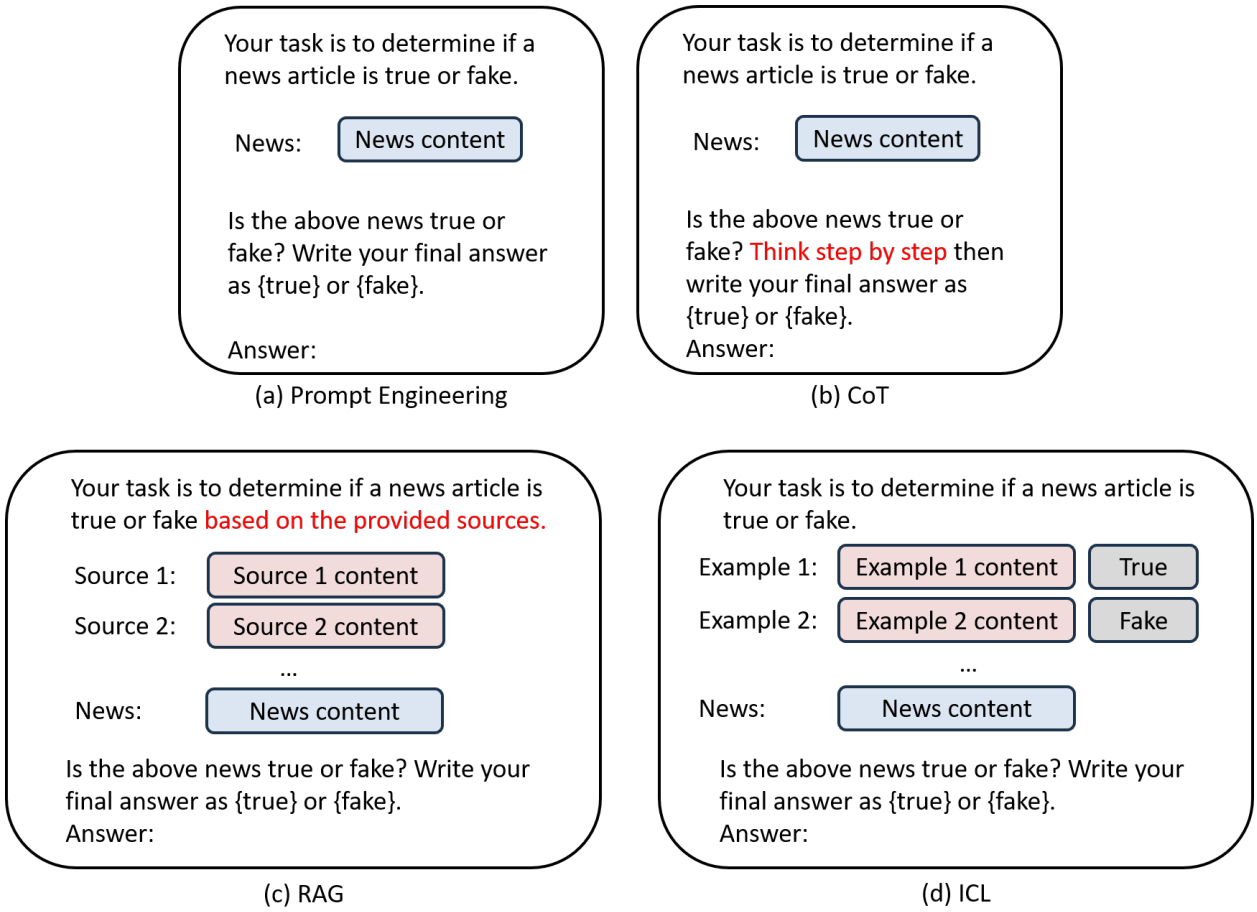
F.5 Performance Comparison between Our Method and LLM-based Approaches for Misinformation Detection

We compared the performance of our method with that of Large Language Model-based approaches (LLM-based approaches) for misinformation detection. The news articles in our COVID dataset were published in year 2020. Therefore, any LLM trained on data after 2020 may possess background knowledge for understanding and classifying these news (as true or fake). To ensure a fair comparison between LLMs and our method, the reviewer suggests to enforce a cutoff date for LLM training (i.e., restricting the LLMs to training data before 2020). Re-training a modern LLM from scratch using only data before 2020 would be a solution to enforce the cutoff date. However, this approach is prohibitively costly and beyond the scope of our study. An alternative is to use earlier LLMs trained on pre-2020 data. However, these models are generally much smaller and less capable. In our experiments, we found that early models such as GPT-2 were unable to follow instructions to determine whether a piece of information is true or fake.[4] Given these constraints, we adopted Mistral-7B-Instruct as our base model. Note that the LLM was released after September 2023. Although the training cutoff date for this LLM has not been disclosed, the model may have an inherent look-ahead advantage over our method when detecting fake news in our COVID dataset.

We implemented four ways of using Mistral-7B-Instruct for misinformation detection, as illustrated in Figure 3. One way is prompt engineering, where the LLM was directly instructed to judge whether

[4] A prompt-based approach (e.g., explicitly instructing an LLM to "only use knowledge before 2020") is unreliable. Since the model parameters are fixed after training, there is no mechanism to constrain the model to rely solely on pre-2020 knowledge. As a result, this approach does not provide a faithful enforcement of the cutoff date.

Figure 3 Using LLM for Misinformation Detection.



Note. Better view in color.

a piece of information is true or fake (Faruk 2024). Another approach is to leverage chain-of-thought reasoning (CoT). Specifically, the model was asked to think step by step before generating the final answer (Kojima et al. 2022). In addition, we employed retrieval-augmented generation (RAG) to assist the LLM in misinformation detection. Specifically, we randomly sampled 2,000 papers published in 2020 about COVID-19 and coronaviruses from the COVID-19 Open Research Dataset[5]. We then utilized the abstracts of these papers as the knowledge base for RAG. For the implementation of RAG, we employed a pretrained BERT to compute document embeddings for both the paper abstracts and the news articles. Next, for each news article, the top three most similar abstracts are retrieved from the knowledge base and provided as additional context to help determine its veracity. Finally, for in-context learning (ICL), we randomly sampled three news articles from the COVID dataset and attached their true labels. These three labeled news were then presented to the LLM as examples.

[5] See `https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/data`.

Table 8 Performance Comparison between Our Method and LLM-based Approaches for Misinformation Detection.

Method	Precision	Recall	F1-score	F2-score
DACA	**0.716** (0.030)	**0.745** (0.049)	**0.719** (0.034)	**0.739** (0.043)
Prompt Engineering	0.664** (0.007)	0.613** (0.014)	0.626** (0.014)	0.616** (0.013)
RAG	0.670** (0.006)	0.671** (0.006)	0.670** (0.006)	0.671** (0.006)
ICL	0.707 (0.027)	0.659** (0.027)	0.670** (0.019)	0.661** (0.025)

Note: Significance levels are denoted by * and ** for 0.05 and 0.01, respectively.

Table 8 presents the precision, recall, F1-score, and F2-score for each compared method, excluding CoT. For CoT, our empirical evaluation shows that the model labels all news articles as true. A similar pattern is observed when using a reasoning model (DeepSeek-R1-Distill-Llama-8B). This behavior could be attributed to the sycophancy of LLMs (Sharma et al. 2024), where models are trained to align with user-provided content, even when it contains misinformation. During the reasoning process, the LLM generates justifications for the claim and consequently outputs a "true" label. As reported in the table, our DACA method significantly outperforms all LLM-based approaches in recall, F1-score, and F2-score. It also surpasses these approaches in precision, though its lead over ICL is not statistically significant. Notably, our method achieves this superior performance despite the possible inherent look-ahead advantage of the LLM.

Appendix G Sensitivity Analysis

G.1 Model Training with Incomplete Source Domain Data

First, we evaluate the performance of our method when trained using a single source domain. Table 9 presents the average performance of DACA (with standard deviations in parentheses) across 25 experiments when trained using both the Politics and Entertainment source domains, as well as the results of DACA (Politics) and DACA (Entertainment) when trained on each domain individually. The results show that using both domains outperforms using either one alone, suggesting that diverse source data enhance model generalizability. Moreover, DACA (Politics) performs better than DACA (Entertainment), indicating that cross-domain learning from the Politics domain contributes more to the performance of our method than learning from the Entertainment domain.

Next, we investigate the effect of training with a subset (ranging from 75% to 25%) of the source domain data on the performance of our method. As an example, consider training with 25% of the source domain data. In this setting, we conducted 25 experiments, each of which randomly sampled

Table 9 Sensitivity Analysis of Source Domains.

	Precision	Recall	F1-score	F2-score
DACA	0.716 (0.030)	0.745 (0.049)	0.719 (0.034)	0.739 (0.043)
DACA (Politics)	0.685** (0.054)	0.727 (0.042)	0.692* (0.044)	0.718* (0.045)
DACA (Entertainment)	0.665** (0.028)	0.697** (0.047)	0.672** (0.032)	0.690** (0.043)

Note: Significance levels are denoted by * and ** for 0.05 and 0.01, respectively.

25% of the data from the Politics and Entertainment domains to train our method. Table 10 reports the average performance of our method (with standard deviations in parentheses) in each setting. As expected, the performance of our method declines as the proportion of the source domain data used for training decreases from 75% to 25%. At both the 50% and 25% levels, the performance of our method drops significantly across all metrics.

Table 10 Sensitivity Analysis of Partial Source Domain Data.

	Precision	Recall	F1-score	F2-score
DACA	0.716 (0.030)	0.745 (0.049)	0.719 (0.034)	0.739 (0.043)
DACA (75% of Source Data)	0.708 (0.039)	0.724* (0.033)	0.704 (0.038)	0.721 (0.039)
DACA (50% of Source Data)	0.681** (0.030)	0.700** (0.041)	0.684** (0.035)	0.696** (0.037)
DACA (25% of Source Data)	0.669** (0.044)	0.681** (0.031)	0.665** (0.038)	0.679** (0.031)

Note: Significance levels are denoted by * and ** for 0.05 and 0.01, respectively.

Finally, we examine the impact of using partially labeled source domain data on model performance and report the evaluation results in Table 11. In this case, we used all the source domain data for training, but only a portion of it (e.g., 75%) was labeled. For example, in an experiment, we randomly sampled 25% of the source domain instances and removed their labels, resulting in 75% of the instances being labeled. Among the three modules of our method, the classification and concept alignment modules require labeled data as input. Since labeled source domain data provides valuable signals for training these two modules, the performance of our method degrades as the percentage of labeled source domain data decreases from 75% to 25%.

Table 11 Sensitivity Analysis of Partially Labeled Source Domain Data.

	Precision	Recall	F1-score	F2-score
DACA	0.716 (0.030)	0.745 (0.049)	0.719 (0.034)	0.739 (0.043)
DACA (75% Labeled Source Data)	0.696 (0.048)	0.720* (0.037)	0.695* (0.034)	0.715* (0.034)
DACA (50% Labeled Source Data)	0.678** (0.049)	0.699** (0.038)	0.681** (0.031)	0.695** (0.035)
DACA (25% Labeled Source Data)	0.662** (0.036)	0.681** (0.035)	0.654** (0.035)	0.677** (0.040)

Note: Significance levels are denoted by * and ** for 0.05 and 0.01, respectively.

G.2 Heterogeneity in Source-Target Domain Closeness for Domain Adaption

We investigate the impact of the closeness between source and target domains on the performance of our method. To this end, we conducted additional sensitivity analyses using a public dataset of Chinese news (Nan et al. 2021), which spans a broad range of domains. Specifically, we employed news data in five different domains, as summarized in Table 12. Using labeled news from a source domain listed in Table 12, we applied our DACA method to predict the label for each piece of unlabeled news in the target COVID domain. Table 13 presents the average precision, recall, F1-score, and F2-score (with standard deviations in parentheses) for each source domain across 25 experiments. For example, when using labeled news from the healthcare domain as the source to predict labels of unlabeled news in the target COVID domain, our method achieves a precision of 0.824, a recall of 0.834, an F1-score of 0.822, and an F2-score of 0.832.

Table 12 Summary Statistics of Chinese News Dataset (Five Domains).

Domain	Number of true news	Number of fake news
COVID (Infodemic / Target)	201	124
Healthcare (Source)	284	391
Politics (Source)	306	546
Entertainment (Source)	1000	440
Military (Source)	121	222

The experimental results in Table 13 suggest that concept alignment and domain adaptation are more effective when source and target domains are more closely related. As shown, the performance of our method increases as we change the source domain from military (least related to the COVID

domain) to healthcare (most related to the COVID domain).[6] In addition, the experimental results also indicate that a distant source domain (e.g., military in the experiments) still provides signals for classifying completely unlabeled news in the target COVID domain.

Table 13 Performance of Our Method Across Different Source Domains.

Source Domain	Precision	Recall	F1-score	F2-score
Healthcare	0.824 (0.016)	0.834 (0.012)	0.822 (0.014)	0.832 (0.013)
Politics	0.750 (0.047)	0.724 (0.036)	0.719 (0.038)	0.729 (0.038)
Entertainment	0.698 (0.059)	0.702 (0.047)	0.700 (0.058)	0.701 (0.055)
Military	0.699 (0.037)	0.687 (0.055)	0.666 (0.042)	0.689 (0.060)

Labeled information (true or false information) in a source domain provides both signals and noise for classifying unlabeled information in the target domain. Signals arise because misinformation shares common characteristics across domains, such as emotional news headlines and inconsistencies between news headlines and news content (Zhou et al. 2020). However, according to Theorem 1, covariate and concept shifts between source and target domains introduce noise. Our method leverages signals in a source domain and mitigates noise in the domain to classify unlabeled information in the target domain. Therefore, as reported in Table 13, a distant source domain (i.e., military) is still useful for classifying information in the target domain. Moreover, a source domain that is more closely related to the target domain incurs smaller covariate and concept shifts between them. As a result, our method is more effective in classifying unlabeled target domain information when applied to a source domain that is more closely related to the target domain.

We also evaluated the performance of two benchmarks under different source-target domain pairs: CANMD (Yue et al. 2022) and DANN (Li et al. 2021). The performance of these methods decreases as the source domain becomes more distant from the target COVID domain. For example, the recall of CANMD drops from 0.788 to 0.647 as the source domain changes from healthcare to military. The performance variations of these methods are consistent with those of our method. The evaluation results further demonstrate that a domain adaptation method is more effective when source and target domains it employs are more closely related.

[6] To compute the distance between the target COVID domain and each source domain, we first embedded each Chinese news article in each domain as a numerical vector using BERT-wwm (Cui et al. 2021)). Next, we calculated the distance between domains using the single-kernel Maximum Mean Discrepancy (MMD) (Borgwardt et al. 2006). Specifically, the MMD distances between the COVID domain and the healthcare, politics, entertainment, and military domains are 0.071, 0.113, 0.178, and 0.245, respectively.

G.3 Performance Evaluation under a Chronological Setting

We evaluated the performance of our method from a chronological perspective. As a first step, we attempted to retrieve the publication date for each news article in the COVID domain. To this end, we accessed the raw data through a mirror version provided by Opendatalab[7]. We then extracted the publication date of a news article from its corresponding source-page URL. For example, the publication date of the article with the URL `https://www.snopes.com/news/2020/03/19/stopping-coronavirus-what-does-the-evidence-say-is-best/` can be directly identified as 2020-03-19. Using this approach, we successfully identified the publication dates of 2,703 articles with valid URLs containing date information. Among these articles, the earliest publication date is 2020-01-01, and the latest is 2020-06-30. We then partitioned these articles into two stages: Stage 1, which contains 755 articles published between 2020-01-01 and 2020-03-31, and Stage 2, which contains 1,948 articles published between 2020-04-01 and 2020-06-30. We observed a higher proportion of fake news in Stage 1 (43%) than in Stage 2 (31%).

Table 14 Performance Comparison between DACA and MK-MMD across Chronological Stages.

Method	Precision	Recall	F1-score	F2-score
DACA(Stage 1)	**0.670** (0.069)	**0.663** (0.074)	**0.667** (0.075)	**0.665** (0.071)
MK-MMD(Stage 1)	0.625** (0.072)	0.623** (0.064)	0.624** (0.062)	0.624** (0.068)
DACA(Stage 2)	**0.721** (0.057)	**0.716** (0.074)	**0.717** (0.072)	**0.717** (0.060)
MK-MMD(Stage 2)	0.666** (0.052)	0.671** (0.054)	0.668** (0.052)	0.670** (0.052)

Note: Significance levels are denoted by * and ** for 0.05 and 0.01, respectively.

We conducted additional experiments to compare the performance of our DACA method with that of the best-performing benchmark (MK-MMD) under chronological and cumulative settings. Specifically, for Stage 1, we trained each method using COVID news articles published up to 2020-03-31, together with source-domain news articles, and evaluated its performance on Stage 1 COVID news articles. For Stage 2, we trained each method using COVID news articles published up to 2020-06-30, along with source-domain news articles, and evaluated its performance on Stage 2 COVID news articles. Table 14 reports the average precision, recall, F1-score, and F2-score of DACA and MK-MMD in these two stages. As reported, our method significantly outperforms MK-MMD across all

[7] See `https://github.com/opendatalab`. We used the mirror version because the original Google Drive link (`https://drive.google.com/drive/folders/1gd4AvT6BxPRtymmNd9Z7ukyaVhae5s7U?usp=sharing`) associated with the dataset has expired.

metrics. Specifically, our method improves precision, recall, F1-score, and F2-score by 7.16%, 6.40%, 6.77%, and 6.55%, respectively, in Stage 1, and by 8.22%, 6.76%, 7.33%, and 6.96%, respectively, in Stage 2.

References

Bandari R, Asur S, Huberman B (2012) The Pulse of News in Social Media: Forecasting Popularity. Proceedings of the International AAAI Conference on Web and Social Media 6(1):26–33.

Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW (2010) A theory of learning from different domains. Machine Learning 79:151–175.

Borgwardt KM, Gretton A, Rasch MJ, Kriegel HP, Schölkopf B, Smola AJ (2006) Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics 22(14):e49–e57.

Chen L, Chen H, Wei Z, Jin X, Tan X, Jin Y, Chen E (2022) Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7181–7190.

Clarke J, Chen H, Du D, Hu YJ (2021) Fake News, Investor Attention, and Market Reaction. Information Systems Research 32(1):35–52, ISSN 1047-7047, 1526-5536.

Cui Y, Che W, Liu T, Qin B, Yang Z (2021) Pre-training with whole word masking for chinese bert. IEEE/ACM Transactions on Audio, Speech, and Language Processing 29:3504–3514.

Doob JL (1940) Regularity properties of certain families of chance variables. Transactions of the American Mathematical Society 47(3):455–486.

Faruk TB (2024) Evaluating the performance of large language models in scientific claim detection and classification. arXiv preprint arXiv:2412.16486 .

Foret P, Kleiner A, Mobahi H, Neyshabur B (2021) Sharpness-aware minimization for efficiently improving generalization. 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021 (OpenReview.net).

Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, March M, Lempitsky V (2016) Domain-adversarial training of neural networks. Journal of Machine Learning Research 17(59):1–35.

Kifer D, Ben-David S, Gehrke J (2004) Detecting change in data streams. VLDB, volume 4, 180–191 (Toronto, Canada).

Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y (2022) Large language models are zero-shot reasoners. Advances in neural information processing systems 35:22199–22213.

Li Y, Jiang B, Shu K, Liu H (2020) Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. arXiv preprint arXiv:2011.04088 .

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

Liu X, Yoo C, Xing F, Oh H, El Fakhri G, Kang JW, Woo J, et al. (2022) Deep unsupervised domain adaptation: A review of recent advances and perspectives. APSIPA Transactions on Signal and Information Processing 11(1).

Liu Z, Zhang X, Yang K, Xie Q, Huang J, Ananiadou S (2025) FMDLlama: Financial Misinformation Detection Based on Large Language Models. Companion Proceedings of the ACM on Web Conference 2025, 1153–1157, WWW '25 (New York, NY, USA: Association for Computing Machinery), ISBN 979-8-4007-1331-6, URL `http://dx.doi.org/10.1145/3701716.3715599`.

Long M, Cao Y, Wang J, Jordan M (2015) Learning transferable features with deep adaptation networks. Proceedings of the 32nd International Conference on Machine Learning, 97–105 (PMLR).

Long M, Zhu H, Wang J, Jordan MI (2017) Deep transfer learning with joint adaptation networks. Proceedings of the 34th International Conference on Machine Learning, 2208–2217 (PMLR).

Lu W, Tong Y, Ye Z (2025) DAMMFND: Domain-Aware Multimodal Multi-view Fake News Detection. Proceedings of the AAAI Conference on Artificial Intelligence 39(1):559–567, ISSN 2374-3468, URL `http://dx.doi.org/10.1609/aaai.v39i1.32036`, number: 1.

Mosallanezhad A, Karami M, Shu K, Mancenido MV, Liu H (2022) Domain adaptive fake news detection via reinforcement learning. Proceedings of the ACM web conference 2022, 3632–3640.

Nan Q, Cao J, Zhu Y, Wang Y, Li J (2021) Mdfend: Multi-domain fake news detection. Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 3343–3347.

Nan Q, Wang D, Zhu Y, Sheng Q, Shi Y, Cao J, Li J (2022) Improving fake news detection of influential domain via domain-and instance-level transfer. Proceedings of the 29th International Conference on Computational Linguistics, 2834–2848.

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

Peng L, Jian S, Kan Z, Qiao L, Li D (2024) Not all fake news is semantically similar: Contextual semantic representation learning for multimodal fake news detection. Information Processing & Management 61(1):103564.

Rangapur A, Wang H, Shu K (2023) Investigating Online Financial Misinformation and Its Consequences: A Computational Perspective. URL `http://dx.doi.org/10.48550/arXiv.2309.12363`, arXiv:2309.12363 [cs].

Rangwani H, Aithal SK, Mishra M, Jain A, Radhakrishnan VB (2022) A closer look at smoothness in domain adversarial training. Proceedings of the 39th International Conference on Machine Learning, 18378–18399 (PMLR).

Saito K, Watanabe K, Ushiku Y, Harada T (2018) Maximum classifier discrepancy for unsupervised domain adaptation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3723–3732.

Sharma M, Tong M, Korbak T, Duvenaud D, Askell A, Bowman SR, DURMUS E, Hatfield-Dodds Z, Johnston SR, Kravec SM, et al. (2024) Towards understanding sycophancy in language models. The Twelfth International Conference on Learning Representations.

Shu K, Mahudeswaran D, Wang S, Lee D, Liu H (2020) Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big data 8(3):171–188.

Sun T, Qian Z, Dong S, Li P, Zhu Q (2022) Rumor Detection on Social Media with Graph Adversarial Contrastive Learning. Proceedings of the ACM Web Conference 2022, 2789–2797, WWW '22 (New York, NY, USA: Association for Computing Machinery), ISBN 978-1-4503-9096-5, URL http://dx.doi.org/10.1145/3485447.3511999.

Thorne J, Vlachos A, Christodoulopoulos C, Mittal A (2018) Fever: a large-scale dataset for fact extraction and verification. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 809–819.

Wei X, Zhang Z, Zhang M, Chen W, Zeng DD (2022) Combining Crowd and Machine Intelligence to Detect False News on Social Media. MIS Quarterly 46(2):977–1008, URL http://dx.doi.org/https://doi.org/10.25300/MISQ/2022/16256.

Wu L, Long Y, Gao C, Wang Z, Zhang Y (2023) Mfir: Multimodal fusion and inconsistency reasoning for explainable fake news detection. Information Fusion 100:101944.

Yue Z, Zeng H, Kou Z, Shang L, Wang D (2022) Contrastive domain adaptation for early misinformation detection: A case study on covid-19. Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2423–2433.

Zhang L, Zhang X, Zhou Z, Huang F, Li C (2024) Reinforced Adaptive Knowledge Learning for Multimodal Fake News Detection. Proceedings of the AAAI Conference on Artificial Intelligence 38(15):16777–16785, ISSN 2374-3468, URL http://dx.doi.org/10.1609/aaai.v38i15.29618, number: 15.

Zhang X, Cao J, Li X, Sheng Q, Zhong L, Shu K (2021) Mining dual emotion for fake news detection. Proceedings of the web conference 2021, 3465–3476.

Zhou X, Jain A, Phoha VV, Zafarani R (2020) Fake news early detection: A theory-driven model. Digital Threats: Research and Practice 1(2):1–25.

Zhu Y, Sheng Q, Cao J, Nan Q, Shu K, Wu M, Wang J, Zhuang F (2023) Memory-Guided Multi-View Multi-Domain Fake News Detection. IEEE Transactions on Knowledge and Data Engineering 35(7):7178–7191, ISSN 1041-4347, URL http://dx.doi.org/10.1109/TKDE.2022.3185151.

Zubiaga A, Liakata M, Procter R (2016) Learning Reporting Dynamics during Breaking News for Rumour Detection in Social Media. URL http://dx.doi.org/10.48550/arXiv.1610.07363, arXiv:1610.07363 [cs].