Improving Fake News Detection of Influential Domain via Domain- and Instance-Level Transfer

Qiong Nan[1,2], Danding Wang[1], Yongchun Zhu[1,2], Qiang Sheng[1,2]
Yuhui Shi[1,2], Juan Cao[1,2],[*] Jintao Li[1]

[1]Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences

{nanqiong19z,wangdanding,zhuyongchun18s}@ict.ac.cn
{shengqiang18z,caojuan,jtli}@ict.ac.cn
shiyuhui221@mails.ucas.ac.cn

Abstract

Both real and fake news in various domains, such as politics, health, and entertainment are spread via online social media every day, necessitating fake news detection for multiple domains. Among them, fake news in specific domains like politics and health has more serious potential negative impacts on the real world (e.g., the infodemic led by COVID-19 misinformation). Previous studies focus on multi-domain fake news detection, by equally mining and modeling the correlation between domains. However, these multi-domain methods suffer from a seesaw problem: the performance of some domains is often improved at the cost of hurting the performance of other domains, which could lead to an unsatisfying performance in specific domains. To address this issue, we propose a Domain- and Instance-level Transfer Framework for Fake News Detection (DITFEND), which could improve the performance of specific target domains. To transfer coarse-grained domain-level knowledge, we train a general model with data of all domains from the meta-learning perspective. To transfer fine-grained instance-level knowledge and adapt the general model to a target domain, we train a language model on the target domain to evaluate the transferability of each data instance in source domains and re-weigh each instance's contribution. Offline experiments on two datasets demonstrate the effectiveness of DITFEND. Online experiments show that DITFEND brings additional improvements over the base models in a real-world scenario.

1 Introduction

With the rapid popularization of the Internet, more and more people tend to acquire news through social media platforms, such as Weibo[1] and Twitter[2]. Due to the above phenomenon, fake news

---

[*]Corresponding author
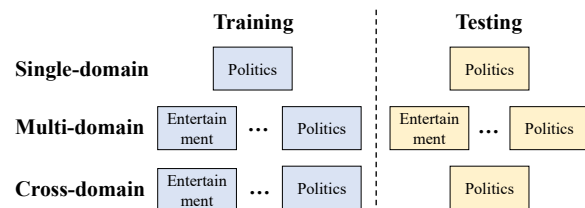[1]https://weibo.com
[2]https://twitter.com

Figure 1: Illustration of the differences among single-domain, multi-domain, and cross-domain (**ours**) fake news detection.

has spread widely all over the world. During the 2016 U.S. presidential election campaign, the top 20 frequently discussed fake election stories generated 8,711,000 engagements on Facebook, which is larger than the total of 7,367,000 for the top 20 most-discussed election stories posted by 19 major news websites (Silverman, 2016). The wide spread of fake news may break the authenticity balance of the news ecosystem (Shu et al., 2017), and it not only misleads many people but also leads to the social mobs and social panic (Chen, 2020). Therefore, fake news detection is of significant importance.

Real-world news platforms categorize news pieces by topic into various domains, e.g., politics, health, and entertainment. Although it is necessary to detect fake news in every domain, the societal effects of fake news from some specific domains are more serious. For example, during the U.S. election, fake news in the political domain may have dictated election results (Allcott and Gentzkow, 2017), and Vosoughi et al. (2018) find that false political news travels deeper more quickly and more broadly, reaches more people, and is more vital than other categories of false information. In the COVID-19 infodemic (Bursztyn et al., 2020), thousands of fake news pieces have caused social panic (Chen, 2020) and weakened the effect of pandemic countermeasures (Bursztyn et al., 2020). Therefore, it is important to detect fake news for these specific news domains, e.g., politics (Wang, 2017; Jin et al., 2017), health (Bang

et al., 2021; Shi et al., 2020). Some researchers pay attention to fake news detection in the influential news domains by modeling each domain separately (Wang, 2017; Allcott and Gentzkow, 2017; Bovet and Makse, 2019; Dai et al., 2020; Cui et al., 2020; Zhou et al., 2020; Shang et al., 2022). However, a single-domain dataset could only contain limited data, e.g., the PolitiFact dataset of the political domain only has 948 samples (Shu et al., 2020). Fortunately, news pieces in different domains are correlated (Nan et al., 2021; Silva et al., 2021), so it is promising to exploit data of other domains to improve performance for a target domain.

Recently, some works (Wang et al., 2018; Silva et al., 2021; Nan et al., 2021) simultaneously model various domains to improve the overall performance of all domains. However, these multi-domain methods suffer from a serious seesaw phenomenon which could cause the performance of some domains to be improved at the cost of hurting the performance of other domains (Tang et al., 2020). For example, EANN (Wang et al., 2018) performs quite well in *Military* (with f1-score of 0.9274), whereas the results is not satisfactory in *Politics* (with f1-score of 0.8705) (Nan et al., 2021). Moreover, it is hard to guide these multi-domain models to improve the performance of a specific target domain due to the lack of a target-oriented design. In this paper, we focus on exploiting news pieces of other domains to improve the detection performance of a certain target domain, called **cross-domain fake news detection**, which can bring additional gains for target domains compared to multi-domain methods. The differences among single-domain, multi-domain, and cross-domain fake news detection are shown in Figure 1. To solve cross-domain fake news detection, we adopt two key ideas:

Transfer domain-level knowledge. It is necessary to transfer knowledge from multiple domains because news pieces in different domains are correlated. Moreover, to alleviate the seesaw phenomenon, improving the generalization ability of a multi-domain model, which can adapt fast to the target domain is necessary.

Transfer instance-level knowledge. The transferability varies from instance to instance. For example, "*A politician claimed COVID-19 is less lethal than flu*" and "*New York Officials welcome immigrants, legal or illegal*" are two different news pieces of politics. The former is more relevant to

health while the latter is irrelevant. In other words, the former is more transferable. Therefore, it is important to quantify the transferability of source instances, in order to decrease the impact of irrelevant instances.

Along this line, we propose a Domain- and Instance-level Transfer framework for Fake News Detection (DITFEND). To transfer domain-level knowledge, we exploit data from multiple domains to train a general model from the meta-learning perspective, which contains common knowledge and can adapt fast to a specific domain. To transfer instance-level knowledge, we first learn a domain-adaptive language model which is endowed with characteristics of the target domain. To weigh the contribution to the target domain of every instance in source domains, we adopt an index, **perplexity** of the domain-adaptive language model, to quantify the transferability of these instances. Because in information theory perplexity can measure how well a probability model predicts an instance, in other words, low perplexity on a news piece indicates that the instance is highly related to the knowledge contained in the language model. Finally, we adapt the general model with instances from the target domain and weighted instances from source domains, in order to achieve satisfying performance on the target domain.

The main contributions of this paper can be summarized as follows:

- We investigate the importance of cross-domain fake news detection with multiple sources for target domains for the first time.

- We propose a Domain- and Instance-level Transfer Framework to improve fake news detection of target domains.

- We evaluate our proposed DITFEND on both English and Chinese real-word fake news datasets, and experiments demonstrate the effectiveness of DITFEND.

2 Related Work

2.1 Fake News Detection

Fake news detection aims at automatically classifying a news piece as real or fake. Existing methods can be generally grouped into two clusters: social-context-based methods and content-based methods (Shu et al., 2017). For social-context-based methods, some analyze propagation patterns to mine structural signals for fake news de-

tection (Jin et al., 2014; Liu and Wu, 2018; Shu et al., 2019; Mosallanezhad et al., 2022; Naumzik and Feuerriegel, 2022), others use the wisdom of crowds, such as emotion and stance, to detect fake news (Zhang et al., 2021; Jin et al., 2016), and Sheng et al. (2022) captures the environmental signals to detect fake news posts. For content-based methods, some extract evidence from external sources (Vlachos and Riedel, 2014; Shi and Weninger, 2016; Chen et al., 2022; Xu et al., 2022; Sheng et al., 2021), while others only analyze news itself and focus on better constructing features (Przybyla, 2020; Wang et al., 2018), which is within the scope of our research.

Since real-world news platforms categorize news pieces by topic into various domains, some researchers pay attention to the fake news detection performance of each domain, especially for that with serious societal effects. Some only take one specific domain into consideration and perform single-domain fake news detection (Wang, 2017; Allcott and Gentzkow, 2017; Bovet and Makse, 2019; Dai et al., 2020; Cui et al., 2020; Zhou et al., 2020; Shang et al., 2022), however, they ignore useful information from other domains. Some methods simultaneously model various domains to improve the overall performance of all domains (multi-domain fake news detection) (Wang et al., 2018; Silva et al., 2021; Nan et al., 2021; Zhu et al., 2022), however, due to the seesaw phenomenon, the performance on some target domains suffer from degeneration. Huang et al. (2021) and Mosallanezhad et al. (2022) propose to use a domain adaptation strategy for cross-domain fake news detection, however, they only transfer knowledge from one single source domain to ensure the model's detection performance on the target domain. However, in the practical scenario, news pieces from multiple source domains are inherently correlated with the target domain. It is straightforward to combine all source domains into one single domain and perform cross-domain fake news detection as (Huang et al., 2021) and (Mosallanezhad et al., 2022) do, but the improvement may not be significant. Hence, it is necessary to find a better way to make full use of all source domains to improve the performance of a target domain.

2.2 Transfer Learning

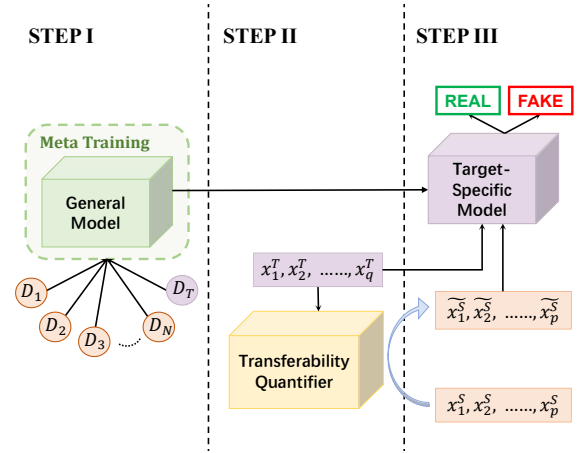Transfer learning aims to leverage knowledge from a source domain to improve the learning perfor-

Figure 2: The overall framework of our DITFEND model. Step I is the general model training procedure, step II is the transferability quantifying procedure, and step III is the target domain adaptation procedure. The three steps correspond to Section 3.2, 3.3 and 3.4 respectively.

mance or minimize the number of labeled examples required in a target domain (Pan and Yang, 2009; Zhuang et al., 2020). Recently, transfer learning has been widely adopted for natural language processing (Devlin et al., 2019; Liu et al., 2019; Lewis et al., 2020), e.g., sentiment classification (Peng et al., 2018), neural machine translation (Kim et al., 2019), style transfer (Yang et al., 2018). Meanwhile, meta learning serves as a paradigm that can be used to improve transfer learning problems (Hospedales et al., 2021). Benefit from its ability to integrate prior experience as well as generality to all domains, meta learning has been widely adopted in many applications (Li et al., 2018; Wang et al., 2021). In this paper, we propose a novel transfer framework based on meta-learning for cross-domain fake news detection.

3 DITFEND: Domain- and Instance-Level Transfer for Fake News Detection

3.1 Problem Statement

In the scenario of cross-domain fake news detection, there is one target domain $\mathcal{D}_T$ and a collection of N source domains $\mathcal{D}_S = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_N\}$. For each domain, we have a collection of news pieces. For a piece of news P, we pad it with additional characters $[CLS]$ and $[SEP]$. Then we tokenize the sentence into m tokens $\{t_1, t_2, ..., t_m\}$, and encode it to get the corresponding embedding vector e. To distinguish between different domains, we denote e_T as the embedding vector of the target domain

Algorithm 1 General Model Training Procedure.

Input: Given N source domains and one target domain.
Input: The base model $\mathcal{M}(\theta)$.
Input: Learning rate α and β.

1. randomly initialize θ
2. **while** not coverage **do**:
3.   sample batch of training tasks $\{\mathcal{T}_1, ..., \mathcal{T}_n\}$
4.   **for** all $\mathcal{T}_d \in \{\mathcal{T}_1, ..., \mathcal{T}_n\}$ **do**:
5. $\mathcal{T}_d$ contains two disjoint sets $\mathcal{D}_d^s$ and $\mathcal{D}_d^q$
6.   evaluate loss $\mathcal{L}_d^s(\theta)$ with $\mathcal{D}_d^s$
7.   compute updated parameter
   $\theta_d = \theta - \alpha \frac{\partial \mathcal{L}_d^s(\theta)}{\partial \theta}$
8.   evaluate loss $\mathcal{L}_d^q(\theta_d)$ with $\mathcal{D}_d^q$
9.   **end**
10.   update $\theta = \theta - \beta \sum_{\mathcal{T}_d \in \{\mathcal{T}_1, ..., \mathcal{T}_n\}} \frac{\partial \mathcal{L}_d^q(\theta_d)}{\partial \theta}$
11. **end**

---

and $\boldsymbol{e_S} \in \{\boldsymbol{e_1}, \boldsymbol{e_2}, ..., \boldsymbol{e_N}\}$ of source domains.

The core idea of our framework is to leverage news pieces from all domains to train a target-adaptive fake news detector. To achieve this, in the first stage, we train a general model with news pieces from all domains, such that the model can alleviate the seesaw phenomenon (i.e., domain-level transfer); In the second stage, we evaluate and quantify the transferability of instances from source domains; And in the third stage, we adapt the general model to the target domain. The overall framework is shown in Figure 2.

3.2 General Model Training

For domain-level transfer, we take the advantage of meta-learning and train a general model $\mathcal{M}(\theta)$ to aggregate the knowledge from all the domains.

In each iteration of the parameter update, we draw a batch of training tasks $\{\mathcal{T}_1, ..., \mathcal{T}_n\}$, and for each task $\mathcal{T}_d \in \{\mathcal{T}_1, ..., \mathcal{T}_n\}$, we divide it into two disjoint sets: a support set $\mathcal{D}_d^s$ and a query set $\mathcal{D}_d^q$. The model is trained with samples from support set $\mathcal{D}_d^s$ and feedback with the corresponding loss $\mathcal{L}_d^s$ from $\mathcal{D}_d^s$:

$$\mathcal{L}_d^s(\theta) = \frac{1}{m_s} \sum_{i=1}^{m_s} -y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i),$$
(1)

where m_s is the number of the data in the current support set. We adopt cross-entropy loss because it is widely used as the optimization goal for binary fake news detection (Wang et al., 2018; Silva et al., 2021; Nan et al., 2021). And then we use $\mathcal{L}_d^s$ to optimize the parameter of the current task $\mathcal{T}_d$ via

gradient descent:

$$\theta_d = \theta - \alpha \frac{\partial \mathcal{L}_d^s(\theta)}{\partial \theta},$$
(2)

where α is the learning rate of the training process within each task, and θ_d is the optimized model parameters for the current task $\mathcal{T}_d$.

And then the model is tested on samples from the current query set $\mathcal{D}_d^q$. The model is improved by considering how the test error $\mathcal{L}_d^q$ on $\mathcal{D}_d^q$ changes with respect to the parameters:

$$\mathcal{L}_d^q(\theta_d) = \frac{1}{m_q} \sum_{i=1}^{m_q} -y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i),$$
(3)

where m_q is the number of data in the current query set.

In effect, the testing error on $\mathcal{D}_d^q$ serves as the training error of the meta-learning process.

After we loop over all tasks in $\{\mathcal{T}_1, ..., \mathcal{T}_n\}$, the base model's parameters θ can be updated as follows:

$$\theta = \theta - \beta \nabla_\theta \sum_{i=1}^{n} \mathcal{L}_i^q,$$
(4)

where β is the learning rate of the meta-learning process. The procedure of general model training is summarized in Alg 1.

3.3 Transferability Quantifying

In order to endow the language model with target domain knowledge, we perform the Masked Language Modeling task on the target domain to get a domain-adaptive language model. This second phase of training the language model can bring significant performance improvement on the following task based on the language model (Gururangan et al., 2020). And we utilize the domain-adaptive language model to evaluate the transferability of source instances.

Domain-adaptive Language Model Training. Let $\mathcal{D}_T$ be a dataset of news pieces from the target domain T, where P_T is a news piece in $\mathcal{D}_T$ containing n tokens $\{w_1, ..., w_n\}$. We replace 15% tokens in the input sequence with the $[MASK]$ token, a random token, or the original token. The model is required to predict the masked tokens based on the other tokens in P_T. The training objective is to minimize the cross-entropy loss of the language model in predicting the masked tokens of the target data:

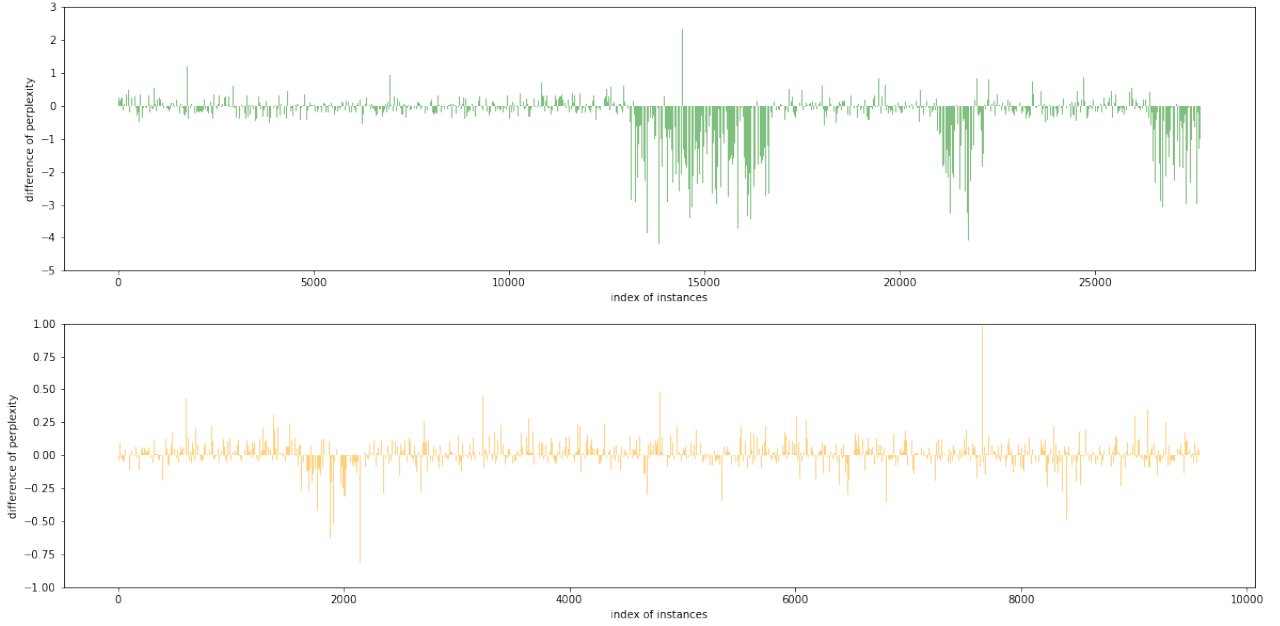$$\min_{\theta_m} \sum_{x \in D_t} \mathcal{L}_m(x, \theta_m),$$
(5)

Figure 3: The D-value of different target language models' perplexity on the same batch of source instances (Top: Chinese dataset; Bottom: English dataset).

where θ_m is the parameters of the language model (*bert-base-uncased* for English and *bert-base-chinese* for Chinese).

After performing the Masked Language Modeling task on the target domain, we can effectively endow the corresponding language model with knowledge of the target domain.

Perplexity is an indicator that can measure how well a language model predicts a sentence, i.e., the higher the probability of prediction is, the better the language model is, and the lower the language model's perplexity is. Many researchers use it to evaluate language models (Belinkov and Glass, 2019; Yogatama et al., 2018). We exploit perplexity from another perspective – the better a language model can predict a sentence, the better this sentence fits the knowledge endowed in the language model. Therefore, it is intuitive to compute the perplexity of the language model on a given sample to quantify the sample's transferability, i.e., the lower perplexity indicates stronger transferability.

Let $\mathcal{D}_S$ indicate one of the source datasets, where P_S is a news piece in $\mathcal{D}_S$ containing m tokens $\{w_0, w_2, ..., w_{m-1}\}$. We pad it with two tokens, i.e., $[CLS]$ and $[SEP]$. A masked sentence P_{mask} is generated by masking a word/character in the sentence as follows:

$$\mathcal{P}_{mask} = <[CLS], ..., w_{i-1}, [MASK], w_{i+1}, ..., [SEP]>, \quad (6)$$

where $0 \leq i \leq m - 1$. Then the target-adaptive

language model is utilized to predict the probability of the right words w_i in the $[MASK]$ position:

$$prob(w_i) = MLM(\mathcal{P}, w_i). \quad (7)$$

After we calculated the probability of all the words in the sentence, the perplexity of the language model on the whole sentence is calculated as follows:

$$pp = \sqrt[N]{\prod_{i=0}^{N-1} \frac{1}{prob(w_i)}}. \quad (8)$$

To show the difference in the target-adaptive language model when assigning different target domains, we compute the difference value of different target-adaptive language models' perplexity on source instances. And we visualize the distribution in Figure 3 in detail. From the distribution, we can see that there exists an obvious difference in both the Chinese and English datasets.

Finally, we quantify the transferability of instances from source domains as follows:

$$w = 1/pp, \quad (9)$$

where w is the indicator of the transferability of the sample. In this way, we quantify the transferability of source instances, i.e., assigning bigger weights to samples with lower perplexity.

3.4 Target Domain Adaptation

In this section, we aim to adapt the general fake news detection model to the target domain. In Section 3.3, we have assigned each sample in source domains with a weight value to indicate its transferability. To make full use of source domain samples, we re-weigh them based on their transferability, along with the target domain samples to train the general model.

We exploit the general model trained via meta-learning (Section 3.2), and optimize it by minimizing the cross-entropy loss as follows:

$$\mathcal{L}^{ce}(y, \hat{y}) = -ylog\hat{y} - (1-y)log(1-\hat{y})$$
$$\mathcal{L} = \mathbb{E}_{(x,y)\sim p_s(x,y)}w(x)\mathcal{L}^{ce}(y, \hat{y}) \quad (10)$$
$$+ \mathbb{E}_{(x,y)\sim p_t(x,y)}\mathcal{L}^{ce}(y, \hat{y}),$$

where y is the ground truth, $\hat{y}$ is the predicted label, and $w(x)$ is the indicator of the transferability of the instance x obtained in Section 3.3.

4 Experiments

In this section, we aim to answer the following evaluation questions:

- **EQ1:** Can DITFEND improve the performance of a target domain when coordinates with different base models?

- **EQ2:** How effective are domain-level transfer and instance-level transfer?

- **EQ3:** Can DITFEND perform well in a real-world fake news detection scenario?

4.1 Datasets

We evaluate DITFEND on both Chinese and English datasets. And the statistics of datasets are listed in Tables 1 and 2.

Table 1: Statistics of the Chinese dataset

Domain	Science	Military	Edu.	Disaster	Politics
# Fake	93	222	248	591	546
# Real	143	121	243	185	306
Total	236	343	491	776	852

Domain	Health	Finance	Ent.	Society	All
# Fake	515	362	440	1,471	4,488
# Real	485	959	1,000	1,198	4,640
Total	1,000	1,321	1,440	2,669	9,128

English Dataset. We combine two datasets (PolitiFact and GossipCop) in FakeNewsNet (Shu et al., 2020) and COVID (Li et al., 2020) into

Table 2: Statistics of the English Dataset

Domain	PolitiFact	GossipCop	COVID	All
# Fake	420	4,947	1,317	7,483
# Real	528	16,694	4,750	22,864
Total	948	21,641	6,067	30,347

an English dataset, which contains three domains, namely PolitiFact, GossipCop and COVID.

Chinese Dataset (Nan et al., 2021). multi-domain fake news detection dataset collected from Sina Weibo. There are 9 domains in total, which is Science, Military, Education, Disaster, Politics, Health, Finance, Entertainment, and Society. In our experiments, in order to evaluate the impact of the domain number, we sample 3 domains (Politics, Health, Entertainment) and 6 domains (Education, Disaster, Health, Finance, Entertainment, Society) from the Chinese dataset to construct the other two multi-domain datasets, i.e., Chinese 3-domain and Chinese 6-domain.

4.2 Baseline Models

In this paper, we focus on textual content-based fake news detection. Technically, our DITFEND framework can coordinate with any text-based models that produce post-representation. Thus, we use 8 representative text-based models as our base models, which can be divided into three groups: (1) text classification models: **TextCNN** (Kim, 2014), **BiGRU** (Ma et al., 2016), and **RoBERTa** (Liu et al., 2019; Cui et al., 2019); (2) transfer learning models: **MMOE** (Ma et al., 2018) and **MOSE** (Qin et al., 2020); (3) multi-domain fake news detection models: **EANNᴛ** (Wang et al., 2018), **EDDFN** (Silva et al., 2021), and **MDFEND** (Nan et al., 2021). Details about the base models are as follows:

TextCNN (Kim, 2014): Convolutional Neural Networks (CNNs) have been proven to gain remarkably strong performance on the task of text classification. In our experiment, we use filter windows of 1, 2, 3, 5, 10 with 64 feature maps each.

BiGRU (Ma et al., 2016): It is a widely used models in natural language processing applications. Different from (Ma et al., 2016), we treat each piece of news as a sequential input to a one-layer BiGRU model.

RoBERTa (Liu et al., 2019; Cui et al., 2019): It is a robustly optimized BERT (Devlin et al., 2019) pre-trained model. We utilize it to encode tokens of news content and feed the extracted embedding

Table 3: Performance comparison of base models with and without DITFEND on English dataset and Chinese 3-domain dataset.

Method	English dataset						Ch-3 dataset					
	target: COVID			target: Politifact			target: Health			target: Politics		
	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc
BiGRU	0.7448	0.9114	0.8606	0.7339	0.8213	0.7375	0.8577	0.9367	0.8580	0.8384	0.9032	0.8637
+*DITFEND*	**0.9219**	**0.9890**	**0.9501**	**0.8476**	**0.9043**	**0.8477**	**0.8792**	**0.9548**	**0.8795**	**0.8528**	**0.9261**	**0.8772**
TextCNN	0.8322	0.9397	0.8955	0.7040	0.8046	0.7064	0.8716	0.9552	0.8720	**0.8579**	0.9067	**0.8859**
+*DITFEND*	**0.8642**	**0.9706**	**0.9184**	**0.7913**	**0.8740**	**0.7913**	**0.8878**	**0.9582**	**0.8880**	0.8563	**0.9127**	0.8807
RoBERTa	0.9014	0.9770	0.9377	0.7967	0.9078	0.7989	0.8955	0.9641	0.8955	0.8300	0.8948	0.8628
+*DITFEND*	**0.9360**	**0.9901**	**0.9578**	**0.8608**	**0.9183**	**0.8609**	**0.9105**	**0.9708**	**0.9105**	**0.8445**	**0.9085**	**0.8725**
EANN_T	0.8836	0.9751	0.9282	0.7558	0.8612	0.7584	0.9189	**0.9787**	0.9190	0.8405	0.9074	0.8690
+*DITFEND*	**0.8883**	**0.9825**	**0.9310**	**0.8040**	**0.9074**	**0.8046**	**0.9219**	0.9760	**0.9220**	**0.8574**	**0.9136**	**0.8854**
MMOE	**0.9379**	0.9883	0.9588	0.8477	0.9408	0.8486	**0.9215**	0.9639	**0.9215**	**0.8779**	0.9388	**0.8982**
+*DITFEND*	0.9361	**0.9911**	**0.9600**	**0.8613**	**0.9515**	**0.8615**	0.9034	**0.9657**	0.9035	0.8523	**0.9398**	0.8766
MOSE	0.9326	0.9879	0.9588	0.8576	0.9447	0.8590	0.9023	0.9683	0.9025	0.8564	0.9138	0.8795
+*DITFEND*	**0.9586**	**0.9880**	**0.9712**	**0.8732**	**0.9553**	**0.8732**	**0.9069**	**0.9711**	**0.9070**	**0.8642**	**0.9220**	**0.8865**
EDDFN	0.9306	0.9891	0.9547	0.8505	0.9432	0.8509	0.9235	**0.9735**	0.9235	0.8440	0.9207	0.8702
+*DITFEND*	**0.9401**	**0.9912**	**0.9600**	**0.8720**	**0.9466**	**0.8725**	**0.9245**	0.9731	**0.9245**	**0.8486**	**0.9251**	**0.8731**
MDFEND	0.9331	0.9874	0.9565	0.8473	0.9391	0.8485	0.9419	0.9855	0.9420	0.8555	0.9259	0.8854
+*DITFEND*	**0.9485**	**0.9934**	**0.9700**	**0.8589**	**0.9500**	**0.8593**	**0.9530**	**0.9856**	**0.9530**	**0.8663**	**0.9368**	**0.8895**

into an MLP to obtain the final prediction. *roberta-base-chinese* and *roberta-base-uncased* (Liu et al., 2019; Cui et al., 2019) are exploited for Chinese and English datasets, respectively.

EANN_T (Wang et al., 2018): It is a model that aims to learn event-agnostic features, which uses a TextCNN module for text representation and adopts an auxiliary event discriminator for adversarial learning. In our experiments, we only use the textual branch, and rather than event-level, we perform domain-level adversarial training (i.e., use the discriminator to classify different domains) to learn domain-shared features.

MMOE (Ma et al., 2018) and **MOSE** (Qin et al., 2020): These two models are proposed for multi-task learning. In our experiments, we assume that fake news in different domains are different tasks, and use the two models for multi-domain modeling.

EDDFN (Silva et al., 2021): It is proposed for multi-domain fake news detection, which models different domains implicitly and jointly preserves domain-specific and cross-domain knowledge.

MDFEND (Nan et al., 2021): It is a multi-domain fake news detection model, which utilizes a domain gate to aggregate multiple representations extracted by mixture-of-experts for multi-domain fake news detection.

4.3 Experiment Settings

Training Procedure. I. We use samples from all domains to train a general model from the meta-learning perspective. **II.** We use samples from the target domain to train a domain-adaptive language model via *Masked Language Modeling* task. **III.** We calculate the perplexity of the domain-adaptive language model on samples from source domains, which is used to re-weigh the samples. **IV.** We use samples from the target domain and source domains (assigned with the corresponding weights) to train the general model to adapt to the target domain.

Target domain assignment. In practice, fake news pieces of Politics, Health, and Finance have more serious influence. For Ch-3 and Ch-9 dataset, we choose *Politics* and *Health* as the target domain, respectively. For Ch-6 dataset, we choose *Health* and *Finance* as the target domain,respectively. For English dataset, we choose *PolitiFact* and *COVID* as the target domain, respectively.

Evaluation Metrics. We treat the fake news detection problem as a binary classification task. We report macro F1 score (F1), accuracy (Acc), and Area Under ROC (AUC). For the online tests, where the number of fake news is much lower than real news, we should detect fake news as accurately as possible without misclassifying real news. Thus, we further report standardized partial AUC ($SPAUC_{FPR \leq 0.1}$).

Implementation Details. We limit the max length of the sentence to 170 tokens for Chinese, and 300 tokens for English. Following the settings in (Nan et al., 2021), we tokenized sentences with jieba[3] (for Chinese), nltk (Bird, 2006) (for English), and embed them with Word2Vec (Le and

---

[3]https://github.com/fxsjy/ jieba

Mikolov, 2014; Mikolov et al., 2013) for BiGRU and TextCNN. We use RoBERTa and the corresponding tokenizer for other models. For the two embedding types, we fix the dimension of embeddings to 768 for RoBERTa (Devlin et al., 2019) and 200 for Word2Vec (Le and Mikolov, 2014; Mikolov et al., 2013). For a fair comparison, we set the same hyperparameters for all base models. The MLP (Multi-Layer Perceptron) module used in these models contains one dense layer (384 hidden units).

4.4 Performance Comparison (EQ1)

To answer **EQ1**, the DITFEND framework coordinates with each base model and compares with it respectively on both the Chinese dataset and English dataset. Experiment results are shown in Table 3 and Table 7 (Appendix A.1). The experimental results further reveal several insightful observations:

- DITFEND can coordinate with various base models. From the results compared with these base models, DITFEND can mostly bring improvements of the fake news detection performance on target domains, which indicates DITFEND has satisfying compatibility.

- The improvement on the English dataset is bigger than the Chinese dataset. The main reason could be that the English dataset contains more samples than Chinese dataset, which can bring benefits for knowledge transfer. During target domain adaptation, the model can learn more transferable information about the target domain.

- We find that for the same target domains (Health and Politics) in Ch-3 dataset and Ch-9 dataset, the performance of the target domain in Ch-9 dataset exceeds the former with most base models, while on the other hand, for the same target domain (Health) in Ch-6 and Ch-9, the performances are too close to call, which indicates that more domains could bring more transferable knowledge to some extent, but the number of domain is not proportionate with performance.

4.5 Analysis (EQ2)

To answer **EQ2**, we conduct both an ablation study and a case study. In the ablation study, we evaluate the effectiveness of domain-level transfer and instance-level transfer respectively; In case study,

we aim to intuitively illustrate that our transferability quantifying strategy can pick out related source news pieces according to one specific domain.

Table 4: Ablation study on Chinese 9-domain dataset with RoBERTa and EDDFN.

Methods	target: Health			target: Politics		
	F1	AUC	Acc	F1	AUC	Acc
RoBERTa	0.9090	0.9611	0.9090	0.8366	0.9034	0.8637
+*DITFEND*	**0.9115**	**0.9739**	**0.9115**	**0.8775**	**0.9199**	**0.8982**
w/o meta	0.9096	0.9654	0.9087	0.8557	0.9144	0.8795
w/o sources	0.9100	0.9712	0.9100	0.8687	0.9183	0.8928
EDDFN	0.9379	0.9807	0.9380	0.8478	0.9292	0.8754
+*DITFEND*	**0.9399**	**0.9821**	**0.9400**	**0.8507**	**0.9308**	0.8772
w/o meta	0.9380	0.9810	0.9380	0.8480	0.9295	0.8760
w/o sources	0.9385	0.9815	0.9383	0.8482	0.9302	**0.8798**

4.5.1 Ablation Study

We evaluate two ablation experiment groups based on two representative fake news detection base models to evaluate the effectiveness of different modules in DITFEND framework. Table 4 shows the experimental results. To evaluate the effectiveness of domain-level transfer, we replace the meta-training procedure with classical training and use the weighted source data and target data to train the general model to adapt to the target domain. From the results, we can see that *w/o meta-training* performs worse than the whole DITFEND framework but better than the original model. To evaluate the effectiveness of instance-level transfer, we abandon instances from source domains when performing domain adaptation. The ablation results show that without instances from source domains, performance on target domains drops a little bit. Comparing the degree of decline of the two ablation experiments, we can also conclude that meta-training procedure brings more boost compared to instance from source domains.

4.5.2 Case Study

To further verify whether our transferability quantifier can pick out useful samples from source domains effectively, we conduct a case study on both the Chinese dataset and the English dataset. For each dataset, we have two assigned target domains (i.e., *Health* and *Politics* for Chinese, *COVID* and *PolitiFact* for English). According to Section 3.2, each instance has two different transferability indicators depending on two different target domains. Therefore, we use the corresponding indicator to choose samples relevant to the target domain, which can illustrate the interpretability of

instance-level transfer to some extent. Representative samples are listed in Table 5. Taking *Politics* domain (as the target domain) in the Chinese dataset for example, we successfully pick out some instances from other domains, which contain some knowledge related to *Politics*.

Table 5: Some representative examples selected depending on Transferability Quantifier. Words or phrases underlined are some indicators related to the target domain.

Target: Health
I. The team will undergo nucleic acid tests in the coming days as they resume individual training. [entertainment]   II. No one was injured when a test carriage derailed. [disaster]   III. Bask in inflatable dolls caused police cars and ambulances all out in broad daylight. [society]
Target: Finance
I. It has become the first Tencent smart medical joint innovation base in China. [health]   II. The central bank claimed to accelerate finance-technology regulatory framework. [science]   III. There are barriers to electronic use for older people in digital lifestyle. [society]
Target: Politics
I. Trump adviser urges Taiwan to come forward with advanced weapons program. [military]   II. Shandong provincial government and Tencent signed a strategic cooperation agreement. [Finance]   III. Beijing municipal government announces the hukou score to the society every year. [society]

4.6 Online Tests (EQ3)

To verify the real benefits of DITFEND bringing to the online system, we conduct online testing experiments. Different from the offline datasets, this online dataset is much more skewed (30,977 real: 774 fake $\approx$ 40:1). The testing set contains all news pieces in a whole month (ranging from 2021/10/10 to 2021/11/10), and the training set contains news pieces published before 2021/10/10. Both the training set and the test set come from the same online system[4]. The DITFEND coordinates with two online baselines: TextCNN and EANN, respectively. In real-world scenarios, the number of fake news items is much smaller than real news items, which means that we should detect as many fake news items as possible without misclassifying real news items. In other words, the objective is to improve the True Positive Rate (TPR) on the basis of a low False Positive Rate (FPR). Thus, beyond AUC and F1, we follow (McClish, 1989; Zhu

---

et al., 2020) and adopt standardized partial AUC ($SPAUC_{FPR \leq 0.1}$). The online results in Table 6 demonstrate that DITFEND achieves a significant improvement on AUC and SPAUC against the baselines, with greater improvements in SPAUC.

Table 6: Online test performance. The results below are a relative improvement over the two baseline models.

Improvement on	SPAUC	AUC
TextCNN	+2.07%	+1.40%
EANN	+3.40%	+2.90%

5 Conclusion and Future Work

Conclusion. In this paper, we propose DITFEND, a Domain- and Instance-level Transfer framework to improve fake news detection in an influential domain. For domain-level transfer, we adopt meta-learning to learn common knowledge across domains, in which way we can learn a general model; For instance-level transfer, we transform the problem of transferability evaluation to the task of character/word prediction by a language model endowed with target knowledge. Experiments on the Chinese dataset and English dataset demonstrate the effectiveness of our DITFEND framework over several fake news detection models. Ablation studies and case studies further evaluate the effectiveness and interpretability of each module in DITFEND. Online testing results verified the practical performance of our DITFEND framework.

Future Work. (1) In this work, the domain label of the target domain is unknown in advance. We will investigate how to handle the situation when the domain label of the target domain and/or source domains is unknown; (2) Temporal distribution shift is more challenging, and we plan to investigate how to adapt fake news detection models to the data in the future; (3) We will explore how to select source domains based on the analysis of transferability between different domains.

Acknowledgements

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.

Yejin Bang, Etsuko Ishii, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2021. Model generalization on covid-19 fake news detection. In *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 128–140. Springer.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Steven Bird. 2006. NLTK: The Natural Language Toolkit.

Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):1–14.

Leonardo Bursztyn, Aakaash Rao, Christopher P Roth, and David H Yanagizawa-Drott. 2020. Misinformation during a pandemic. Technical report, National Bureau of Economic Research.

Qingqing Chen. 2020. Coronavirus rumors trigger irrational behaviors among chinese netizens. https://www.globaltimes.cn/content/1178157.shtml. 2020.

Zhendong Chen, Siu Cheung Hui, Fuzhen Zhuang, Lejian Liao, Fei Li, Meihuizi Jia, and Jiaqi Li. 2022. Evidencenet: Evidence fusion network for fact verification. In *Proceedings of the ACM Web Conference 2022*, pages 2636–2645.

Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. DE-TERRENT: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 492–502. Association for Computing Machinery.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Enyan Dai, Yiwei Sun, and Suhang Wang. 2020. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 853–862.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Timothy M Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J Storkey. 2021. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yinqiu Huang, Min Gao, Jia Wang, and Kai Shu. 2021. Dafd: Domain adaptation framework for fake news detection. In *International Conference on Neural Information Processing*, pages 305–316. Springer.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, Yu Wang, and Jiebo Luo. 2017. Detection and analysis of 2016 US presidential election related rumors on twitter. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, pages 14–24. Springer.

Zhiwei Jin, Juan Cao, Yu-Gang Jiang, and Yongdong Zhang. 2014. News credibility evaluation on microblog with a hierarchical propagation model. In *2014 IEEE international conference on data mining*, pages 230–239. IEEE.

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

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart:

Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2018. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 3490–3497.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. *arXiv preprint arXiv:2011.04088*.

Yang Liu and Yi-Fang Brook Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Thirty-second AAAI Conference on Artificial Intelligence*.

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

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1930–1939.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3818–3824. AAAI Press.

Donna Katzman McClish. 1989. Analyzing a portion of the roc curve. *Medical decision making*, 9(3):190–195.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. MDFEND: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery.

Christof Naumzik and Stefan Feuerriegel. 2022. Detecting false rumors from retweet dynamics on social media. In *Proceedings of the ACM Web Conference 2022*, pages 2798–2809.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Minlong Peng, Qi Zhang, Yu-gang Jiang, and Xuan-Jing Huang. 2018. Cross-domain sentiment classification with target domain specific information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2505–2513.

Piotr Przybyla. 2020. Capturing the style of fake news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 490–497. AAAI Press.

Zhen Qin, Yicheng Cheng, Zhe Zhao, Zhe Chen, Donald Metzler, and Jingzheng Qin. 2020. Multitask mixture of sequential experts for user activity streams. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3083–3091.

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

Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom out and observe: News environment perception for fake news detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4543–4556, Dublin, Ireland. Association for Computational Linguistics.

Qiang Sheng, Xueyao Zhang, Juan Cao, and Lei Zhong. 2021. Integrating pattern-and fact-based fake news detection via model preference learning. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1640–1650, Virtual Event, Queensland, Australia. Association for Computing Machinery.

Anqi Shi, Zheng Qu, Qingyao Jia, and Chen Lyu. 2020. Rumor detection of covid-19 pandemic on online social networks. In *2020 IEEE/ACM Symposium on Edge Computing (SEC)*, pages 376–381. IEEE.

Baoxu Shi and Tim Weninger. 2016. Fact checking in heterogeneous information networks. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 101–102.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. dEFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 395–405, Anchorage, AK, USA. Association for Computing Machinery.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 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.

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

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

Craig Silverman. 2016. This analysis shows how viral fake election news stories outperformed real news on facebook. Accessed: November, 2016.

Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. 2020. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Fourteenth ACM Conference on Recommender Systems*, pages 269–278.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 849–857, London, United Kingdom. Association for Computing Machinery.

Yaqing Wang, Fenglong Ma, Haoyu Wang, Kishlay Jha, and Jing Gao. 2021. Multimodal emergent fake news detection via meta neural process networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3708–3716.

Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM Web Conference 2022*, pages 2501–2510.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7298–7309.

Dani Yogatama, Yishu Miao, Gabor Melis, Wang Ling, Adhiguna Kuncoro, Chris Dyer, and Phil Blunsom. 2018. Memory architectures in recurrent neural network language models. In *International Conference on Learning Representations*.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*, pages 3465–3476, Ljubljana, Slovenia. Association for Computing Machinery.

Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3205–3212.

Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. 2022. Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*.

Yongchun Zhu, Dongbo Xi, Bowen Song, Fuzhen Zhuang, Shuai Chen, Xi Gu, and Qing He. 2020. Modeling users' behavior sequences with hierarchical explainable network for cross-domain fraud detection. In *Proceedings of The Web Conference 2020*, pages 928–938.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.

A Supplementary Experiments

A.1 Experiments on Chinese 6-domain and Chinese 9-domain Dataset

We list our experimental results on the Chinese 6-domain and Chinese 9-domain datasets in Table 7.

A.2 Experiments on Other Target Domains

In this section, we set other domains, which have not been chosen as target domains in Section 4, and evaluate the performance of the proposed DIT-FEND on these domains. Experimental results are shown in Table 8.

Table 7: Performance comparison of base models with and without DITFEND on Chinese 6-domain dataset and Chinese 9-domain dataset.

Method	Ch-6 dataset						Ch-9 dataset					
	target: Health			target: Finance			target: Health			target: Politics		
	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc
BiGRU	0.8626	0.9442	0.8630	0.8254	0.9281	0.8642	**0.8868**	0.9574	**0.8870**	0.8356	0.9119	0.8590
+*DITFEND*	**0.8991**	**0.9676**	**0.8992**	**0.8431**	**0.9343**	**0.8755**	0.8789	**0.9630**	0.8794	**0.8749**	**0.9334**	**0.8925**
TextCNN	0.8832	**0.9670**	0.8835	**0.8646**	**0.9498**	**0.8943**	0.8768	0.9556	0.8770	0.8561	0.9225	0.8813
+*DITFEND*	**0.8922**	0.9631	**0.8925**	0.8581	0.9436	0.8887	**0.9093**	**0.9742**	**0.9094**	**0.8768**	**0.9235**	**0.8962**
RoBERTa	0.9100	0.9644	0.9100	0.8700	**0.9553**	0.8989	0.9090	0.9611	0.9090	0.8366	0.9034	0.8637
+*DITFEND*	**0.9175**	**0.9769**	**0.9175**	**0.8768**	0.9538	**0.9044**	**0.9115**	**0.9739**	**0.9115**	**0.8775**	**0.9199**	**0.8982**
EANN	0.9150	0.9761	0.9150	**0.8621**	**0.9483**	0.8906	0.9150	0.9762	0.9150	0.8705	0.9176	0.8918
+*DITFEND*	**0.9224**	**0.9767**	**0.9225**	0.8538	0.9466	**0.8918**	**0.9250**	**0.9797**	**0.9250**	**0.8822**	**0.9367**	**0.9035**
MMOE	0.9260	0.9754	0.9260	0.8546	0.9501	0.8887	**0.9364**	**0.9774**	**0.9365**	0.8620	0.9314	0.8842
+*DITFEND*	**0.9271**	**0.9756**	**0.9271**	**0.8611**	**0.9563**	**0.8906**	0.9199	0.9655	0.9200	**0.8633**	**0.9357**	**0.8860**
MOSE	0.9118	0.9720	0.9120	**0.8639**	0.9500	**0.8921**	0.9179	0.9700	**0.9160**	0.8673	0.9388	0.8918
+*DITFEND*	**0.9200**	**0.9735**	**0.9200**	0.8502	**0.9533**	0.8830	0.9050	**0.9725**	0.9050	**0.8681**	**0.9422**	**0.8972**
EDDFN	0.9280	0.9774	0.9280	0.8456	0.9436	0.8830	0.9379	0.9807	0.9380	0.8478	0.9292	0.8754
+*DITFEND*	**0.9350**	**0.9821**	**0.9350**	**0.8801**	0.9434	**0.9019**	**0.9399**	**0.9821**	**0.9400**	**0.8507**	**0.9308**	**0.8772**
MDFEND	0.9430	0.9851	0.9430	0.8749	0.9610	0.9023	0.9425	0.9846	0.9425	0.8774	0.9370	0.8994
+*DITFEND*	**0.9451**	**0.9876**	**0.9450**	**0.8800**	**0.9649**	**0.9100**	**0.9500**	**0.9864**	**0.9500**	**0.8986**	**0.9541**	**0.9181**

Table 8: Performance comparison (f1-score) of base models with and without DITFEND when assigning other target domains on Chinese 9-domain dataset.

target domain	science	military	Edu.	Disaster	Politics	Health	Finance	Ent.	Society
BiGRU	0.7269	0.8724	0.8138	0.7935	0.8356	**0.8868**	0.8291	0.8629	**0.8485**
+*DITFEND*	**0.8061**	**0.8604**	**0.9089**	**0.8017**	**0.8749**	0.8789	**0.8665**	**0.8840**	0.8483
TextCNN	0.7254	0.8839	0.8362	0.8222	0.8561	0.8768	0.8638	0.8456	0.8540
+*DITFEND*	**0.7283**	**0.8882**	**0.8829**	**0.8301**	**0.8768**	**0.9093**	**0.8665**	**0.8840**	**0.8769**
RoBERTa	0.7777	0.9072	0.8331	0.8512	0.8366	0.9090	0.8735	0.8769	0.8577
+*DITFEND*	**0.8107**	**0.9129**	**0.8384**	**0.8574**	**0.8775**	**0.9115**	**0.8889**	**0.8889**	**0.8881**
EANN$_T$	0.8225	0.9274	**0.8624**	0.8666	0.8705	0.9150	0.8710	0.8957	**0.8877**
+*DITFEND*	**0.8271**	**0.9419**	0.8582	**0.9060**	**0.8833**	**0.9250**	**0.8833**	**0.9100**	0.8874
MMOE	**0.8755**	0.9112	0.8706	0.8770	0.8620	**0.9364**	0.8567	0.8886	0.8750
+*DITFEND*	0.8587	**0.9186**	**0.8769**	**0.8836**	**0.8792**	0.9350	**0.8786**	**0.8923**	**0.8895**
MOSE	0.8502	0.8858	0.8815	**0.8672**	0.8808	0.9179	0.8672	0.8913	0.8729
+*DITFEND*	**0.8656**	**0.9121**	**0.8918**	0.8627	**0.8897**	**0.9200**	**0.8846**	**0.8987**	**0.8832**
EDDFN	**0.8186**	0.9137	0.8676	0.8786	0.8478	0.9379	0.8636	0.8832	0.8689
+*DITFEND*	0.8061	0.8834	**0.8887**	**0.8855**	**0.8492**	**0.9399**	**0.8647**	**0.8889**	**0.8754**
MDFEND	0.8301	0.9389	0.8917	0.9003	0.8865	0.9400	0.8951	0.9066	0.8980
+*DITFEND*	**0.8426**	**0.9419**	**0.9091**	**0.9145**	**0.8895**	**0.9550**	**0.8984**	**0.9100**	**0.8991**

A.3 Experiments on New Target Domains

In this section, we perform experiments to exclude target domain data when training the general model, in order to testify how our proposed framework performs when the target domain is unseen during general model training procedure.

We show the experimental results in Table 9 and Table 10. From the experimental results, we can find that our framework DITFEND can bring significant improvements to all domains.

Table 9: Performance comparison of base models with and without DITFEND on English dataset and Chinese 3-domain dataset when the target domain is unseen during meta-training procedure.

| Method | English dataset | | | | | | Ch-3 dataset | | | | | |
| | target: COVID | | | target: Politifact | | | target: Health | | | target: Politics | | |
	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc
BiGRU	0.7448	0.9114	0.8606	0.7339	0.8213	0.7375	0.8577	0.9367	0.8580	0.8384	0.9032	0.8637
+DITFEND	**0.8949**	**0.9650**	**0.8950**	**0.8406**	**0.9052**	**0.8406**	**0.8746**	**0.9609**	**0.8750**	**0.8633**	**0.9316**	**0.8889**
TextCNN	0.8322	0.9397	0.8955	0.7040	0.8046	0.7064	0.8716	0.9552	0.8720	0.8579	0.9067	**0.8859**
+DITFEND	**0.8549**	**0.9658**	**0.9042**	**0.7851**	**0.8632**	**0.7850**	**0.8949**	**0.9650**	**0.8950**	**0.8587**	**0.9615**	0.8810
RoBERTa	0.9014	0.9770	0.9377	0.7967	0.9078	0.7989	0.8955	0.9641	0.8955	0.8300	0.8948	0.8628
+DITFEND	**0.9305**	**0.9800**	**0.9509**	**0.8493**	**0.9335**	**0.8497**	**0.9250**	**0.9743**	**0.9250**	**0.8302**	**0.9027**	**0.8655**
EANN_T	0.8836	0.9751	0.9282	0.7558	0.8612	0.7584	0.9189	**0.9787**	0.9190	0.8405	0.9074	0.8690
+DITFEND	**0.9507**	**0.9896**	**0.9667**	**0.8664**	**0.9597**	**0.8671**	**0.9200**	0.9768	**0.9200**	**0.8615**	**0.9177**	**0.8889**
MMOE	0.9379	0.9883	**0.9588**	0.8477	0.9408	0.8486	0.9215	0.9679	0.9215	**0.8779**	**0.9388**	**0.8982**
+DITFEND	0.9356	**0.9907**	0.9586	**0.8542**	**0.9482**	**0.8555**	**0.9299**	**0.9735**	**0.9300**	0.8659	0.9331	0.8896
MOSE	0.9326	0.9879	0.9588	0.8576	0.9447	0.8590	**0.9023**	0.9683	**0.9025**	0.8564	**0.9138**	0.8795
+DITFEND	**0.9547**	**0.9889**	**0.9692**	**0.8725**	**0.9528**	**0.8728**	0.9010	**0.9685**	0.9012	**0.8568**	0.9137	**0.8812**
EDDFN	0.9306	0.9891	0.9547	0.8505	0.9432	0.8509	**0.9235**	0.9735	**0.9235**	0.8440	0.9207	**0.8702**
+DITFEND	**0.9383**	**0.9899**	**0.9586**	**0.8721**	**0.9466**	**0.8728**	0.9150	**0.9762**	0.9150	**0.8441**	**0.9337**	0.8663
MDFEND	0.9331	0.9874	0.9565	0.8473	0.9391	0.8485	0.9419	0.9855	0.9420	**0.8555**	0.9259	**0.8854**
+DITFEND	**0.9482**	**0.9932**	**0.9659**	**0.8599**	**0.9503**	**0.8613**	**0.9550**	**0.9877**	**0.9550**	0.8507	**0.9442**	0.8772

Table 10: Performance comparison of base models with and without DITFEND on Chinese 6-domain dataset and Chinese 9-domain dataset when the target domain is unseen during meta-training procedure.

| Method | Ch-6 dataset | | | | | | Ch-9 dataset | | | | | |
| | target: Health | | | target: Finance | | | target: Health | | | target: Politics | | |
	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc
BiGRU	0.8626	0.9442	0.8630	0.8254	0.9281	0.8642	**0.8868**	0.9574	**0.8870**	0.8356	0.9119	0.8590
+DITFEND	**0.8875**	**0.9548**	**0.8875**	**0.8400**	**0.9295**	**0.8710**	0.8700	**0.9612**	0.8715	**0.8541**	**0.9305**	**0.8816**
TextCNN	0.8832	**0.9670**	0.8835	**0.8646**	**0.9498**	**0.8943**	0.8768	0.9556	0.8770	0.8561	0.9225	0.8813
+DITFEND	**0.8901**	0.9608	**0.8902**	0.8547	0.9431	0.8814	**0.8854**	**0.9715**	**0.8855**	**0.8650**	**0.9230**	**0.8879**
RoBERTa	0.9100	0.9644	0.9100	0.8700	**0.9553**	0.8989	0.9090	0.9611	0.9090	0.8366	0.9034	0.8637
+DITFEND	**0.9249**	**0.9666**	**0.9250**	**0.8781**	0.9347	**0.9057**	**0.9149**	**0.9723**	**0.9150**	**0.8999**	**0.9294**	**0.9181**
EANN_T	0.9150	**0.9761**	0.9150	0.8621	0.9483	0.8906	0.9150	**0.9762**	0.9150	0.8705	0.9176	0.8918
+DITFEND	**0.9184**	0.9753	**0.9184**	**0.8789**	**0.9649**	**0.9094**	**0.9196**	0.9676	**0.9196**	**0.8733**	**0.9178**	**0.8940**
MMOE	**0.9260**	**0.9754**	**0.9260**	0.8546	0.9501	0.8887	**0.9364**	**0.9774**	**0.9365**	**0.8620**	0.9314	0.8842
+DITFEND	0.9145	0.9732	0.9150	**0.8548**	**0.9503**	**0.8888**	0.9200	0.9663	0.9200	0.8496	**0.9459**	**0.8843**
MOSE	**0.9118**	0.9720	**0.9120**	0.8639	0.9500	0.8921	**0.9179**	0.9700	**0.9160**	**0.8673**	**0.9388**	**0.8918**
+DITFEND	0.9110	**0.9725**	0.9112	**0.8640**	**0.9512**	**0.8925**	0.9048	**0.9715**	0.9049	0.8540	0.9386	0.8713
EDDFN	0.9280	**0.9774**	0.9280	0.8456	0.9436	0.8830	0.9379	**0.9807**	0.9380	**0.8478**	**0.9292**	**0.8754**
+DITFEND	**0.9299**	0.9745	**0.9300**	**0.8471**	**0.9497**	**0.8868**	**0.9399**	0.9768	**0.9400**	0.8240	0.8951	0.8596
MDFEND	0.9430	0.9851	0.9430	0.8749	**0.9610**	0.9023	0.9425	0.9846	0.9425	**0.8774**	0.9370	**0.8994**
+DITFEND	**0.9450**	**0.9900**	**0.9450**	**0.8792**	0.9486	**0.9057**	**0.9600**	**0.9892**	**0.9600**	0.8713	**0.9452**	0.8947