Cross-Domain Fake News Detection based on Dual-Granularity Adversarial Training

Wenjie Wei[1], Yanyue Zhang[1], Jinyan Li[2], Panfei Liu[2], Deyu Zhou[*1]
[1]School of Computer Science and Engineering, Southeast University, China
[2]Huawei Technologies Co., Ltd
{230228509, yanyuez98, d.zhou}@seu.edu.cn, ljyjy86@hotmail.com, panfeil@nuaa.edu.cn

Abstract

Cross-domain fake news detection, aiming to detect fake news in unseen domains, has achieved promising results with the help of pre-trained language models. Existing approaches mainly relied on extracting domain-independent representations or modeling domain discrepancies to achieve domain adaptation. However, we found that the relationship between entities in a piece of news and its corresponding label (fake or real) fluctuates among different domains. Such discrepancy is ignored by existing methods, leading to model entity bias. Therefore, in this paper, we propose a novel cross-domain fake news detection method based on dual-granularity adversarial training from the perspective of document-level and entity-level. Specifically, both the news pieces and their entities are modeled individually to construct an encoder that can generate domain-independent representations using adversarial training. Moreover, the dual-granularity soft prompt, consisting of two independent learnable segments trained on the source domains, is employed to make the model easily adapt to the unseen target domains. In addition, MultiFC, a released dataset for cross domain fake news detection, is not suitable for the evaluation due to its unreasonable domain construction rules. We artificially reconstructed the dataset and named it New-MultiFC, which is a more domain-discriminative dataset. Experimental results on both the newly constructed New-MultiFC and FND3 show the effectiveness of the proposed approach, achieving the state-of-the-art results in unseen domains.

1 Introduction

With the vigorous development of social platforms, people are more inclined to express opinions or consult information on the Internet. However, the increasing information is accompanied by more fake news, which has caused negative impacts on our lives. Therefore, it is crucial to detect fake news automatically and precisely.

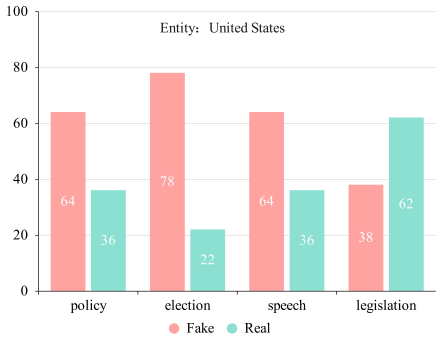


Figure 1: The label distribution of two named entities in four domains in the New-MultiFC dataset.

Previous work mainly focused on single-domain fake news detection, and achieved good results, especially after the application of pre-training models such as BERT (Devlin et al., 2018). However, in the actual scenario, the news may come from multiple domains, with discrepancies in text semantics, such as writing style, word and label distribution. This may cause models trained on source domains to fail to adapt to unknown target domains. In order to make the model more adaptable on target domains, some fake news detection approaches have focused on cross-domain scenarios, which can be classified into two categories: domain-independent-representations based methods (Wang et al., 2018; Hardalov et al., 2021; Mosallanezhad et al., 2022; Wu and Shi, 2022) and domain-differences-modeling based methods (Tang et al., 2020; Nan et al., 2022b,a; Zhu et al., 2022b). The former focuses on extracting the independent representations to remove the discrepancies between domains, while the latter devotes to model the discrepancies between domains to achieve the improved performance in unseen domains.

Although these works have achieved certain re-

* Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 9407–9417
January 19–24, 2025.

sults in solving the domain adaptability problem on fake news detection tasks, they ignore the discrepancy of specific words in different domains, such as entities.

Through statistics, we found that the relationship between a certain entity in a piece of news and its corresponding label (fake or real) fluctuates among different domains. As shown in Figure 1, in domain policy, election and speech from the New-MultiFC dataset, about 61% of news pieces containing 'United States' are fake, on average. However, in domain legislation, only 38% of news pieces containing 'United States' are fake. Without modeling the fluctuation, models trained on the first three domains might predict the news containing 'United States' in domain legislation as fake with high probability.

To solve the problem mentioned, we propose cross-domain **FA**ke News **DE**tection based on **D**ual-Granularity Adversarial Training, which achieves domain adaptation from the dual-granularity perspective (document-level and entity-level). Following domain-independent-representations based methods (Wang et al., 2018; Hardalov et al., 2021), FADED adopts the domain adversarial training framework to achieve domain adaptation. In detail, FADED adopts domain adversarial training for both text semantics and entity words to obtain an encoder that can extract domain-independent representations for text and entities. Besides, inspired by the approach of Wu and Shi (2022), which utilizes prompts to make model adapt to unseen domains, we introduced prompts to the model from two perspectives for cross-domain fake news detection. In detail, we propose dual-granularity soft prompt, which uses two independent tunable vectors as prompts for text and entities, respectively.

In addition, based on the MultiFC dataset which is a challenging fake news detection task released by Augenstein et al. (2019), we constructed a new dataset called New-MultiFC. The domain division rules on the MultiFC dataset are source-based, that is, samples from different websites belong to different domains. As an example from MultiFC dataset shown in Table 1, although texts on different websites are written by different authors, they may describe the same topic and have similar language styles. The basis for dividing domains in cross-domain scenarios is mainly based on topics, which ensures the existence of differences among domains. Therefore, we manually divide the MultiFC based on topics to construct a more domain-discriminative dataset, which will be described in Section 4 and Appendix A to B.

Our main contributions can be summarized as:

1) We propose FADED, which achieves domain adaptation from the perspective of dual-granularity: document-level and entity-level.

2) We introduce New-MultiFC, a more domain-discriminative dataset reconstructed on MultiFC divided by topics.

3) Experimental results on two datasets show that FADED outperforms the current State-Of-The-Art approaches.

2 Related Work

Cross-domain fake news detection aims to improve the performance of models to detect fake news in unseen domains. Existing methods can be generally grouped into two clusters: domain-independent-representations based methods and domain-differences-modeling based methods.

Domain-independent-representations based methods focus on extracting the independent representations to remove the discrepancies between domains. Wang et al. (2018) proposed an end-to-end framework - Event Countering Neural Network (EANN), which can extract the invariant features of events, thus facilitating the detection of fake news. Hardalov et al. (2021) proposed a novel framework (MoLE) that combines domain-adaptation and label embeddings for learning heterogeneous target labels. Based on adversarial framework, Mosallanezhad et al. (2022) proposed to utilize reinforcement learning process, which removes domain differences from the representation space. Recently, some efforts are devoted to adopts prompts in cross-domain news detection, such as the model proposed by Wu and Shi (2022), which adopts separate soft prompt instead of hard templates to learn different vectors for different domains based on adversarial training.

Domain-differences-modeling based methods mainly focus on modeling the discrepancies between domains to achieve the improved performance in unseen domains. For knowledge transfer from source domains to target domains, Nan et al. (2022b) proposed a pipeline method which first trains a general model with data of all domains and then utilizes confusion values calculated on MLM task to evaluate the transferability of each sample. For modeling the discrepancies between domains,

Domain	Website	Sample	Topic
abbc	abc	The claim: Environment Minister Greg Hunt says the Coalition's emissions reduction fund, at $13.95 per tonne of carbon, is around 1 per cent of the cost of reducing carbon under the former Labor government's carbon pricing scheme, which he says cost $1,300 a tonne.	policy
vogo	voiceofsandiego	Statement: "In total, the City Auditor's Office identified $7,425,271 in potential monetary recoveries and cost savings for the City, which equates to $3 in potential savings for every $1 of audit costs," wrote Eduardo Luna, San Diego auditor, in a memorandum to City Hall's audit committee.	policy
pomt	politifact_stmt	The nonpartisan CBO, Congressional Budget Office, has said that the No. 1 policy decision that brought us to the need to prevent the nation from defaulting on our debt for the first time in history were the Bush tax cuts in 2001 and 2003.	policy

Table 1: An example from MultiFC dataset, where samples come from different domains but share the same topic.

Nan et al. (2022a) also proposed a novel model, which adopts a domain-gate to select useful experts of MoE (Jacobs et al., 1991). Zhu et al. (2022b) encoded the input from a multi-view perspective, and predict its probability distribution in domain space to obtain the relationships between domains.

Our work mainly falls into the first group, which improves the performance in unseen fake news detection domains by extracting domain-independent representations. Inspired by the work of Zhu et al. (2022a), we find that not only text semantics between different domains have discrepancies, but also the correlations between entities and labels, which may induce models to capture the entity bias in domains. Thus, our proposed FADED model removes text semantic discrepancies and entity bias from dual-granularity: document-level and entity-level. Following Wu and Shi (2022), FADED adopts dual-granularity soft prompt, which consist of two parts of prompts that learned for text and entity respectively, while conducting domain adversarial training for both text and entities to extract domain-independent representations. In addition, based on a challenging fake news detection dataset MultiFC which may not have an explicit distinction in domains due to unreasonable domain division rules, we re-divided it based on topics to construct a more domain-discriminative dataset called NewMultiFC.

3 Method

3.1 Problem Statement

Given a news piece with $|x|$ words as $x = \{x_1, x_2, ..., x_{|x|}\}$, its entity set $e = \{e_1, e_2, ..., e_{|e|}\}$ and relevant evidences collected are denoted as $T = \{T_1, T_2, ..., T_{|T|}\}$, in which each evidence is represented by $t = \{t_1, t_2, ..., t_{|t|}\}$. Each news piece has a ground-truth label $y \in \{0, 1\}$, where 0 and 1 denote the new piece is fake and real, respectively. In addition, each news piece is categorized into a single domain with a domain label $d \in \{Domain_1, Domain_2, ..., Domain_n\}$, where n indicates the number of domains. Cross-domain fake news detection aims to train a model which can effectively detect fake news in unseen domains.

3.2 Overall Architecture

The overall architecture of the proposed FADED is shown in Figure 2, which is composed by five part (which are bolded and italicized in Figure 2): (1) Model Input; (2) BERT Encoder; (3) Fake News Classifier; (4) Dual-Granularity Domain Classifiers; (5) Domain Adversarial Training.

Given the input, the model detects whether the samples are fake news by the Fake News Classifier. Meanwhile, the model also predicts the domain label of each sample by the Dual-Granularity Domain Classifiers, which two take text and entity representations as input respectively. Then, the BERT Encoder and soft prompt are tuned through the domain adversarial training to remove text semantic differences and entity bias in domains. Moreover, effective prompts for text and entities are also learned.

The following sections will describe each component in detail except BERT Encoder, which is widely known.

3.3 Model Input

As shown in the lower half of the figure, the model input is mainly composed of four parts: dual-granularity soft prompt, news pieces, entities and related evidence set. It should be noted that dual-granularity soft prompt is split into two independent tunable segments for learning prompts in continuous space for text and entities, respectively. The

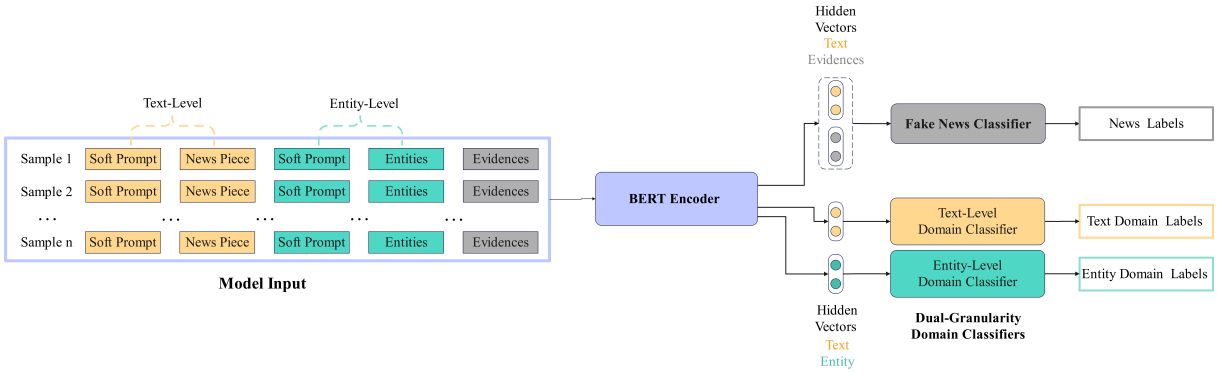


Figure 2: Overall structure of the proposed method.

model input is represented as:

$$\begin{aligned} input = & \big[E(\text{“}[CLS]\text{”}); \\ & \{v_1, v_2, ..., v_K\}; E(\{x_1, x_2, ..., x_{|x|}\}); \\ & \{v_1^{'}, v_2^{'}, ..., v_K^{'}\}; E(\{e_1, e_2, ..., e_{|e|}\}); \\ & E(\{T_1\}; \{T_2\}; ...; \{T_{|T|}\}); \\ & E(\text{“}[SEP]\text{”})\big] \end{aligned} \tag{1}$$

Where E(.) denotes the embedding function, $\{.\}$ represents the token set of a sentence, $(;)$ indicates the concatenate operation and $v = \{v_1, v_2, ..., v_K\}$, $v^{'} = \{v_1^{'}, v_2^{'}, ..., v_K^{'}\}$ indicates K tunable vectors as soft prompt for text and entities, respectively.

3.4 Fake News Classifier

The Fake News Classifier detects fake news based on evidence ranking, which aims to assign an importance distribution to evidences by learning the compatibility between a news piece and each evidence. It ranks evidences by their utility for the veracity prediction task, and then uses the resulting ranking to obtain a weighted combination of all news-evidence pairs. Since no direct labels are available to rank evidence, the model needs to use prediction accuracy to learn the importance distribution implicitly.

To combine the news and its evidences, we refer to the work of Mou et al., 2015 to joint the news and each evidence representation as follows:

$$Score_i = \big[h_x; h_{T_i}; h_x - h_{T_i}; h_x \cdot h_{T_i}\big] \tag{2}$$

Where h_x and h_{T_i} denote the representation (extracted from BERT Encoder) of the news piece x and its i-th evidence from the related evidence set. Here, for news piece and each relevant evidence, we use the representation of its [CLS] token as input to the model.

All joint news-evidence representations ($Score_i$) are then projected into a $|T|$-dimension vector via a fully connected layer FC, followed by a non-linear activation function f_1 and softmax, to obtain the importance distribution:

$$\begin{aligned} o = & softmax\Big[f_1(FC(Score_1)); \\ & ...; f_1(FC(Score_{|T|}))\Big] \end{aligned} \tag{3}$$

Final prediction for news piece x is then obtained by a non-linear projection layer $f_2(FC(.))$ and softmax, with the news text and weighted evidence representations as input:

$$P(y|x) = softmax\big[f_2(FC(h_x; h_T \cdot o))\big] \tag{4}$$

Where h_T denotes the related evidence set of the news piece x, and which can be represented as $h_T = \{h_{T_1}, h_{T_2}...h_{T_i}\}$.

3.5 Dual-Granularity Domain Classifiers

Given a news sample x, domain classifiers aim to predict its domain label $d \in \{Domain_1, Domain_2, ..., Domain_n\}$. In this paper, domain classifiers are mainly used for adversarial training to achieve domain adaptive.

Since there are no direct labels for evidences and some of them may come from different domains, we utilize only the news piece representation as input to classify its domain label:

$$P(d|x) = softmax[f(FC(h_x))] \tag{5}$$

Where f is a non-linear activation function.

Inspired by Zhu et al., samples in different domains have discrepancies not only in semantics representations, but also reflected in word usage,

emotional expression and writing styles. We consider that keywords in news, such as entities or events, may also have discrepancies in different domains. As shown in Figure 1, due data research we find that the correlations between entities and labels also have discrepancies. These discrepancies are ignored by existing methods, which may induce the model to detect fake news by using the entity bias in domains. Thus, to remove the text semantic discrepancies and entity bias in domains, we adopt Dual-Granularity Domain Classifiers during adversarial training, which is composed of two domain classifiers (document-level and entity-level) to predict the domain labels of samples with text and entities as input respectively.

Given a news piece x, its entity set $e = \{e_1, e_2, ..., e_{|e|}\}$ and domain label $d_x \in \{Domain_1, Domain_2, ..., Domain_n\}$, the learning objective of the two classifiers is to correctly predict the domain label:

$$argmaxP(d_x|x) = softmax[f_3(FC(h_x))] \quad (6)$$

$$argmaxP(d_x|e) = softmax[f_4(FC(h_e))] \quad (7)$$

Where h_e is the representation of the entity in the given news (extracted from BERT), f_3 and f_4 are non-linear activation functions.

3.6 Domain Adversarial Training

The main purpose of adopting the domain adversarial training is to remove the text semantic discrepancies and entity bias in different domains by training an encoder that can extract domain-independent representations for both text and entities. For different components of the model, we designed corresponding loss functions to achieve the purpose mentioned. The domain adversarial training is mainly realized by optimizing the loss function of both the encoder and soft prompt.

We use the following cross-entropy loss function to update fake news classifier:

$$\begin{aligned} L_{fake} = -\sum_{i=1}^{N} \big[& y_i \cdot log(P(y|x)) \\ & -(1-y_i) \cdot log(1-P(y|x)) \big] \end{aligned} \quad (8)$$

Where N is the number of news samples, and $y_i \in \{0, 1\}$ denotes the ground truth label ranging from 1 as the positive label and 0 as the negative label.

For document-level domain classifier, we use the following cross-entropy loss function to update the classifier:

$$L_{dom-t} = -\sum_{i=1}^{N}\sum_{j=1}^{M} \big[y_{ij} \cdot log(P(d_x|x)) \big] \quad (9)$$

Where M is the size of the domain label space, and $y_{ij} \in \{0, 1\}$ denotes the ground truth label ranging from 1 as the positive label and 0 as the negative label.

For entity-level domain classifier, we use the following cross-entropy loss function to update the classifier:

$$L_{dom-e} = -\sum_{i=1}^{N}\sum_{j=1}^{M} \big[y_{ij} \cdot log(P(d_e|e)) \big] \quad (10)$$

$y_{ij} \in \{0, 1\}$ also denotes the ground truth label ranging from 1 as the positive label and 0 as the negative label.

To train an encoder that can extract domain-independent representations for both text and entities, we update the BERT Encoder by optimizing the following loss function:

$$L_{enc} = L_{fake} - \lambda_1 L_{dom-t} - \lambda_2 L_{dom-e} \quad (11)$$

Where λ_1 and λ_2 are weight coefficients.

From the equation (11), the domain adversarial training can be seen as two-player minimax game where the domain classifiers tend to minimize the domain prediction loss so as to make the domain predictors strong, while the encoder tends to maximize the domain prediction loss so as to weaken the domain classifiers. Through the domain adversarial training to obtain an encoder which can extract representations that domain-independent.

Two vectors in continuous space are concatenated with samples as soft prompt for text and entities. During the domain adversarial training, they are updated with the following loss functions, respectively:

$$L_{pro-t} = L_{fake} + \mu_1 L_{dom-t} \quad (12)$$

$$L_{pro-e} = L_{fake} + \mu_2 L_{dom-e} \quad (13)$$

Where μ_1 and μ_2 are are weight coefficient. Unlike the encoder, two loss functions of soft prompt enable the adding vectors to learn text and entity prompts for fake news detection, while also containing some domain-related information. It brings a

Domain	Website	Insts	labels
ranz	radionz	21	2
bove	—	295	2
abbc	abc	436	3
huca	huffingtonpostca	34	3
mpws	mpnews	47	3
peck	pesacheck	65	3
faan	—	111	3
clck	climatefeedback	38	3
fani	—	20	3
chct	checkyourfact	355	4
obry	observatory	59	4
vees	verafiles	504	4
faly	—	111	5
goop	gossipcop	2943	6
pose	politifact_promise	1361	6
thet	theferret	79	6
thal	thejournal	163	7
afck	africacheck	433	7
hoer	hoaxslayer	1310	7
para	pandora	222	7
wast	washingtonpost	201	7
vogo	voiceofsandiego	654	8
pomt	politifact_stmt	15390	9
farg	factcheckorg	485	11
snes	snopes	6455	12
tron	truthorfiction	3423	27

Table 2: Total number of instances and unique labels per domain in MultiFC, as well as the websites where the data for each domain resides.

greater difficulty for removing the text semantic discrepancies and entity bias in domains, which may urge the model to train a more powerful encoder for extracting domain-independent representations for both text and entities.

4 Dataset Construction

Our improved new dataset is based on the original dataset-MultiFC, which is a proposed challenging task in fake news detection. Table 2 lists the statistics of MultiFC dataset, mainly including the number, labels, and source websites of each domain. As can be seen from the first two columns in the table, the domain division of MultiFC is based on the website, which is obviously different from the traditional topic-based division. Although the texts of different websites are written by different authors, they may describe the same topic and the discrepancies are not obvious. Therefore, in order to make the domains in MultiFC dataset more discriminative, we re-divided it based on topics and named the resulting new dataset New-MultiFC. More details are described in Appendix A and B.

5 Experiments

In this section, we conduct experiments to answer the following research questions:
RQ1 Does the proposed FADED outperform other methods on cross-domain datasets?
RQ2 What are the effects of different granularity and components in the proposed FADED?
RQ3 Does FADED effectively remove the text semantic discrepancies and entity bias in domains?

Domain	PolitiFact	GossipCop	CoAID
#Fake News	269	1269	135
#Real News	230	2466	1568

Table 3: The statistics of FND-3 dataset, which contains three domains: PolitiFact, GossipCop and CoAID.

5.1 Experimental Settings

5.1.1 Datasets

We conduct experiments on New-MultiFC dataset which under the manual divided method based on MultiFC dataset. Following Silva et al. (2021), we combine three well-known disinformation datasets: PolitiFact (Shu et al., 2020), GossipCop (Shu et al., 2020) and CoAID (Li et al., 2020) to produce a cross-domain news dataset, which we named FND-3. Table 3 shows the statistics of FND-3 dataset.

5.1.2 Experimental Details

We divide all datasets into train/ validation/ test sets with keep the domain distribution in each set. For the test set, we select all samples from unseen domains to ensure that the task is similar to the actual scenario. For parameter settings, following Devlin et al. (2018), we truncate the input length to 512 and set the vector dimension to 768. For each claim, we select 5 pieces from its evidence set as input, which are concatenated with claim. For soft prompt, we add 20 vectors in continuous space to the input as prompts. Both fake news classifier and two domain classifiers adopt a similar feed-forward neural network with a single hidden layer of 256 neurons. During training, the initial learning rate for the Adam optimizer (Kingma and Ba, 2014) is tuned by grid searches from 1e-6 to 1e-2.

5.1.3 Baselines

To demonstrate the effectiveness of our proposed model FADED, we compare it with several existing methods in three groups:

(1) Single-domain methods, which separately train models for each domain, including:

- **BiGRU** (Jing et al., 2016), is a widely used baseline for fake news detection. We adopt a one-layer BiGRU with a hidden size of 512.
- **TextCNN** (Kim, 2014), is an effective and commonly used text encoder. We implement TextCNN with 5 kernels and 64 channels.

- **BERT** (Devlin et al., 2018), is a pre-training model, which is widely used in various tasks and serves as a commonly baseline.

(2) Domain-differences-modeling based methods, which model the differences by learning inter domain weights, including:

- **MDFEND** (Nan et al., 2022a), is a recent cross-domain fake news detection model, which adopts a Domain-Gate to select useful experts of MoE (Jacobs et al., 1991).
- **M^{3}FEND** (Zhu et al., 2022b), is the latest cross-domain fake news detection model, which encodes the news piece from a multi-view perspective and adopts a Domain Memory Bank to enrich information for samples.

(3) Domain-independent-representations based methods, which focus on removing the differences by extracting the domain-independent representations, including:

- **EANN** (Wang et al., 2018), which proposes an Event Adversarial Neural Network to learn event-invariant representations.
- **REAL-FND** (Mosallanezhad et al., 2022), is a novel domain adversarial model based on reinforcement learning.
- **ASPT** (Wu and Shi, 2022), which is the basic model for our work, it adopts soft prompt tuned on adversarial training framework.

5.2 Model Comparison (RQ1)

We compare proposed FADED with eight baselines on both New-MultiFC and FND-3 dataset. For two corpus, we select one domain for testing and the remaining as the train set per test epoch. The main results are shown in Table 4 and 5, and we have the following observations:

(1) On all tasks, the poor performance of the single-domain approaches reflects their failure to achieve domain adaptation. Compared with single-domain group, the cross-domain methods have achieved significant performance improvement on both datasets, which shows that jointly training data of multiple domains is helpful for detecting fake news in unseen domains.

(2) For cross-domain methods, the third group outperforms the second in unseen domains, which demonstrates that domain-independent representations are more important for cross-domain fake news detection. The reason could be that it is more effective to extract domain-independent representations directly than learning the differences among domains due to the big gap between training and testing sets.

(3) Compared with all baselines, our proposed FADED achieves the best experimental results on both datasets, which shows the effectiveness of FADED model. The results have answered the first research question (**RQ1**) that FADED does outperform other methods on cross-domain datasets, and the contribution of each component is demonstrated in ablation study section.

5.3 Ablation Study (RQ2)

In this section, we analyze the effects of different granularity and components in our proposed FADED and conduct a ablation study on both New-MultiFC and FND-3 dataset with the average F1 score as shown in Table 6.

First, we conduct experiments to explore the contributions of different granularity by designing two kinds of models: w/o Text-Domain and w/o Entity-Domain, which remove document-level and entity-level domain adversarial training from FADED, respectively. From the results we find that both granularity are beneficial for cross-domain fake news detection, especially the document-level, which is also the core modeling object of most related methods (Wang et al., 2018; Mosallanezhad et al., 2022). The entity-level domain adversarial training also contributes to improve performance on detecting fake news, which shows the effectiveness of removing the entity bias in different domains.

Then, we conduct experiments to testify the effectiveness of dual-granularity soft prompt, which separated by half-to-half for text and entities. We introduce two kinds of models: w/o Text-Prompt and w/o Entity-Prompt, which remove text and entities soft prompt, respectively. The performance drop on removing either type of soft prompt indicates its effectiveness, and it also demonstrates that soft prompt set for different targets need to be separated from each other to prevent mutual influence, such as text and entities.

According to the ablation experiments above, the answers to the second research question (**RQ2**) are follows:

1. Both granularity have a great contribution for the performance improvement, especially the document-level.

Model	policy		election		legislation		meeting		speech		examination		others	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
BiGRU	0.498	0.509	0.467	0.479	0.481	0.492	0.480	0.489	0.439	0.450	0.423	0.431	0.472	0.483
TextCNN	0.513	0.521	0.484	0.493	0.496	0.505	0.509	0.518	0.477	0.488	0.455	0.469	0.500	0.508
BERT	0.679	0.698	0.637	0.648	0.631	0.644	0.668	0.685	0.646	0.661	0.611	0.623	0.676	0.689
MDFEND	0.695	0.713	0.647	0.663	0.642	0.657	0.674	0.691	0.652	0.670	0.631	0.645	0.686	0.695
M^{3}DFEND	0.693	0.712	0.652	0.668	0.640	0.654	0.677	0.690	0.659	0.674	0.627	0.640	0.690	0.698
EANN	0.704	0.724	0.658	0.66	0.650	0.666	0.680	0.697	0.674	0.690	0.630	0.647	0.691	0.707
REAL-FND	0.699	0.718	0.651	0.663	0.647	0.660	0.683	0.696	0.669	0.681	0.634	0.650	0.687	0.701
ASPT	0.713	0.736	0.670	0.679	0.660	0.677	0.690	0.701	0.678	0.693	0.667	0.679	0.706	0.722
FADED	**0.732**	**0.757**	**0.679**	**0.690**	**0.671**	**0.687**	**0.705**	**0.711**	**0.684**	**0.695**	**0.665**	**0.678**	**0.721**	**0.733**

Table 4: Results on the New-MultiFC dataset, tested on one domain and trained on the remaining domains.

Model	PolitiFact		GossipCop		CoAID	
	F1	AUC	F1	AUC	F1	AUC
BiGRU	0.407	0.431	0.426	0.456	0.457	0.489
TextCNN	0.422	0.444	0.453	0.481	0.439	0.476
BERT	0.665	0.678	0.532	0.544	0.703	0.710
MDFEND	0.707	0.714	0.576	0.591	0.738	0.750
M^{3}DFEND	0.712	0.719	0.591	0.601	0.735	0.747
EANN	0.714	0.717	0.585	0.599	0.733	0.746
REAL-FND	0.719	0.725	0.597	0.608	0.748	0.761
ASPT	0.721	0.730	0.601	0.609	0.744	0.760
FADED	**0.735**	**0.750**	**0.607**	**0.614**	**0.763**	**0.775**

Table 5: Results on the FND-3 dataset, tested on one domain and trained on the remaining domains.

Dataset	New-MultiFC		FND-3	
	F1	AUC	F1	AUC
FADED	**0.694**	**0.707**	**0.702**	**0.713**
w/o Text-Domain	0.657	0.670	0.661	0.674
w/o Entity-Domain	0.678	0.692	0.685	0.699
w/o Text-Prompt	0.683	0.695	0.686	0.697
w/o Entity-Prompt	0.687	0.700	0.691	0.704

Table 6: Results of ablation study.

2. Soft prompt for text and entities is also beneficial for cross-domain fake news detection.

5.4 Analysis (RQ3)

In this section, we conduct experiments to verify the effectiveness of the model in removing text semantic discrepancies and entity bias in domains. Considering that the goal of domain adversarial training is to extract the domain-independent representations, we conduct experiments on FND-3 dataset to testify whether the representations extracted by our proposed FADED are confusing for domain classifiers (we use two classifiers that have been trained on other domain classification tasks). The results are shown in Table 7, from which we have the following observations:

(1) For both document-level and entity-level domain classifiers, it's obviously easier to classify domain labels with the representations extracted by BERT as the input, which demonstrates that extracting domain-independent representations based on adversarial training is effective for removing text semantic discrepancies and entity bias in domains.

(2) Compared with the basic model ASPT, the representations extracted by FADED are far more confusing for entity-level domain classifier, which verifies that FADED has removed some entity bias in domains. Moreover, for document-level domain classifier, the accuracy of using the representations extracted by FADED as input is lower compared with ASPT model, which shows that removing the entity bias in domains may also eliminate some text semantic discrepancies.

Classifier	Cls-text	Cls-entity
BERT	0.935	0.832
ASPT	0.633	0.557
FADED	**0.589**	**0.307**

Table 7: The classification accuracy of the domain classifiers with the representations extracted from BERT, basic model ASPT and our proposed FADED as input.

(3) Comparing the experimental results of two granularity domain classifiers, we find that text semantic discrepancies are more difficult to remove. The reason could be that compared with entities, text have complex context and topics.

According to the experiments above, the answer to the last research question (**RQ3**) is that FADED effectively remove the text semantic discrepancies and entity bias in domains to some extent.

6 Conclusion

In this paper, we propose a novel cross-domain fake news detection method based on dual-granularity adversarial training named FADED, which achieves domain adaptive from the perspective of dual-granularity: document-level and entity-level. Based on adversarial training, FADED removes text semantic discrepancies and entity bias in domains. Besides, it adopts soft prompt, which composed by two parts of prompts that learned for text and entities, respectively. A new dataset named New-MultiFC is also proposed, which builds on MultiFC dataset by dividing the domain based on topics. On two cross-domain fake news detection datasets, our proposed FADED has achieved the state-of-the-art results in unseen domains.

Limitations

As shown in equation (5), due to no direct labels for related evidences, we utilize only the news piece representation as input to classifier the domain label of a certain sample. However, the news and its evidences may belong to different domains, which is not considered in this paper. If the news and relevant evidences cannot be modeled separately, this may lead to the failure of the model to effectively remove the domain discrepancies for the evidences. Therefore, in future work, we plan to explore whether domain discrepancies are exist in relevant evidences and how to remove these discrepancies if exist.

References

I. Augenstein, C. Lioma, D. Wang, L. C. Lima, C. Hansen, C. Hansen, and J. G. Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

M. Hardalov, A. Arora, P. Nakov, and I. Augenstein. 2021. Cross-domain label-adaptive stance detection.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation*.

M. Jing, G. Wei, P. Mitra, S. Kwon, and M. Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *International Joint Conference on Artificial Intelligence*.

Y. Kim. 2014. Convolutional neural networks for sentence classification. *Eprint Arxiv*.

D. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *Computer Science*.

Y. Li, B. Jiang, K. Shu, and H. Liu. 2020. Mm-covid: A multilingual and multidimensional data repository for combatingcovid-19 fake new.

A. Mosallanezhad, M. Karami, K. Shu, M. V. Mancenido, and H. Liu. 2022. Domain adaptive fake news detection via reinforcement learning.

L. Mou, R. Men, G. Li, Y. Xu, L. Zhang, R. Yan, and Z. Jin. 2015. Natural language inference by tree-based convolution and heuristic matching. *Computer Science*.

Q. Nan, J. Cao, Y. Zhu, Y. Wang, and J. Li. 2022a. Mdfend: Multi-domain fake news detection.

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

K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. 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.

A. Silva, L. Luo, S. Karunasekera, and C. Leckie. 2021. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data.

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 *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys '20, page 269–278, New York, NY, USA. Association for Computing Machinery.

Y. Wang, F. Ma, Z. Jin, Y. Ye, and K. Jha. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Acm Sigkdd International Conference*.

Hui Wu and Xiaodong Shi. 2022. Adversarial soft prompt tuning for cross-domain sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2438–2447, Dublin, Ireland. Association for Computational Linguistics.

Y. Zhu, Q. Sheng, J. Cao, S. Li, D. Wang, and F. Zhuang. 2022a. Generalizing to the future: Mitigating entity bias in fake news detection.

Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. 2022b. Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–14.

A Data Processing

Since the MultiFC datasets are basically political texts, we have divided them into finer-grained domains, such as policy, legislation, etc. Besides, in order to normalize the data set, all labels are unified into three types: fake, real or unknown.

We use ChatGPT and manual methods to divide the domain of the MultiFC dataset, respectively. For the former, use a script to call the ChatGPT interface, take text as input, and ask ChatGPT to predict its domain. The form is as follows: {The text is: [input], and what domain or topic does this

IAA	A & B	A & C	B & C	A & B & C
First Round	0.843	0.844	0.825	0.798
Second Round	0.859	0.877	0.850	0.811
Third Round	0.884	0.916	0.854	0.818
SUM	0.862	0.879	0.843	0.809

Table 8: The inter-annotator agreement (IAA) rate for three rounds of the new dataset domain labels, where A, B, C represent three annotators.

article belong to?}. For the latter, we used a manual method to divide the data into domains, and repeated inspections and corrections by three people. In addition to personal understanding, the main basis for manual division is domain keywords and co-occurrence words. For example, a text containing the keywords: law and court, may be subject to policy or legislation domain, but through the co-occurrence words: law-legislature, it can basically be identified as the domain of legislation.

It should be noted that there are two reasons why we choose ChatGPT to divide the domain here: 1) Due to the superiority of its parameters, its prediction results can be used as a comparison guarantee for manual division results. 2) Explore the accuracy of the large model on the current domain division task.

B Data Analysis

Method	Domain	Insts	labels
ChatGPT	policy	14654	3
	election	7941	3
	legislation	4333	3
	meeting	1497	3
	speech	2123	3
	examination	1987	3
	others	1875	3
Manual	policy	14489	3
	election	7895	3
	legislation	4080	3
	meeting	1673	3
	speech	2666	3
	examination	1364	3
	others	2243	3

Table 9: The statistics of the resulting new datasets after domain division by ChatGPT (top) and mannual (bottom).

In order to ensure the rationality and accuracy of the manually annotated data, three annotators performed three rounds of annotation on the original dataset (MultiFC). The inter-annotator agreement rate for the domain labels of the constructed dataset (New-MultiFC) is shown in Table 8.

The datasets after domain division by the above two methods are shown in Table 9, from which we

True-label	False-label	Acc
examination	policy	0.32
	election	0.08
	legislation	0.14
	meeting	0.06
	speech	0.29
	others	0.11
False-label	**True-label**	**Acc**
examination	policy	0.39
	election	0.12
	legislation	0.17
	meeting	0.03
	speech	0.24
	others	0.05

Table 10: The proportion of ChatGPT's prediction errors in the examination domain: samples that belong to examination domain are predicted to be other domains (top); samples that belong to other domains are predicted to be examination domain (bottom).

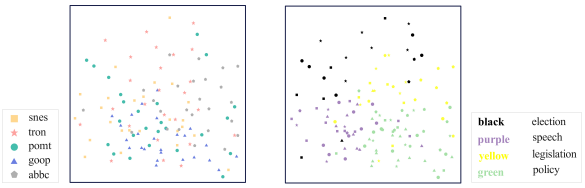
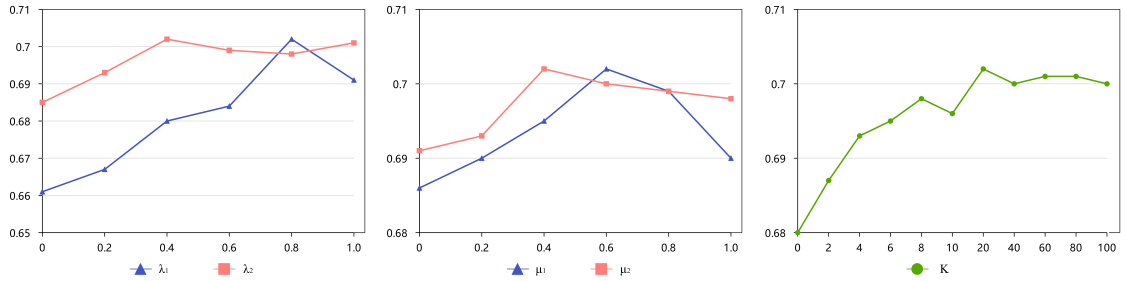


Figure 3: Visual representation of the sampling samples on the MultiFC and New-MultiFC datasets based on their respective domain divisions.

can observe that except for the large difference in the number of samples in the examination domain, the rest are very close. In order to explore why this difference exists, we selected the samples that belong to examination domain in the manual division, but were predicted by ChatGPT to other domains. On the other hand, we also selected samples that belong to other domains under manual classification, but were predicted to be examination domain by ChatGPT. Through manual discrimination, we found that for these samples, manual division is more reliable, so we only need to analyze the reasons for ChatGPT's prediction errors. Table 10 shows the proportion of ChatGPT's prediction errors in the examination domain, from which we observe that prediction errors of examination domain are concentrated on policy and speech domain. Based on word statistics, we speculate that this may be due to the high probability of overlapping words in examination, policy and speech domains, and they are all closely related to laws or important events. Therefore, we believe that ChatGPT still has a lot of room for exploration on some more complex tasks that require reasoning.

In addition, in order to verify the domain discrimination of the new dataset, we use T-SNE to visual-



(a) Effect of the λ_1 and λ_2 parameter (b) Effect of the μ_1 and μ_2 parameter (c) Effect of the K parameter

Figure 4: The impact of loss function parameters and the number of learnable vectors in continuous space, which serve as soft prompt.

ize some samples in MultiFC dataset, as shown in the Figure 3. After re-dividing the MultiFC dataset, the domain boundaries are clearer, which means that the new dataset is more domain-discriminative.

C Hyper-parameter Sensitivity

We test the sensitivity of multiple hyper-parameters on the FND-3 dataset, including 1) λ_1 and λ_2, which are weight coefficients in the loss function of Fake News Classifier; 2) μ_1 and μ_2, which are weight coefficients in the loss functions of Dual-Granularity Domain Classifiers; 3) K, is the number of learnable vectors which utilized as soft prompt. As shown in Figure 4, hyper-parameters have a certain impact on the model performance, and with properly tuned, FDDGA can achieve satisfying performance. From the experimental results, we have the following observations:

(1) According to Figure 4(a) and 4(b), we find that the model is more sensitive to text-level domain adversarial training than entity-level. The reason could be that compared with entities, the gap between text in different domains is much bigger, which requires more complex functions and parameters to model.

(2) According to the orange polyline in the first two sub-figures, we find that for domain adversarial training on entity-level, when the weight coefficients reach a certain threshold, the model performance tends to be stable, which is different from the text-level. To some extent, this also demonstrates that the gap of entities in different domains are smaller than that of text.

(3) From Figure 4(c), we observe that for soft prompt, not the longer the length, the better the performance. When the K value increased to 20, the model performance reaches the peak, and further increase will cause some degradation. The reason could be that excessive learnable tokens may lead to over-fitting, it means that the model trends to capture the local features, which are not generalized for unseen domains.