Out-of-Distribution Evidence-Aware Fake News Detection via Dual Adversarial Debiasing

Qiang Liu, *Member, IEEE*, Junfei Wu, Shu Wu, *Senior Member, IEEE*, and Liang Wang, *Fellow, IEEE*

***Abstract*—Evidence-aware fake news detection aims to conduct reasoning between news and evidences, which are retrieved based on news content, to find uniformity or inconsistency. However, we find evidence-aware detection models suffer from biases, i.e., spurious correlations between news/evidence contents and true/fake news labels, and are hard to be generalized to Out-Of-Distribution (OOD) situations. To deal with this, we propose a novel Dual Adversarial Learning (DAL) approach. We incorporate news-aspect and evidence-aspect debiasing discriminators, whose targets are both true/fake news labels, in DAL. Then, DAL reversely optimizes news-aspect and evidence-aspect debiasing discriminators to mitigate the impact of news and evidence content biases. At the same time, DAL also optimizes the main fake news predictor, so that the news-evidence interaction module can be learned. This process allows us to teach evidence-aware fake news detection models to better conduct news-evidence reasoning, and minimize the impact of content biases. To be noted, our proposed DAL approach is a plug-and-play module that works well with existing backbones. We conduct comprehensive experiments under two OOD settings, and plug DAL in four evidence-aware fake news detection backbones. Results demonstrate that, DAL significantly and stably outperforms the original backbones and some competitive debiasing methods.**

***Index Terms*—Fake news detection, evidence-aware, out-of-distribution, debiasing, adversarial learning.**

I. Introduction

With the development of online media, users can access to information more easily, and messages can be spread more rapidly. However, at the same time, fake statement about news can also be spread to the public more easily and widely. This leads to potential harm to the society, and may cause severe consequences. For example, rumors about Covid-19 have seriously affected the public health [1]. Thus, it is necessary to conduct research on automatic fake news detection [2], which remains a challenging task.

In the past decade, the task of fake news detection has been widely studied from different perspectives. Most research works focus on extracting and modeling patterns between true news and fake news based on different types of related features, including textual contents [3], [4], [5], multimodal contents [6], [7], [8] and propagation structures [9], [10], [11], [12], to detect fake news on online media. On the other hand, evidence-aware fake news detection [13], [14], [15], [16] aims to conduct textual reasoning between news, which is the claim of a fact, and evidences, which are retrieved from news platforms according to the content of the news. With the help of evidences, compared with other types of models, evidence-aware fake news detection models are more reliable and interpretable. In this work, we focus on the evidence-aware fake news detection task.

In real-world fake news detection systems, we usually face the Out-Of-Distribution (OOD) problem [17], [18], [19], [20], in which training phase and testing phase share different data distributions. We usually train fake news detection models on data from limited platforms, and need to apply them to generalized platforms. Meanwhile, during training of fake news detection models, we usually only have data from limited topics or events in constrained time periods, and require the models to generalize to more environments.

As datasets for training are usually with biased data distribution, evidence-aware fake news detection models may mistakenly learn spurious correlations between news contents and true/fake news labels. Considering the content relevance between news and evidences, spurious correlations between evidence contents and true/fake news labels [17] may also be captured. These spurious correlations bring evidence-aware fake news detection models severe OOD problems. We denote above biases as *news content bias* and *evidence content bias* respectively. Both biases limit evidence-aware fake news detection models to well reason between news and evidences. For example, if training data contains news about an accident event which is mostly fake news, the models have chance to learn that the topics and key-words about the accident refer to fake news, instead of conducting news-evidence reasoning. Obviously, above learned correlations do not hold when data distribution changes. Meanwhile, when training data and testing data share different topic distributions, e.g., entertainment and politics, correlations between some key-words in the training data and true/fake news labels may be captured, which may not exist during the testing phase. Moreover, in Section III-B, with empirical experiments, we prove that existing evidence-aware models have trouble in OOD environments. Accordingly, it is necessary to mitigate the

Manuscript received 27 April 2023; revised 12 October 2023; accepted 10 April 2024. Date of publication 24 April 2024; date of current version 27 September 2024. This work was supported by the National Natural Science Foundation of China under Grant 62206291, Grant 62141608, and Grant 62236010. Recommended for acceptance by G. Wang. *(Corresponding author: Shu Wu.)*

The authors are with the Center for Research on Intelligent Perception and Computing (CRIPAC), State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100045, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: qiang.liu@nlpr.ia.ac.cn; junfei.wu@cripac.ia.ac.cn; shu.wu@nlpr.ia.ac.cn; wangliang@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TKDE.2024.3390431

impact of news content bias and evidence content bias, and better teach detection models to conduct news-evidence reasoning.

However, the OOD problem has not been sufficiently studied in fake news detection, especially for the evidence-aware task [17]. Some approaches investigate multi-domain detection [21], [22] or cross-domain detection [18], [23], [24], but still require observations in target domains. Some multimodal fake news detection models [19], [20] focus on the situation that training samples and testing samples come from different events. However, they only focus on common features among different events or topics, but neglect the specific reasoning path in evidence-aware fake news detection. Recently, counterfactual inference [25] has been applied to debiasing evidence-aware fake news detection models [26]. However, it is performed during the testing phase, and hard to be adaptively optimized. Meanwhile, there is a similar task of evidence-aware fake news detection called fact verification [27], [28], [29]. Some efforts have made for debiasing fact verification models [30], [31]. However, biases in fact verification are different from those in evidence-aware fake news detection. The fact verification datasets are usually human-annotated, and negative samples are generated via data augmentation in claims by adding words such as "not." Thus, the biases in fact verification are the spurious correlations between some negative words in claims and labels. On the other hand, biases in evidence-aware fake news detection come from some topics and events in both news and evidence contents.

In this paper, we aim to mitigate both news and evidence content biases, and obtain detection models with great OOD generalization ability. Inspired by domain-adversarial training [32], [33], we propose a novel *Dual Adversarial Learning (DAL)* approach, for debiasing evidence-aware fake news detection models. The proposed DAL approach is a plug-and-play module, which can be applied in various evidence-aware fake news detection models. (1) For content of a piece of news, we conduct mean pooling on word embeddings of the news extracted in a detection model with no evidence content information, and obtain news representation. For content of each retrieved evidence, similarly, we conduct mean pooling on word embeddings of the evidence extracted in a detection model with no news content information, and obtain evidence representation. (2) In dual aspects, we remove the spurious correlations between news/evidence contents and true/fake news labels. Specifically, we use the news representation and evidence representations to construct news-aspect and evidence-aspect debiasing discriminators respectively, via simple multi-layer perception networks. (3) We conduct word-level and sentence-level interaction between news representation and evidence representations, as existing evidence-aware models do, to construct a main fake news predictor. (4) Simultaneously, we positively optimize the main fake news predictor, while reversely optimize the news-aspect and evidence-aspect debiasing discriminators. In this way, during the learning of the evidence-aware fake news detector, we can mitigate the spurious correlations between news/evidence contents and true/fake news labels as much as possible.

Furthermore, we conduct experiments under two OOD settings, i.e., cross-platform and cross-topic, and plug the proposed DAL approach in four evidence-aware fake news detection backbones. Our approach can significantly and stably outperform the original detection backbones and several state-of-the-art debiasing baselines, which shows the effectiveness of DAL for debiasing evidence-aware fake news detection models in OOD environments.

Our main contributions can be summarized as follows:

- We introduce news content bias and evidence content bias in evidence-aware fake news detection, and propose to mitigate them for training detection models with better OOD generalizing ability.
- We propose a plug-and-play dual adversarial learning approach, which incorporates both news-aspect debiasing and evidence-aspect debiasing modules.
- Comprehensive experiments are conducted to demonstrate the superiority of DAL in different OOD environments, and promote existing evidence-aware fake news detection backbones.

The rest of the paper is organized as follows. In Section II, we review some related work on fake news detection and debiasing methods. Then, in Section III, we present causal view analysis, and then conduct some empirical experiments to show the performances of existing evidence-aware models in OOD settings. Section IV details our proposed DAL approach for debiasing evidence-aware fake news detection models. In Section V, we conduct empirical experiments to verify the effectiveness of DAL. Finally, Section VI concludes our work.

II. Related Work

In this section, we first briefly review some related works on different types of fake news detection tasks, i.e., content-based, pattern-based and evidence-aware. Then, we introduce some debiasing methods which have been used in fake news detection tasks.

A. Content-Based Fake News Detection

Content-based fake news detection solely relies on content features for identifying misinformation. The considered content features can be grouped into two categories: textual features and multimodal features.

Fake news detection models based on textual features aim to find linguistic patterns of true/fake news [3], [34]. Existing research works on text-based fake news detection attempt to analyze style [5] or emotion [4], [35] of news. Meanwhile, BERT [36] has been widely used for detecting fake news recently [37]. And the NEP model [38] investigates popularity and novelty of news to detect misinformation from macro and micro environments respectively.

On the other hand, multimodal fake news detection considers both texts and images of news [6], [39] for identifying misinformation. The major problem to be investigated and solved is the interaction and fusion among multimodal features, and methods such as recurrent neural networks [6], multimodal variational autoencoder [7] and multimodal attention [8] have been applied. Moreover, some works focus discovering the inconsistency among different modalities for multimodal fake news detection [40], [41], [42].

B. Propagation-Based Fake News Detection

Propagation-based fake news detection aims to identify misinformation with feedback in online social media, such as reposts, likes, and comments [9], [10]. The core problem of propagation-based fake news detection is to aggregate the propagation history of a piece of news, and plenty of works have been published. Among them, models based on recurrent neural networks [11], [43], convolutional neural networks [44], [45], attentive networks [46], and generative adversarial networks [47] have been proposed.

With the development of Graph Neural Network (GNN) [48], recent works on propagation-based fake news detection mainly model propagation structures as graphs and applies GNN for misinformation identification [12], [49], [50]. Furthermore, some works based on contrastive learning [51] or hypergraph [52] have been further proposed. For better interpretability, reasoning over subgraph has been investigated [53], [54]. Meanwhile, for capturing temporal characteristics, dynamic graph has been constructed and applied for rumor detection [55].

C. Evidence-Aware Fake News Detection

Evidence-aware fake news detection seeks to explore the semantic similarity or conflict between news and related evidences to identify misinformation [13]. DeClare [14] employs BiLSTMs to embed the news and evidences. Then, it computes news representation via mean pooling, and computes news-aware attention for each word in evidences. HAN [56] computes the sentence-level coherence and entailment scores between news and evidences. EHIAN [57] incorporates self-attention for word-level interaction. MAC [15] proposes hierarchical mutli-head attention networks to model interactions between news and evidences. Then, some works propose to hierarchically conduct word-level and sentence-level interactions [58], [59]. Recently, GET [16] exploits graph-structure reasoning between news content and evidence contents, and GETRAL [60] further incorporates contrastive learning for representation leaning of news and evidences. Moreover, the way integrating pattern-based and evidence-aware fake news detection has also been investigated [61]. Meanwhile, there is research showing that, evidence-aware fake news detection models can hardly learn about news-evidence reasoning, but capture spurious correlations between news/evidence contents and true/fake news labels [17].

D. Debiasing Methods for Fake News Detection

Some research works have tried to debias different types of fake news detection models or models in some related tasks.

The causal theory has been applied for debiasing fake news detection models or fact checking models. Under the causal intervention framework [62], [63], [64], the sample weighting strategy [65], which downsamples the contribution of biased samples during loss computation, has been adopted for debiasing fake news detection models. For example, PeE [66] constructs a bias-only model, and downsamples the samples' spurious class distribution. For debiasing fact verification models, ReW [31] downsamples the samples containing n-grams highly correlated with labels. Meanwhile, counterfactual inference [25], which subtracts the effects of spurious correlations estimated by a bias-only model, is applied in eliminating entity bias [67], or debiasing evidence-aware fake news detection models [26] and fact checking models [68].

Besides, data augmentation strategies aim to generate some unbiased samples and incorporate them for training [69]. For the fact verification task, CrossAug [30] proposes a cross contrastive augmentation strategy, in which original claim text is modified to be negative, and evidences are changed to support the modified claim, to deal with the spurious correlations between some negative words in claim and labels. Recently, prompt learning has also been used for dealing with cross-language and cross-domain problems in propagation-based fake news detection [70]. Moreover, inspired by domain-adversarial training strategy [32], [33], EANN [19] incorporates an event discriminator to mitigate the correlations between input features and domain information. And similar strategy is used for fake news video detection [20]. Compared with our proposed DAL approach, EANN directly applies the domain-adversarial training strategy for dealing with the general domain shift problem. In contrast, DAL deeply investigates the specific biases in evidence-aware fake news detection models and designs suitable adversarial losses for mitigating them.

III. Analysis

In this section, we present a causal view analysis of evidence-aware fake news detection models, and then conduct some empirical experiments to show the performances of existing detection models in OOD settings.

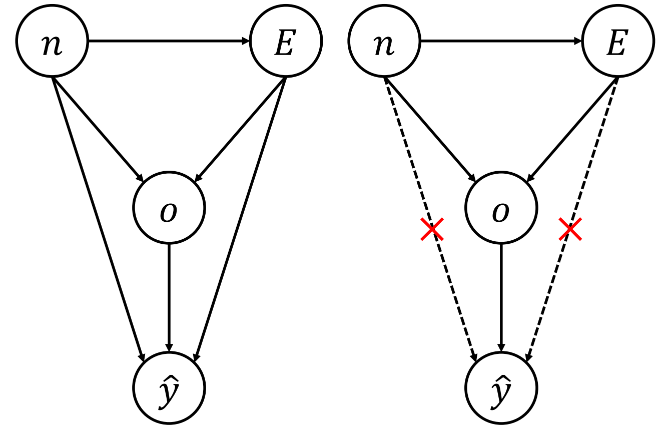
A. Causal Diagrams

We have news n, and the corresponding evidences $E = \{e_0, e_1, e_2, \ldots\}$. Via news-evidence interaction, we obtain interaction feature o, and the final prediction $\hat{y}$ on true/fake news label. As in some causal inference-based debiasing approaches, we rely on causal diagrams for investigating the specific biases in evidence-aware fake news detection models and correspondingly designing debiasing strategies. In Fig. 1, we illustrate the causal diagrams of evidence-aware fake news detection.

Evidence-aware fake news detection models usually conduct interaction between news and evidences and obtain interaction feature, i.e., $n \rightarrow o$ and $E \rightarrow o$. We usually wish the models can well perform news-evidence reasoning and give predictions, i.e., $o \rightarrow \hat{y}$. However, due to biases in data, the models may learn the spurious correlation between news content and true/fake news labels, i.e., $n \rightarrow \hat{y}$, without conducting new-evidence reasoning. Meanwhile, considering evidences are retrieved according to news contents, contents of news and evidences are highly correlated, and they may share similar topics or key-words, i.e., $n \rightarrow E$. Due to biased data distribution, these topics and key-words may be mistakenly associated with true/fake labels,

TABLE I
EMPIRICAL EXPERIMENTAL RESULTS, IN WHICH TRAINING AND TESTING ARE CONDUCTED ON THE SAME DATASET WITH RANDOM SPLITTING

Backbone	Input	Training: PolitiFact, Testing: PolitiFact		Training: Snopes, Testing: Snopes	
		F1-Macro	F1-Micro	F1-Macro	F1-Micro
BERT	News	0.6258	0.6367	0.6232	0.6867
	Envidences	0.6413	0.6433	0.6598	0.7235
	News+Envidences	0.6529	0.6624	0.6709	0.7481
DeClare	News	0.6251	0.6303	0.6159	0.6634
	Envidences	0.6392	0.6440	0.6522	0.7305
	News+Envidences	0.6508	0.6590	0.6642	0.7572
MAC	News	0.6244	0.6344	0.6090	0.6703
	Envidences	0.6455	0.6465	0.6637	0.7167
	News+Envidences	0.6609	0.6642	0.6725	0.7552
GET	News	0.6299	0.6329	0.6421	0.6999
	Envidences	0.6298	0.6344	0.6610	0.7325
	News+Envidences	0.6567	0.6702	0.6741	0.7545



(a) Causal diagram of existing detection models.

(b) Removing spurious correlations $n \rightarrow \hat{y}$ and $E \rightarrow \hat{y}$.

Fig. 1. Causal diagrams of evidence-aware fake news detection.

which results in the spurious correlation between evidence content and true/fake news labels, i.e., $E \rightarrow \hat{y}$. These spurious correlations may change cross different OOD environments, in which we usually have $p_{\text{train}}(\hat{y}|n) \neq p_{\text{test}}(\hat{y}|n)$ and $p_{\text{train}}(\hat{y}|E) \neq p_{\text{test}}(\hat{y}|E)$. Examples have been already discussed in Section I. Thus, it is necessary to mitigate the impact of spurious correlations $n \rightarrow \hat{y}$ and $E \rightarrow \hat{y}$, so that evidence-aware fake news detection models are able to better reason between news and evidences.

B. Empirical Experiments

In [17], the authors claim that evidence-aware fake detection models can hardly conduct news-evidence reasoning, but merely capture biases in contents. To clarify our motivation, we conduct further empirical experiments. We consider two datasets, i.e., PoliticFact and Snopes [17], and four detection models, i.e., BERT [36], DeClare [14], MAC [15] and GET [16]. In Table I, training and testing are conducted on the same platform, i.e., dataset. Meanwhile, in Table II, we conduct cross-platform training and testing. We have three types of input features to the models: only news, only evidences, and both news and evidences. Obviously, performances with both features are very close to those with only news features or evidence features. Moreover, comparing results in Table I and results in Table II, we can conclude that, the out-of-distribution problem severely affects the performances of evidence-aware fake news detection models. Accordingly, we need to teach models to better conduct news-evidence reasoning, and mitigate news and evidence content biases.

IV. METHODOLOGY

In this section, we detail our proposed DAL approach. We start with problem formulation of evidence-aware fake news detection. Then, we introduce the common structure of existing evidence-aware fake news detection models. Finally, we introduce our dual adversarial debiasing strategy.

A. Problem Formulation

Evidence-aware fake news detection is a binary classification task, in which the detection model aims to predict the probability of true/fake news. We have a set of news to be verified denoted as $\mathcal{N} = \{n_1, n_2, \ldots, n_{|\mathcal{N}|}\}$. According to each news n_i, evidences are retrieved, and denoted as $E_i = \{e_{i,1}, e_{i,2}, \ldots, e_{i,|E_i|}\}$. Respectively, n_i and $e_{i,j}$ contain textual contents of the corresponding news and evidence. The ground-truth true/fake news label of n_i is denoted as $y_i \in \{0, 1\}$. Based on n_i and E_i, we need to give prediction $\hat{y}_i$ on veracity of the news. To be noted, in this work, we focus on solving the OOD problem, in which training samples and testing samples are collected from different platforms, or share different topic distributions.

B. Structure of Evidence-Aware Fake News Detection

In this subsection, we summarize the common structure of evidence-aware fake news detection models [13], [14], [15], [16], [56], [57], [58], [59], which is shown in Fig. 2(a). With a piece of news n_i and its corresponding retrieved evidences $E_i = \{e_{i,1}, e_{i,2}, \ldots, e_{i,|E_i|}\}$, a detection model first encodes input features into word-level embeddings as

$$w_{n_i} = Encoder\left(n_i\right), \tag{1}$$

TABLE II
EMPIRICAL EXPERIMENTAL RESULTS, IN WHICH TRAINING AND TESTING ARE CONDUCTED ON DIFFERENT DATASETS

Backbone	Input	Training: Snopes, Testing: PolitiFact		Training: PolitiFact, Testing: Snopes	
		F1-Macro	F1-Micro	F1-Macro	F1-Micro
BERT	News	0.5112	0.5050	0.5353	0.6074
	Evidences	0.5003	0.5099	0.5437	0.6094
	News+Envidences	0.4932	0.5147	0.5569	0.6204
DeClare	News	0.5372	0.5372	0.5313	0.6250
	Evidences	0.5265	0.5265	0.5267	0.6214
	News+Envidences	0.5328	0.5350	0.5488	0.6442
MAC	News	0.5256	0.5287	0.5277	0.6115
	Evidences	0.5414	0.5427	0.5361	0.6036
	News+Envidences	0.5484	0.5497	0.5443	0.6320
GET	News	0.5359	0.5420	0.5307	0.6486
	Evidences	0.5265	0.5265	0.5321	0.6384
	News+Envidences	0.5473	0.5593	0.5420	0.6581

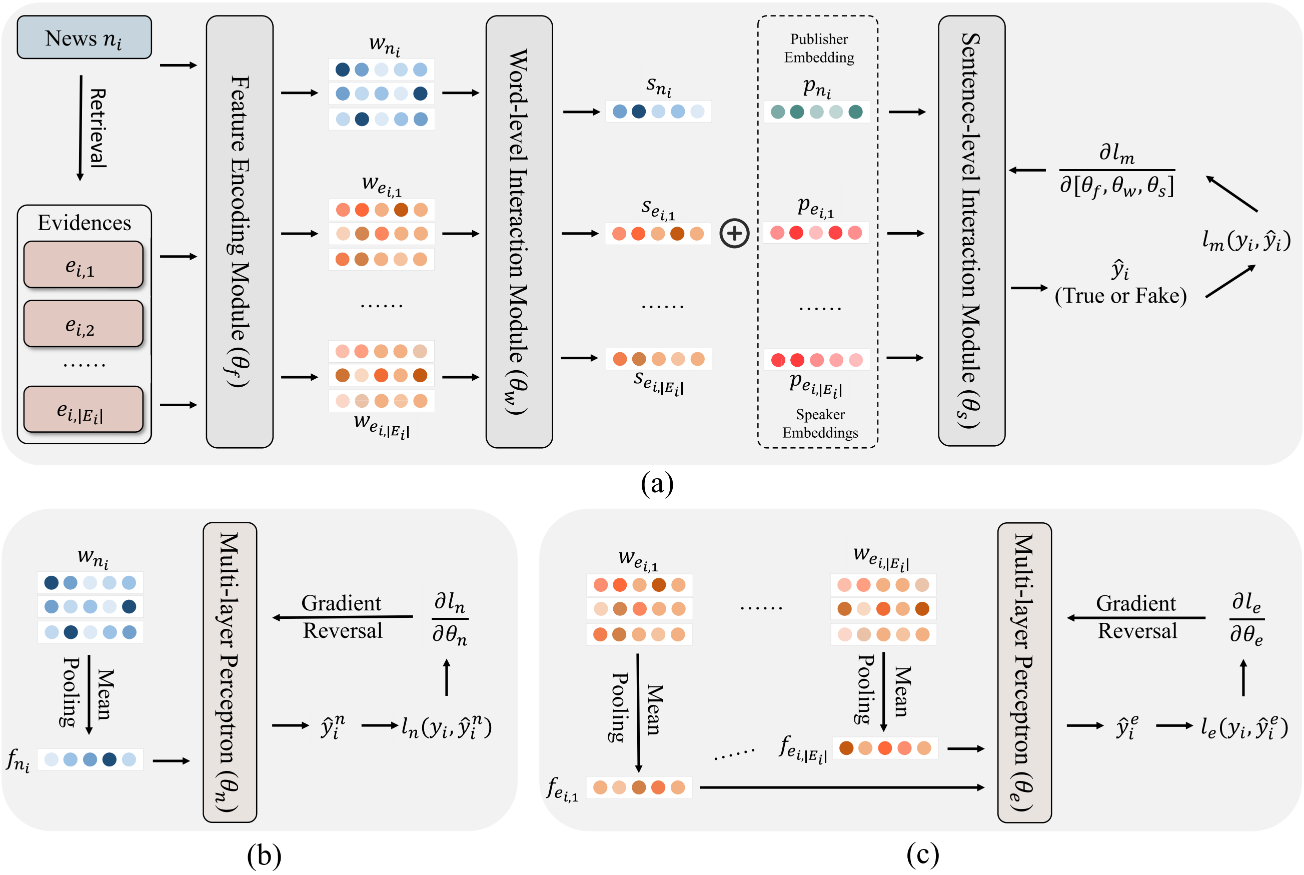


Fig. 2. The overview of DAL with a piece of news n_i and its corresponding evidences $E_i = \{e_{i,1}, e_{i,2}, \ldots, e_{i,|E_i|}\}$: (a) the structure of evidence-aware fake news detection models; (b) the structure of the news-aspect debiasing; (c) the structure of the evidence-aspect debiasing.

$$w_{e_{i,j}} = Encoder\left(e_{i,j}\right), \tag{2}$$

where $w_{n_i} \in \mathcal{R}^{|n_i|\times d_w}$ and $w_{e_{i,j}} \in \mathcal{R}^{|e_{i,j}|\times d_w}$ are word-level embedding matrices for news and evidence respectively, and d_w denotes the embedding dimensionality. We denote the learnable parameters in the features encoder as θ_f.

Then, the detection model conducts word-level news-evidence interaction, to generate sentence-level embeddings. Sentence-level embeddings of news are usually obtained via variety of pooling operations as

$$s_{n_i} = Pooling\left(w_{n_i}\right), \tag{3}$$

where $s_{n_i} \in \mathcal{R}^{d_s}$, and d_s denotes the embedding dimensionality. And the interaction process for generating evidence embeddings can be formulated as

$$s_{e_{i,j}} = WorInter\left(s_{n_i}, w_{e_{i,j}}\right), \tag{4}$$

where $s_{e_{i,j}} \in \mathcal{R}^{d_s}$. This process conducts word-level interaction and reasoning between news and evidences. We denote the learnable parameters in the word-level interaction module as θ_w.

Finally, detection models further conduct sentence-level interaction between news and evidences for the final prediction on the true/fake news label. This process can be formulated as

$$\hat{y}_i = SenInter\left(s_{n_i}, s_{e_{i,1}}, \ldots, s_{e_{i,|E_i|}}\right). \tag{5}$$

We denote the learnable parameters in the sentence-level interaction module as θ_s. Moreover, in some detection models, publisher embedding and speaker embedding is involved to represent the credibility of authors of the news and corresponding evidences. In this situation, (5) can be rewritten as

$$\hat{y}_i = SenInter\left(s_{n_i}, p_{n_i}, s_{e_{i,1}}, p_{e_{i,1}}, \ldots, s_{e_{i,|E_i|}}, p_{e_{i,|E_i|}}\right), \tag{6}$$

where p_{n_i} and $p_{e_{i,j}}$ denote publisher embedding and speaker embedding respectively.

We aim to propose a plug-and-play debiasing approach that can be applied for existing evidence-aware fake news detection models with feature encoding module, word-level interaction module and sentence-level interaction module as common components. To make the structure of our approach clearer, we conclude several representative evidence-aware fake news detection models as follows. HAN [56] uses GRU [71] for feature encoding, adopts coherence and entailment attention for sentence-level interaction, and has no word-level interaction and publisher and speaker embeddings. DeClare [14] uses BiLSTMs [72] for feature encoding, conducts claim-aware attention for word-level interaction, involves publisher and speaker embeddings, and has no sentence-level interaction. MAC [15] uses BiLSTMs for feature encoding, multi-head attention for word-level interaction, multi-head attention for sentence-level interaction, and involves publisher and speaker embeddings. GET [16] incorporates graph neural networks [48] for feature encoding, multi-head attention for word-level and sentence-level interaction, and involves publisher and speaker embeddings.

C. News-Aspect Debiasing Discriminator

According to the analysis in Section III, we need to remove the spurious correlations between news contents and labels. To achieve this, we need first to estimate the dependency of the labels on news contents. We obtain news representation via mean pooling as

$$f_{n_i} = MeanPooling\left(w_{n_i}\right), \tag{7}$$

where $f_{n_i} \in \mathcal{R}^{d_w}$. Then, we incorporate a news-aspect debiasing predictor to estimate the dependency, with Multi-Layer Perceptron (MLP), as

$$\hat{y}_i^n = MLP^n\left(f_{n_i}\right). \tag{8}$$

The learnable parameters in the discriminator are denoted as θ_n. With the discriminator, we can force the model not to accurately predict true/fake news labels based on only news contents, so that we can mitigate the news content bias.

Algorithm 1: Dual Adversarial Learning.

```
Input: News set $\mathcal{N}$ and an evidence-aware model $M$.
Output: Model parameters $\theta_f$, $\theta_w$ and $\theta_s$.
1: Initialize $k \leftarrow 0$ and $k_{best} \leftarrow 0$.
2: Initialize $[\theta_f^{(0)}, \theta_w^{(0)}, \theta_s^{(0)}]$ in $M$, and $[\theta_n^{(0)}, \theta_e^{(0)}]$.
3: repeat
4:     Keep $[\theta_f^{(k)}, \theta_w^{(k)}, \theta_s^{(k)}]$ fixed, and update
       $[\theta_n^{(k+1)}, \theta_e^{(k+1)}]$ according to (13)–(14) on $\mathcal{N}$.
5:     Keep $[\theta_n^{(k+1)}, \theta_e^{(k+1)}]$ fixed, and update
       $[\theta_f^{(k+1)}, \theta_w^{(k+1)}, \theta_s^{(k+1)}]$ according to (17) on $\mathcal{N}$.
6:     $k \leftarrow k + 1$.
7:     Update $k_{best} \leftarrow k$, if better validation results
       reached.
8: until Convergence.
9: return $\theta_f^{(k_{best})}$, $\theta_w^{(k_{best})}$ and $\theta_s^{(k_{best})}$.
```

D. Evidence-Aspect Debiasing Discriminator

Considering the content similarity between news and evidences, as well as the analysis in Section III, we also need to remove the spurious correlations between evidence contents and labels. To ensure that evidence representations used for debiasing contain no information about the news, we take use of the word embeddings $w_{e_{i,j}}$ before news-evidence interactions. Specifically, we also use mean pooling to generate the evidence representation as

$$f_{e_{i,j}} = MeanPooling\left(w_{e_{i,j}}\right), \tag{9}$$

where $f_{e_{i,j}} \in \mathcal{R}^{d_w}$. Then, we incorporate another MLP as our evidence-aspect debiasing discriminator, to estimate the dependency of the labels on each evidence content. The predictor can be formulated as the average of predictions by all the evidences as

$$\hat{y}_i^e = \frac{1}{|E_i|} \sum_{e_{i,j} \in E_i} MLP^e\left(f_{e_{i,j}}\right). \tag{10}$$

The learnable parameters in the discriminator are denoted as θ_e. Similarly, with the discriminator, we are able to mitigate the evidence content bias.

E. Dual Adversarial Learning

Inspired by domain-adversarial training [32], [33], which forces models to contain minimum domain information in an adversarial learning manner, we propose to positively optimize the main fake news predictor, while reversely optimize the news-aspect and evidence-aspect debiasing discriminators. Thus, during the learning of news-evidence reasoning, we can mitigate spurious correlations between news/evidence contents and true/fake news labels.

To avoid trivial solutions, i.e., the reverse optimization of news-aspect and evidence-aspect debiasing discriminators does not affect the parameters in the main fake news predictor (θ_f, θ_w and θ_s), we need to first positively optimize the parameters in the debiasing discriminators (θ_n and θ_e), while freeze the

TABLE III
PERFORMANCE COMPARISON RESULTS UNDER THE CROSS-PLATFORM SETTING

Backbone	Debiasing Approach	Training: Snopes, Testing: PolitiFact		Training: PolitiFact, Testing: Snopes	
		F1-Macro	F1-Micro	F1-Macro	F1-Micro
BERT	None	0.4932	0.5147	0.5569	0.6204
	ReW	0.5024	0.5090	0.5555	0.6262
	PoE	0.5127	0.5065	0.5569	0.6308
	CF	0.5108	0.5093	0.5592	0.6282
	EANN	0.5167	0.5146	0.5541	0.6311
	DAL	**0.5333**	**0.5733**	**0.5654**	**0.6432**
DeClare	None	0.5328	0.5350	0.5488	0.6442
	ReW	0.5366	0.5408	0.5513	0.6468
	PoE	0.5465	0.5535	0.5562	0.6493
	CF	0.5354	0.5367	0.5509	0.6434
	EANN	0.5395	0.5422	0.5556	0.6532
	DAL	**0.5811**	**0.5813**	**0.5785**	**0.6700**
MAC	None	0.5484	0.5497	0.5443	0.6320
	ReW	0.5513	0.5581	0.5478	0.6356
	PoE	0.5593	0.5648	0.5514	0.6376
	CF	0.5498	0.5548	0.5532	0.6386
	EANN	0.5564	0.5668	0.5507	0.6406
	DAL	**0.5808**	**0.5821**	**0.5787**	**0.6581**
GET	None	0.5473	0.5593	0.5420	0.6581
	ReW	0.5483	0.5564	0.5503	0.6553
	PoE	0.5556	0.5648	0.5588	0.6612
	CF	0.5477	0.5674	0.5608	0.6528
	EANN	0.5543	0.5712	0.5658	0.6616
	DAL	**0.5783**	**0.5836**	**0.5805**	**0.6650**

We plug several debiasing approaches in four evidence-aware fake news detection backbones. The best performances for each backbone are indicated by bold font.

other parameters. The news-aspect and evidence-aspect losses respectively are

$$l_n = \frac{1}{|\mathcal{N}|} \sum_{n_i \in \mathcal{N}} CrossEntropy\left(y_i, \hat{y}_i^n\right), \tag{11}$$

$$l_e = \frac{1}{|\mathcal{N}|} \sum_{n_i \in \mathcal{N}} CrossEntropy\left(y_i, \hat{y}_i^e\right), \tag{12}$$

and we optimize them as

$$\theta_n^{(k+1)} = \arg\min_{\theta_n} l_n \left[\theta_f^{(k)}, \theta_w^{(k)}, \theta_s^{(k)}, \theta_n^{(k)}\right], \tag{13}$$

$$\theta_e^{(k+1)} = \arg\min_{\theta_e} l_e \left[\theta_f^{(k)}, \theta_w^{(k)}, \theta_s^{(k)}, \theta_e^{(k)}\right], \tag{14}$$

where k denotes the optimization step.

Then, we freeze θ_n and θ_e, and optimize the main fake news detector. The main loss can be calculated as

$$l_m = \frac{1}{|\mathcal{N}|} \sum_{n_i \in \mathcal{N}} CrossEntropy\left(y_i, \hat{y}_i\right). \tag{15}$$

To mitigate the news and evidence content biases, we construct the overall loss as

$$l = l_m - \alpha l_n - \beta l_e, \tag{16}$$

where α and β are hyper-parameters controlling news-aspect and evidence-aspect debiasing respectively. In this way, gradients of the debiasing predictors are reversed. Then, parameters are optimized as

$$\begin{aligned} &\theta_f^{(k+1)}, \theta_w^{(k+1)}, \theta_s^{(k+1)} \\ &\quad = \arg\min_{\theta_f, \theta_w, \theta_s} l \left[\theta_f^{(k)}, \theta_w^{(k)}, \theta_s^{(k)}, \theta_n^{(k+1)}, \theta_e^{(k+1)}\right]. \end{aligned} \tag{17}$$

Above two optimization processes are performed alternatively, until convergence. The dual adversarial debiasing strategy teaches detection models to better perform reasoning and interaction between news and evidences, rather than relying solely on news or evidence contents. And the whole procedure of DAL is illustrated in Fig. 2 and Algorithm 1.

V. EXPERIMENTS

To evaluate the effectiveness of our proposed DAL approach, we conduct comprehensive experiments under two OOD settings with four state-of-the-art backbones to answer following Research Questions (RQs):

- RQ1: Under OOD environments, how well can DAL improve the backbones, and how does DAL perform compared to previous debiasing approaches?
- RQ2: How effective are the news-aspect debiasing and evidence-aspect debiasing in DAL?
- RQ3: How does DAL perform under different hyperparameter settings?

The following subsections describe the details of the experiments, results and analysis.

TABLE IV
PERFORMANCE COMPARISON RESULTS UNDER THE CROSS-TOPIC SETTING

Backbone	Debiasing Approach	Training: PolitiFact, Testing: PolitiFact		Training: Snopes, Testing: Snopes	
		F1-Macro	F1-Micro	F1-Macro	F1-Micro
BERT	None	0.6206	0.6206	0.5993	0.6916
	ReW	0.6223	0.6242	0.6258	0.7143
	PoE	0.6287	0.6318	0.6512	0.7456
	CF	0.6338	0.6357	0.6374	0.7043
	EANN	0.6268	0.6281	0.6477	0.7237
	DAL	**0.6541**	**0.6566**	**0.6774**	**0.7894**
DeClare	None	0.6005	0.6286	0.5941	0.6026
	ReW	0.6041	0.6296	0.5985	0.6095
	PoE	0.6125	0.6339	0.6246	0.6458
	CF	0.6153	0.6366	0.6167	0.6387
	EANN	0.6093	0.6342	0.6281	0.6511
	DAL	**0.6451**	**0.6508**	**0.6564**	**0.6977**
MAC	None	0.5732	0.6117	0.6584	0.6809
	ReW	0.5844	0.6174	0.6566	0.6767
	PoE	0.6032	0.6183	0.6582	0.6814
	CF	0.5856	0.6186	0.6541	0.6710
	EANN	0.5818	0.6203	0.6628	0.6926
	DAL	**0.6400**	**0.6405**	**0.6782**	**0.7215**
GET	None	0.6178	0.6358	0.6359	0.6601
	ReW	0.6242	0.6381	0.6254	0.6565
	PoE	0.6219	0.6328	0.6335	0.6659
	CF	0.6263	0.6394	0.6278	0.6739
	EANN	0.6228	0.6388	0.6383	0.6765
	DAL	**0.6350**	**0.6441**	**0.6458**	**0.7086**

We plug several debiasing approaches in four evidence-aware fake news detection backbones. The best performances for each backbone are indicated by bold font.

TABLE V
ABLATION STUDY UNDER THE CROSS-PLATFORM SETTING

Backbone	Debiasing Approach	Training: Snopes, Testing: PolitiFact		Training: PolitiFact, Testing: Snopes	
		F1-Macro	F1-Micro	F1-Macro	F1-Micro
BERT	DAL-news	0.5173	0.5671	0.5583	0.6245
	DAL-env	0.5165	0.5670	0.5520	0.6265
	DAL	**0.5333**	**0.5733**	**0.5654**	**0.6432**
DeClare	DAL-news	0.5715	0.5718	0.5718	0.6456
	DAL-env	0.5571	0.5615	0.5651	0.6505
	DAL	**0.5811**	**0.5813**	**0.5785**	**0.6700**
MAC	DAL-news	0.5694	0.5718	0.5627	0.6482
	DAL-env	0.5662	0.5689	0.5710	0.6482
	DAL	**0.5808**	**0.5821**	**0.5787**	**0.6581**
GET	DAL-news	0.5668	0.5678	0.5657	0.6620
	DAL-env	0.5650	0.5659	0.5703	0.6581
	DAL	**0.5783**	**0.5836**	**0.5805**	**0.6650**

DAL-news and DAL-env indicate dal with only news- and evidence-aspect debiasing respectively.
The best performances for each backbone are indicated by bold font.

A. Experimental Configurations

1) Datasets: We evaluate our DAL on PolitiFact and Snopes datasets which are collected by previous work [17]. The news and corresponding labels are from two major fact-checking websites *PolitiFact*[1] and *Snopes*.[2] And the evidence are top-10 relevant snippets retrieved by the news. As we only consider binary classification, we merge $false$, $mostly\ false$ claims into $false$ class and the others into $true$ for Snopes, and likewise merge $pants\ on\ fire$, $false$, $mostly\ false$ into $false$ claims and the rest to $true$ for PolitiFact.

2) Setups: Following the work [17], [26], We choose F1-Macro and F1-Micro as our evaluation metrics. To evaluate the effectiveness of our method, we construct OOD settings, including *cross-platform* and *cross-topic*. Under the cross-platform setting, which is in line with the work [17], the model is trained and validated on one dataset's training and validation sets respectively, while tested on out-of-dataset testing set (e.g., trained and validated on Snopes while tested on PolitiFact). Under the cross-topic setting, we first apply the LDA algorithm [73] to cluster samples by topics, and split the whole dataset into training, validation, and testing sets to ensure that there is no overlap topics between them. The training, validation and testing samples are from the same dataset, but share different topics. It is worth noting that we only tune the hyperparameters on validation sets to ensure that the model cannot access the data distribution of testing sets. And the model early stops when F1-macro does not increase in 10 epochs.

[1] https://www.politifact.com/
[2] https://www.snopes.com/

TABLE VI
Ablation Study Under the Cross-Topic Setting

Backbone	Debiasing Approach	Training: Snopes, Testing: PolitiFact		Training: PolitiFact, Testing: Snopes	
		F1-Macro	F1-Micro	F1-Macro	F1-Micro
BERT	DAL-news	0.6394	0.6394	0.6423	0.7825
	DAL-env	0.6256	0.6276	0.6352	0.7537
	DAL	**0.6541**	**0.6566**	**0.6774**	**0.7894**
DeClare	DAL-news	0.6299	0.6397	0.6470	0.6839
	DAL-env	0.6130	0.6379	0.6415	0.6822
	DAL	**0.6451**	**0.6508**	**0.6564**	**0.6977**
MAC	DAL-news	0.6286	0.6353	0.6614	0.7086
	DAL-env	0.6293	0.6331	0.6622	0.7096
	DAL	**0.6400**	**0.6405**	**0.6782**	**0.7215**
GET	DAL-news	0.6303	0.6410	0.6364	0.6918
	DAL-env	0.6329	0.6368	0.6260	0.7017
	DAL	**0.6350**	**0.6441**	**0.6458**	**0.7086**

DAL-news and DAL-env indicate dal with only news- and evidence-aspect debiasing respectively.
The best performances for each backbone are indicated by bold font.

3) Detection Backbones: We also choose four state-of-the-art evidence-aware fake news detection model as our backbones:

- *BERT* [36] takes the concatenated news and evidence as input into BERT to model contextual evidence representations.
- *DeClare* [14] uses a news specific attention to compute the weights of evidence words and averages the credibility score of each evidence as the final prediction.
- *MAC* [15] adopts word-level and document-level attention mechanism to model different granularity interactions between news and evidence.
- *GET* [16] incorporates graph neural network to model textual content, in order to mine fine-grained semantics of news and evidence.

4) Debiasing Baselines: To demonstrate the effectiveness of DAL, we compare it with several debiasing methods for fake news detection:

- *ReW* [31] assigns a low weight for biased samples containing n-grams highly correlated to labels.
- *PoE* [66] down-weights the biased samples based on the uneven predictions of the bias-only model.
- *CF* [26] adopts counterfactual inference [25] to subtract the outputs of a bias-only model to obtain debiased predictions.
- *EANN* [19], [20] develops an event adversarial framework to help the model capture event-invariant features in fake news detection.

5) Implementation Details: We set the maximum lengths of news and evidences both as 100, and the optimizer is Adam [74]. Each news is attached with 10 pieces of retrieved relevant evidences from websites other than the corresponding news publisher, which are collected in the original datasets [17]. We report average results with 5 different random seeds. The hyper-parameters α and β are both taken from $\{0.001, 0.01, 0.1, 1.0\}$. Other settings of detection backbones and debiasing baselines follow their original literature.

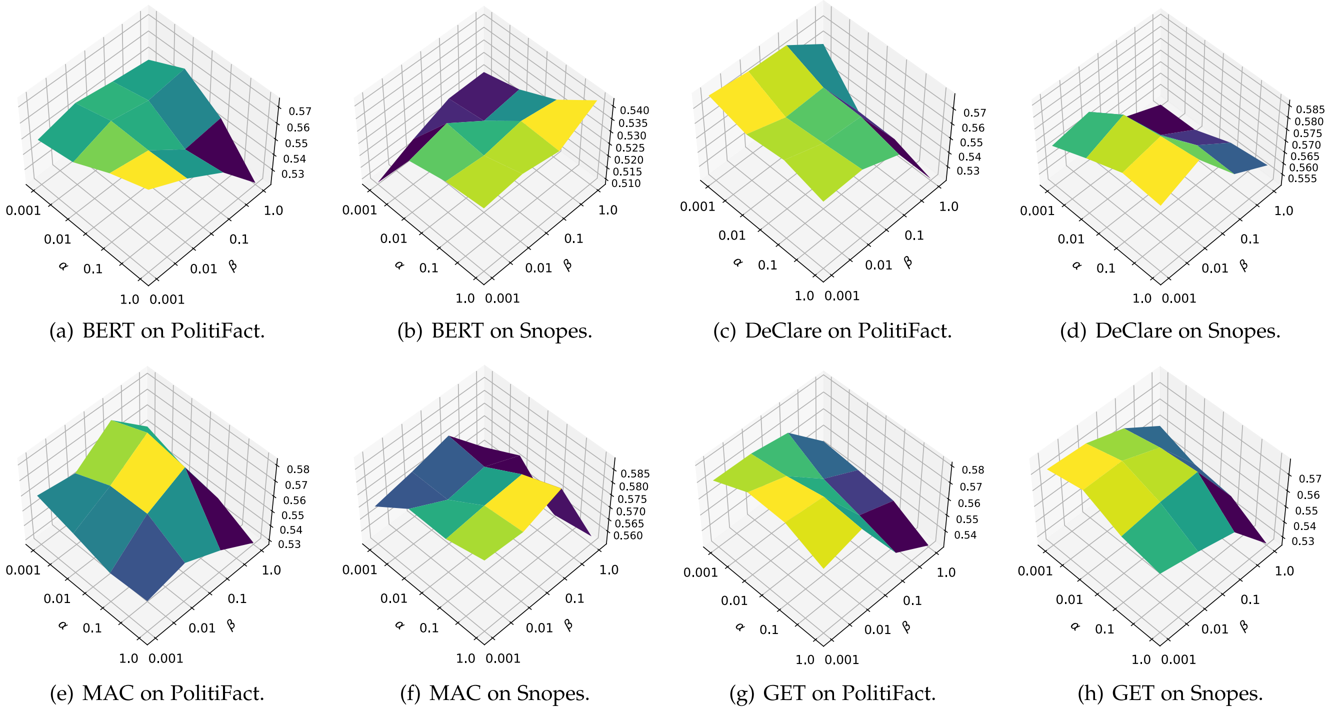
B. Overall Performance (RQ1)

We compare DAL with four debiasing methods on four backbones, under cross-platform setting and cross-topic setting respectively. There is a significant data distribution gap between these two OOD settings, which poses a severe challenge to testing OOD generalization capabilities. From Tables III and IV, we can have the following observations.

First, under the cross-platform setting, the comparison results are summarized in Table III. We can observe that DAL brings significant improvement for each backbone model on both metrics, compared to the marginal improvement brought by other debiasing competitors. Notably, our method achieves the top performance consistently, while the performance gains of the other debiasing methods are unstable on different testing sets. It indicates that the news and evidence content biases in the original datasets can be mitigated to the maximum extent by DAL effectively and the backbones are facilitated better OOD generalization. To be more specific, all baselines over DeClare and MAC model can only outperform the backbone without any debiasing method slightly by around 1% measured by F1-Macro and F1-Micro. In contrast, there is about 3%–4% improvement of DeClare and MAC equipped with DAL measured by each evaluation metric. Similar phenomena can also be observed on the rest two backbones.

Second, Table IV summarizes debiasing performance under the cross-topic setting. The performance of each backbone without any debiasing methods under the cross-topic setting is between that under the in-distribution setting in Table I and the cross-platform setting in Table III, indicating that the distribution shift problem exists while is weaker than that in the cross-platform setting. It is obvious that DAL can still maintain consistent and significant improvement over backbones, compared to its debiasing counterparts. On the other hand, the performance improvement of baselines varies across different datasets and backbones. It is worth noting that most baselines other than EANN over the MAC and GET model even underperform the original backbones on the Snopes dataset.

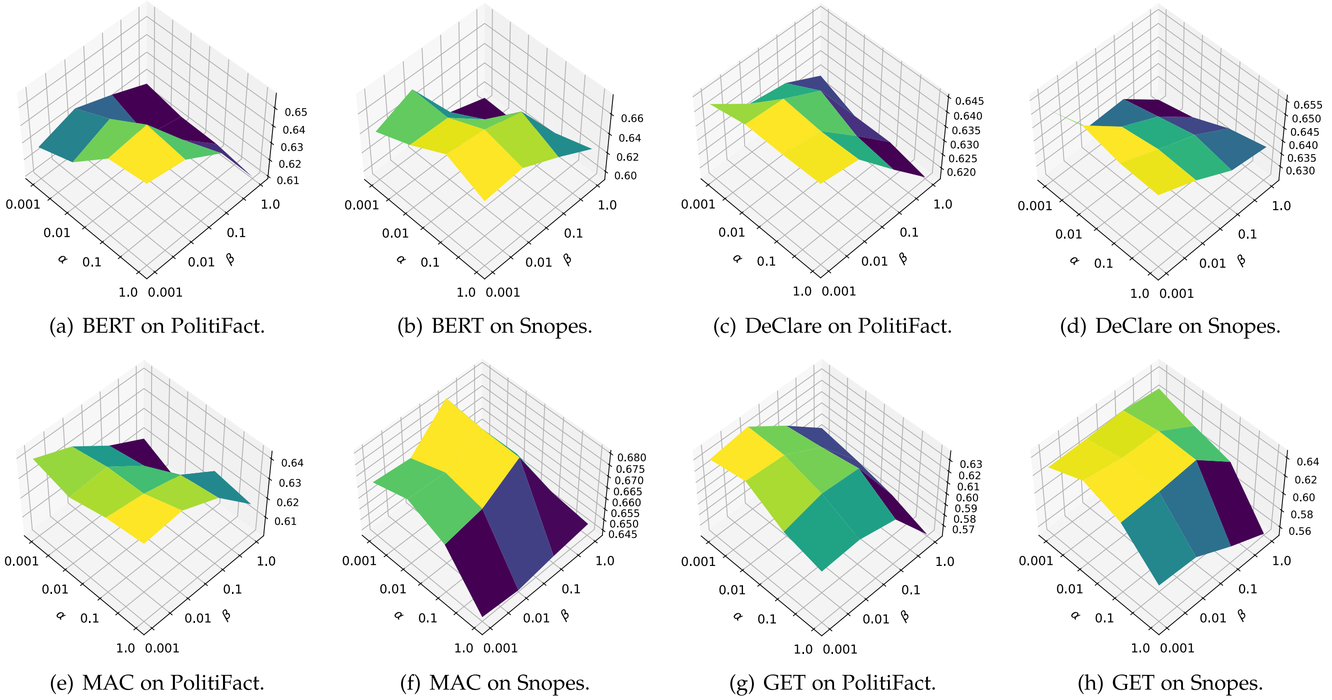
Based on the above two groups of experimental results, we can draw the following conclusions. (1) Results under the cross-platform setting are much lower than those under the cross-topic setting. This may indicate that the cross-platform setting is with more serious distribution shift issues. Besides the difference in topics, different platforms have other distinguishable



(a) BERT on PolitiFact. (b) BERT on Snopes. (c) DeClare on PolitiFact. (d) DeClare on Snopes.

(e) MAC on PolitiFact. (f) MAC on Snopes. (g) GET on PolitiFact. (h) GET on Snopes.

Fig. 3. Sensitivity of hyper-parameters, i.e., α and β, of DAL plugged in four different evidence-aware fake news detection backbones, tested on PoliticFact and Snopes under the cross-platform setting measured by F1-Macro.



(a) BERT on PolitiFact. (b) BERT on Snopes. (c) DeClare on PolitiFact. (d) DeClare on Snopes.

(e) MAC on PolitiFact. (f) MAC on Snopes. (g) GET on PolitiFact. (h) GET on Snopes.

Fig. 4. Sensitivity of hyper-parameters, i.e., α and β, of DAL plugged in four different evidence-aware fake news detection backbones, tested on PoliticFact and Snopes under the cross-topic setting measured by F1-Macro.

factors such as writing styles and audience groups. (2) Sample weighting approaches bring slight and unstable performance improvements. This shows that the estimation of sample weights is with high variance and hard to be stable, as mentioned in causal inference-related literature. Among them, PoE performs better than ReW. This may indicate that weight estimation based on a sole model is more effective than that based on n-gram statistics. (3) The CF approach also brings slight performance improvements and even underperforms the original backbones in some cases. This may be due to the excessive bias removal on the training sets leading to harmful effects. (4) EANN can stably bring improvements over the original backbones. Because it leverages adversarial learning to adaptively mitigate spurious correlations between domain features and labels. (5) DAL significantly outperforms EANN, though they both incorporate adversarial learning for debiasing. Because EANN, which can be seen as an application of domain-adversarial training, only focuses on common features among different events or topics. In contrast, DAL deeply investigates the specific biases in evidence-aware fake news detection models and accordingly designs suitable adversarial losses. In other words, DAL conducts debiasing in the specific reasoning path in evidence-aware fake news detection, which is the main characteristic of this problem. (6) DAL achieves the best performances under all the settings. This shows the superiority of mitigating news and evidence content biases with dual adversarial learning.

C. Ablation Study (RQ2)

To further demonstrate the effects of news-aspect and evidence-aspect debiasing of DAL, we conduct detailed ablation studies with the four backbones under the two OOD settings and the results are shown in Tables V and VI, respectively. We denote the DAL variants as follows: (1) DAL-news: A variant only keeps the news-aspect debiasing. (2) DAL-env: A variant only keeps the evidence-aspect debiasing.

According to the results, we can first observe that backbones with DAL-news and DAL-env both perform better than the backbones significantly, which indicates the existence of bias from news and evidences and the adversarial debiasing of each bias aspect is beneficial and effective. What's more, the effectiveness of bias removal from different aspects is similar for most backbones, while DeClare is the exception where DAL-news outperforms DAL-env under both settings. It can be attributed to the different structures of backbones, which determine different bias poisoning from news-aspect and evidence-aspect.

In addition, it is obvious that the proposed DAL approach outperforms DAL-news and DAL-env greatly for all backbones and test settings. It is reasonable since each aspect of debiasing is effective and demonstrates the importance of mitigating both the impact of news and evidence content biases. Therefore, dual adversarial debiasing can integrate the effectiveness of each single aspect and further go beyond, making the superiority of DAL.

D. Sensitivity Analysis (RQ3)

In this section, we test the sensitivity of DAL to the hyper-parameters α and β for four different backbones under the cross-platform and cross-topic settings, with values taken from $\{0.001, 0.01, 0.1, 1.0\}$. These two hyper-parameters determine the extent of the debiasing of news content bias and evidence content bias, respectively. The testing results are summarized in Figs. 3 and 4.

As shown in Fig. 3, DAL shows different sensitivity for different backbones on different datasets. Taking BERT backbone as an example, the best performance is achieved when $\alpha = 1.0$, $\beta = 0.001$ on PolitiFact and $\alpha = 1.0$, $\beta = 1.0$ on Snopes, which indicates that news-aspect debiasing plays an important role for BERT model. Meanwhile, when $\alpha = 0.001$, $\beta = 0.001$ on PolitiFact and $\alpha = 1.0$, $\beta = 0.001$ on Snops, DeClare equipped with DAL has the highest results, which indicates that a slight extent of evidence-aspect debiasing is beneficial for DeClare. It is mainly due to the bias poison from news and evidence that has distinctions for different backbones, then the optimal hyper-parameters for different backbones are different. We can also observe similar phenomena under the cross-topic setting, as shown in Fig. 4. To be noted, in the rest of our experiments, hyper-parameters are tuned and results are reported according to the best performances on validation sets under different OOD settings.

VI. Conclusion

In this paper, we propose a plug-and-play Dual Adversarial Learning (DAL) approach for debiasing evidence-aware fake news detection models. DAL incorporates a news-aspect debiasing predictor, and an evidence-aspect debiasing predictor for all the corresponding evidences. Via reversely optimizing the news-aspect and evidence-aspect discriminators, while positively optimizing the main fake news predictor, DAL can mitigate the spurious correlations between news/evidence contents and true/fake news labels. Experiments under two different OOD settings and with four evidence-aware backbones, strongly demonstrate the effectiveness and stability of the DAL approach.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions allowing them to improve the quality of this article.

References

[1] S. B. Naeem and R. Bhatti, "The COVID-19 'infodemic': A new front for information professionals," *Health Inf. Libraries J.*, vol. 37, pp. 233–239, 2020.
[2] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proc. Int. Conf. World Wide Web*, 2011, pp. 675–684.
[3] S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas, "Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 647–653.
[4] A. Giachanou, P. Rosso, and F. Crestani, "Leveraging emotional signals for credibility detection," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2019, pp. 877–880.
[5] P. Przybyla, "Capturing the style of fake news," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 490–497.
[6] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proc. ACM Int. Conf. Multimedia*, 2017, pp. 795–816.
[7] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. World Wide Web Conf.*, 2019, pp. 2915–2921.
[8] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 153–162.
[9] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *Proc. IEEE Int. Conf. Data Mining*, 2013, pp. 1103–1108.
[10] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong, "Detect rumors using time series of social context information on microblogging websites," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 1751–1754.
[11] J. Ma et al., "Detecting rumors from microblogs with recurrent neural networks," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 3818–3824.

[12] T. Bian et al., “Rumor detection on social media with bi-directional graph convolutional networks,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 549–556.

[13] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum, “Where the truth lies: Explaining the credibility of emerging claims on the web and social media,” in *Proc. Int. Conf. World Wide Web Companion*, 2017, pp. 1003–1012.

[14] K. Popat, S. Mukherjee, A. Yates, and G. Weikum, “DeClarE: Debunking fake news and false claims using evidence-aware deep learning,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 22–32.

[15] N. Vo and K. Lee, “Hierarchical multi-head attentive network for evidence-aware fake news detection,” in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2021, pp. 965–975.

[16] W. Xu, J. Wu, Q. Liu, S. Wu, and L. Wang, “Evidence-aware fake news detection with graph neural networks,” in *Proc. Web Conf.*, 2022, pp. 2501–2510.

[17] C. Hansen, C. Hansen, and L. C. Lima, “Automatic fake news detection: Are models learning to reason?,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 80–86.

[18] H. Lin, J. Ma, L. Chen, Z. Yang, M. Cheng, and G. Chen, “Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2022, pp. 2543–2556.

[19] Y. Wang et al., “EANN: Event adversarial neural networks for multi-modal fake news detection,” in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 849–857.

[20] H. Choi and Y. Ko, “Using topic modeling and adversarial neural networks for fake news video detection,” in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2021, pp. 2950–2954.

[21] Q. Nan, J. Cao, Y. Zhu, Y. Wang, and J. Li, “MDFEND: Multi-domain fake news detection,” in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2021, pp. 3343–3347.

[22] Y. Zhu et al., “Memory-guided multi-view multi-domain fake news detection,” *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 7, pp. 7178–7191, Jul. 2023.

[23] A. Silva, L. Luo, S. Karunasekera, and C. Leckie, “Embracing domain differences in fake news: Cross-domain fake news detection using multimodal data,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 557–565.

[24] Q. Nan et al., “Improving fake news detection of influential domain via domain-and instance-level transfer,” in *Proc. Int. Conf. Comput. Linguistics*, 2022, pp. 2834–2848.

[25] Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, and J.-R. Wen, “Counterfactual VQA: A cause-effect look at language bias,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12 700–12 710.

[26] J. Wu, Q. Liu, W. Xu, and S. Wu, “Bias mitigation for evidence-aware fake news detection by causal intervention,” in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 2308–2313.

[27] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, “FEVER: A large-scale dataset for fact extraction and verification,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2018, pp. 809–819.

[28] J. Zhou et al., “GEAR: Graph-based evidence aggregating and reasoning for fact verification,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 892–901.

[29] Z. Liu, C. Xiong, M. Sun, and Z. Liu, “Fine-grained fact verification with kernel graph attention network,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7342–7351.

[30] M. Lee, S. Won, J. Kim, H. Lee, C. Park, and K. Jung, “CrossAug: A contrastive data augmentation method for debiasing fact verification models,” in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2021, pp. 3181–3185.

[31] T. Schuster, D. Shah, Y. J. S. Yeo, D. R. F. Ortiz, E. Santus, and R. Barzilay, “Towards debiasing fact verification models,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 3419–3425.

[32] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[33] Y. Ganin et al., “Domain-adversarial training of neural networks,” *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.

[34] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, “Truth of varying shades: Analyzing language in fake news and political fact-checking,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2931–2937.

[35] X. Zhang, J. Cao, X. Li, Q. Sheng, L. Zhong, and K. Shu, “Mining dual emotion for fake news detection,” in *Proc. Web Conf.*, 2021, pp. 3465–3476.

[36] J. D. M.-W. C. Kenton and L. K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.

[37] R. K. Kaliyar, A. Goswami, and P. Narang, “FakeBERT: Fake news detection in social media with a bert-based deep learning approach,” *Multimedia Tools Appl.*, vol. 80, no. 8, pp. 11 765–11 788, 2021.

[38] Q. Sheng, J. Cao, X. Zhang, R. Li, D. Wang, and Y. Zhu, “Zoom out and observe: News environment perception for fake news detection,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 4543–4556.

[39] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, “Novel visual and statistical image features for microblogs news verification,” *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 598–608, Mar. 2017.

[40] R. Tan, B. Plummer, and K. Saenko, “Detecting cross-modal inconsistency to defend against neural fake news,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 2081–2106.

[41] P. Qi et al., “Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues,” in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 1212–1220.

[42] Y. Chen et al., “Cross-modal ambiguity learning for multimodal fake news detection,” in *Proc. Web Conf.*, 2022, pp. 2897–2905.

[43] J. Ma, W. Gao, and K.-F. Wong, “Rumor detection on Twitter with tree-structured recursive neural networks,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1980–1989.

[44] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan, “A convolutional approach for misinformation identification,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 3901–3907.

[45] Y. Liu and Y.-F. Wu, “Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks,” in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 354–361.

[46] Q. Liu, F. Yu, S. Wu, and L. Wang, “Mining significant microblogs for misinformation identification: An attention-based approach,” *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 5, pp. 1–20, 2018.

[47] J. Ma, W. Gao, and K.-F. Wong, “Detect rumors on Twitter by promoting information campaigns with generative adversarial learning,” in *Proc. World Wide Web Conf.*, 2019, pp. 3049–3055.

[48] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–4.

[49] Y.-J. Lu and C.-T. Li, “GCAN: Graph-aware co-attention networks for explainable fake news detection on social media,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 505–514.

[50] V.-H. Nguyen, K. Sugiyama, P. Nakov, and M.-Y. Kan, “FANG: Leveraging social context for fake news detection using graph representation,” in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2020, pp. 1165–1174.

[51] T. Sun, Z. Qian, S. Dong, P. Li, and Q. Zhu, “Rumor detection on social media with graph adversarial contrastive learning,” in *Proc. Web Conf.*, 2022, pp. 2789–2797.

[52] X. Sun et al., “Structure learning via meta-hyperedge for dynamic rumor detection,” *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 9, pp. 9128–9139, Sep. 2023.

[53] Y. Jin et al., “Towards fine-grained reasoning for fake news detection,” in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 5746–5754.

[54] R. Yang, X. Wang, Y. Jin, C. Li, J. Lian, and X. Xie, “Reinforcement subgraph reasoning for fake news detection,” in *Proc. ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2022, pp. 2253–2262.

[55] M. Sun, X. Zhang, J. Zheng, and G. Ma, “DDGCN: Dual dynamic graph convolutional networks for rumor detection on social media,” in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 4611–4619.

[56] J. Ma, W. Gao, S. Joty, and K.-F. Wong, “Sentence-level evidence embedding for claim verification with hierarchical attention networks,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2561–2571.

[57] L. Wu, Y. Rao, X. Yang, W. Wang, and A. Nazir, “Evidence-aware hierarchical interactive attention networks for explainable claim verification,” in *Proc. Int. Conf. Int. Joint Conf. Artif. Intell.*, 2021, pp. 1388–1394.

[58] L. Wu, Y. Rao, L. Sun, and W. He, “Evidence inference networks for interpretable claim verification,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 14 058–14 066.

[59] L. Wu, Y. Rao, Y. Lan, L. Sun, and Z. Qi, “Unified dual-view cognitive model for interpretable claim verification,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 59–68.

[60] J. Wu, W. Xu, Q. Liu, S. Wu, and L. Wang, “Adversarial contrastive learning for evidence-aware fake news detection with graph neural networks,” 2022, *arXiv:2210.05498*.

[61] Q. Sheng, X. Zhang, J. Cao, and L. Zhong, "Integrating pattern-and fact-based fake news detection via model preference learning," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2021, pp. 1640–1650.

[62] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims, "Recommendations as treatments: Debiasing learning and evaluation," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 1670–1679.

[63] Y. Zhang et al., "Causal intervention for leveraging popularity bias in recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 11–20.

[64] L. Hu, Z. Chen, Z. Z. J. Yin, and L. Nie, "Causal inference for leveraging image-text matching bias in multi-modal fake news detection," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 11, pp. 11141–11152, Nov. 2023.

[65] C. Clark, M. Yatskar, and L. Zettlemoyer, "Don't take the easy way out: Ensemble based methods for avoiding known dataset biases," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 4069–4082.

[66] R. K. Mahabadi, Y. Belinkov, and J. Henderson, "End-to-end bias mitigation by modelling biases in corpora," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 8706–8716.

[67] Y. Zhu, Q. Sheng, J. Cao, S. Li, D. Wang, and F. Zhuang, "Generalizing to the future: Mitigating entity bias in fake news detection," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 2120–2125.

[68] W. Xu, Q. Liu, S. Wu, and L. Wang, "Counterfactual debiasing for fact verification," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2023, pp. 6777–6789.

[69] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 6382–6388.

[70] H. Lin et al., "Zero-shot rumor detection with propagation structure via prompt learning," in *Proc. AAAI Conf. Artif. Intell.*, 2023, Art. no. 582.

[71] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in *Proc. IEEE Int. Midwest Symp. Circuits Syst.*, 2017, pp. 1597–1600.

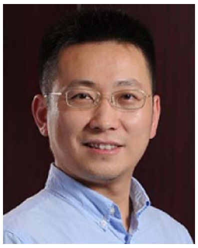
[72] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*.

[73] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. Jan., pp. 993–1022, 2003.

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

Qiang Liu (Member, IEEE) received the PhD degree from the Chinese Academy of Sciences (CASIA). He is an associate professor with the Center for Research on Intelligent Perception and Computing (CRIPAC), State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences. Currently, his research interests include data mining, misinformation detection, LLM safety, and AI for science. He has published papers in top-tier journals and conferences, such as *IEEE Transactions on Knowledge and Data Engineering*, AAAI, NeurIPS, KDD, WWW, SIGIR, CIKM, ICDM, ACL, and EMNLP.

Junfei Wu is currently working toward the PhD degree in computer science with the Center for Research on Intelligent Perception and Computing (CRIPAC), State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences (CASIA). His current research interests mainly include fake news detection.



Shu Wu (Senior Member, IEEE) received the BS degree from Hunan University, China, in 2004, the MS degree from Xiamen University, China, in 2007, and the PhD degree from the Department of Computer Science, University of Sherbrooke, Quebec, Canada, all in computer science. He is an associate professor with the Center for Research on Intelligent Perception and Computing (CRIPAC), State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences (CASIA). He has published more than 50 papers in the areas of data mining and information retrieval in international journals and conferences, such as *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Human-Machine Systems*, AAAI, ICDM, SIGIR, and CIKM. His research interests include data mining, information retrieval, and recommendation.

Liang Wang (Fellow, IEEE) received the BEng and MEng degrees from Anhui University, in 1997 and 2000, respectively, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2004. From 2004 to 2010, he was a research assistant with Imperial College London, U.K., and Monash University, Australia, a research fellow with the University of Melbourne, Australia, and a lecturer with the University of Bath, U.K., respectively. Currently, he is a full professor with the Hundred Talents Program, State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA. His major research interests include machine learning, pattern recognition, and computer vision. He has widely published in highly ranked international journals such as *IEEE Transactions on Pattern Analysis and Machine Intelligence* and *IEEE Transactions on Image Processing*, and leading international conferences such as CVPR, ICCV, and ECCV. He has served as an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Image Processing*, and *Pattern Recognition*. He is an IAPR fellow.