Counterfactual Debiasing for Fact Verification

Weizhi Xu[1,2*] **Qiang Liu**[1,2*] **Shu Wu**[1,2†] **Liang Wang**[1,2]
[1]Center for Research on Intelligent Perception and Computing,
State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
weizhi.xu@cripac.ia.ac.cn, {qiang.liu,shu.wu,wangliang}@nlpr.ia.ac.cn

Abstract

Fact verification aims to automatically judge the veracity of a claim according to several pieces of evidence. Due to the manual construction of datasets, spurious correlations between claim patterns and its veracity (i.e., biases) inevitably exist. Recent studies show that models usually learn such biases instead of understanding the semantic relationship between the claim and evidence. Existing debiasing works can be roughly divided into data-augmentation-based and weight-regularization-based pipeline, where the former is inflexible and the latter relies on the uncertain output on the training stage. Unlike previous works, we propose a novel method from a counterfactual view, namely CLEVER, which is augmentation-free and mitigates biases on the inference stage. Specifically, we train a claim-evidence fusion model and a claim-only model independently. Then, we obtain the final prediction via subtracting output of the claim-only model from output of the claim-evidence fusion model, which counteracts biases in two outputs so that the unbiased part is highlighted. Comprehensive experiments on several datasets have demonstrated the effectiveness of CLEVER.

1 Introduction

Unverified claims have been prevalent online with the dramatic increase of information, which poses a threat to public security over various domains, e.g., public health (Naeem and Bhatti, 2020), politics (Allcott and Gentzkow, 2017), and economics (Kogan et al., 2019). Therefore, fact verification, which aims to automatically predict the veracity of claims based on several collected evidence, has attracted lots of research interests (Liu et al., 2020; Zhong et al., 2020; Vo and Lee, 2021; Jin et al., 2022; Yang et al., 2022).

Existing fact-checking datasets inevitably involve some biases since they are manually collected. For example, Schuster et al. (2019) discover that negation words in claims are highly-correlated with the label 'REFUTES' in the FEVER dataset (Thorne et al., 2018). Such biases may mislead models to explore the spurious correlation between claim patterns and its label without looking into the evidence. In consequence, though models achieve promising performance on biased datasets, they suffer from obvious performance decline on out-of-domain unbiased datasets and are vulnerable to adversarial attacks (Thorne et al., 2019).

To alleviate the aforementioned problems, several debiasing methods have been proposed, which can be mainly grouped into two categories. The first pipeline is based on data augmentation, which utilizes manually-designed schemes, such as word swapping (Wei and Zou, 2019) and span replacement (Lee et al., 2021) to generate additional data for training. However, these methods heavily rely on the quality of augmented data and are difficult to be employed under complicated circumstance, e.g., multi-hop evidence reasoning, due to their inflexible augmentation rules.

The second pipeline aims to downweigh the contribution of biased samples to the training loss of main model, whose inputs are both claim and evidence. Then, the key issue is how to recognize the biased instances. Specifically, Schuster et al. (2019) downweigh the claim involving n-grams that share spurious correlation with labels. Mahabadi et al. (2020) assume instances correctly classified by the bias-only model are biased, where the input of bias-only model is the claim only. Nevertheless, the former lacks the generalization to different types of biases since they only focus on n-grams; the latter relies on the assumption that the outputs of main model and bias-only model regarding the biased instances are similar, which does not always hold (Amirkhani and Pilehvar, 2021).

[*]Equal contribution.
[†]To whom correspondence should be addressed.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6777–6789
July 9-14, 2023 ©2023 Association for Computational Linguistics

Moreover, the inaccurate and unstable outputs of bias-only model during training may mistakenly result in downweighing unbiased samples (Xiong et al., 2021).

Unlike existing works based on augmentation or adjusting the data contribution on the training stage, we propose a novel method from a **C**ounterfactua**L** view for d**E**biasing fact **VER**ification, namely CLEVER, which is augmentation-free and alleviates biases on the inference stage. In general, existing methods fuse the claim and the evidence to make the final prediction, which is equivalent to asking the model to answer a factual question: *What will the output be if the model receives a claim and its corresponding evidence?* Causally, the Total Causal Effect is estimated in this condition, where the output is affected by both the biases in the claim and the claim-evidence interaction information (See the causal graph in Figure 1). In other words, claim biases are entangled with the claim-evidence fused information, making them difficult to be mitigated precisely and thus resulting in a biased output.

To overcome this, we aim to obtain the debiased output by removing claim biases from the Total Causal Effect. Inspired by the progress of counterfactual inference (Sekhon, 2008; Niu et al., 2021), we would expect to ask a counterfactual question: *What would the output be if the model only received a claim?* That is, from a causal perspective, requiring the fact-checking model to learn the Direct Claim Effect solely affected by claim biases. Practically, we first train a claim-evidence fusion model and a claim-only model independently to capture the Total Causal Effect and the Direct Claim Effect, respectively. Then, we subtract the **Direct Claim Effect** from the **Total Causal Effect** on the inference stage to obtain the Total Indirect Effect, which is the final debiased prediction.

Taking Figure 1 as an example, the claim is spuriously correlated with the false label 'REFUTES' due to the phrase 'did not'. Therefore, the Direct Claim Effect inclines to the label 'REFUTES' since it is affected by the claim only. However, though the probability of wrong prediction 'REFUTES' in Total Causal Effect is still the largest, the prediction is turned towards the ground-truth label 'SUPPORTS' via using the Total Indirect Effect as the final output, where the high probability of 'REFUTES' induced by claim biases is counteracted. As biases have been mitigated, the Total Indirect Effect reflects the intrinsic claim-evidence interaction information, leading to an unbiased prediction.

Overall, the main contributions can be summarized as follows:

- We open up a new counterfactual pipeline for debiasing fact verification by analyzing the biased problem from a causal view.

- We propose a novel debiasing method CLEVER, which is augmentation-free and mitigates biases on the inference stage.

- Comprehensive experiments are conducted to validate the effectiveness of CLEVER, where the results demonstrate the superiority and the in-depth analysis provides the rationality.

2 Related Work

In this section, we briefly review the related literature in both domains of fact verification and debiasing strategy.

2.1 Fact Verification

Recent years have witnessed the rapid development of research on fact verification. Since the unified benchmark dataset FEVER along with the shared task were proposed (Thorne et al., 2018), most researchers utilize them to evaluate the model performance. Generally, the fact-checking task mainly consists of three separate parts, i.e., document retrieval, evidence selection, and claim verification. Existing works mainly focus on the last subtask and employ traditional and widely used methods (Hanselowski et al., 2018; **?**) to retrieve relevant documents and evidence. Early works treat fact verification as a natural language inference (NLI) task and apply methods from NLI to perform verification (Chen et al., 2017; Ghaeini et al., 2018). Then, to capture more fine-grained semantic consistency between claims and the evidence, a series of methods have been proposed to promote the claim-evidence interaction by formulating them as graph-structure data (Zhou et al., 2019; Liu et al., 2020; Zhong et al., 2020). Besides, inspired by the strong representation ability of pretrained language models (PLM), some works attempt to fine-tune PLM on fact-checking datasets and achieve promising results (Lee et al., 2020; Subramanian and Lee, 2020). Recently, researchers have paid more attention to explainable fact verification, which requires a model to produce both veracity prediction and

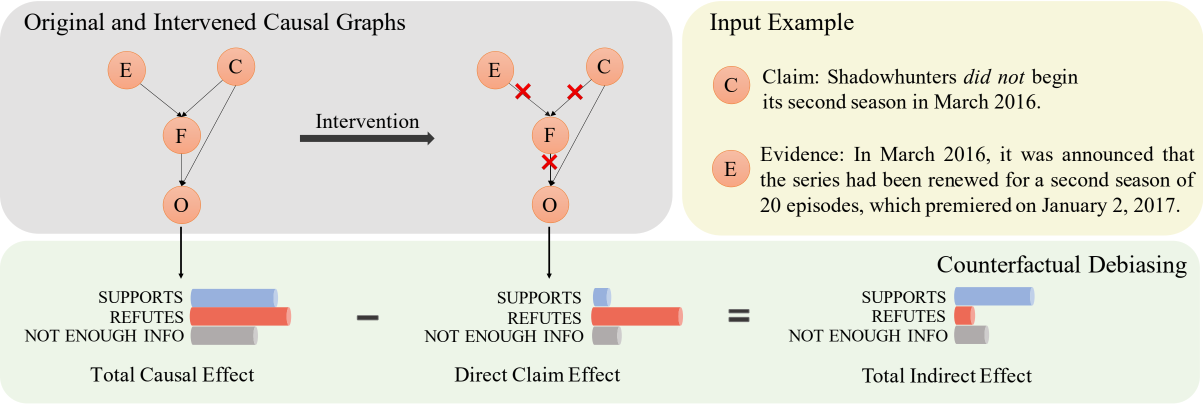

Original and Intervened Causal Graphs
E
C
F
O
Intervention
E
C
F
O
Input Example
C
Claim: Shadowhunters *did not* begin its second season in March 2016.
E
Evidence: In March 2016, it was announced that the series had been renewed for a second season of 20 episodes, which premiered on January 2, 2017.
Counterfactual Debiasing
SUPPORTS
REFUTES
NOT ENOUGH INFO
Total Causal Effect
–
SUPPORTS
REFUTES
NOT ENOUGH INFO
Direct Claim Effect
=
SUPPORTS
REFUTES
NOT ENOUGH INFO
Total Indirect Effect

Figure 1: The causal view of proposed framework CLEVER. The nodes with ‘F’ and ‘O’ denote the claim-evidence fused information and the model output, respectively. We take a typical sample in the biased FEVER dataset as input, where the label is ‘SUPPORTS’ and the strong correlation between the phrase ‘did not’ and label ‘REFUTES’ exists. The output in original graph (Total Causal Effect) is affected by two sources, i.e., claim and claim-evidence fused information. After the intervention via cutting off the fusion path, the output (Direct Claim Effect) is solely influenced by the claim, which contains biases that mislead the model to produce spurious label prediction. To mitigate such biases, a subtraction scheme is proposed to obtain the Total Indirect Effect, which inclines to the true debiased distribution. Note that a path from evidence to output does not exist since there is no obvious bias in the evidence that affects the outcome.

its corresponding explanation (Kotonya and Toni, 2020a,b).

2.2 Debiasing Strategy

Although the aforementioned fact-checking methods have achieved promising performance on the FEVER test set, it is demonstrated that they lack robustness since they learn biases (shortcuts) from claims in datasets instead of performing reasoning over pieces of evidence. To this end, several unbiased and adversarial datasets are proposed to evlatuate the model robustness and reasoning ability (Thorne et al., 2019; Schuster et al., 2019). Existing debiasing strategies in fact verification can be roughly divided into two groups:

1) *Data-augmentation-based pipeline*: In this group, methods aim to generate unbiased samples and incorporate them into training, with the expectation that the proportion of biased instances will be downgraded, resulting in a more unbiased model. In detail, Wei and Zou (2019) utilize random word swapping and synonym replacement to obtain new training data. Lee et al. (2021) design a cross contrastive strategy to augment data, where original claims are modified to be negative using the generation model BART (Lewis et al., 2020) and the evidence are changed via span replacement to support such negative claims.

2) *Weight-regularization-based pipeline*: The motivation of methods in this pipeline is to reduce the contribution of biased samples to the final loss computation, thus models may attach importance to the unbiased data. Next, the problem is transformed into how to filter the biased instances out of the full dataset. Schuster et al. (2019) utilize Local Mutual Information to obtain the n-grams that are highly correlated with a specific label. Then, the claims involving such n-grams are downweighed. Mahabadi et al. (2020) employ a bias-only model to capture biases in claims and assume the unevenness of output label distribution is positively correlated to the confidence of biased instances. However, the confidence estimation is inaccurate observed by some researchers and some calibration methods are further proposed to adjust the estimation (Xiong et al., 2021; Amirkhani and Pilehvar, 2021). Besides, works following this pipeline have also been developed in the related task natural language inference (He et al., 2019; Clark et al., 2019, 2020).

Apart from the mentioned debiasing research pipeline in fact verification, much attention has been paid to incorporating causal inference techniques to obtain more unbiased model. Representative works include counterfactual inference for exposure biases in recommender systems (Tan et al., 2021), implicit knowledge biases and object appearance biases in computer vision (Niu et al., 2021; Sun et al., 2021). However, such pipeline is still under-explored in fact verification. Inspired by these works, we open up a new debiasing pipeline

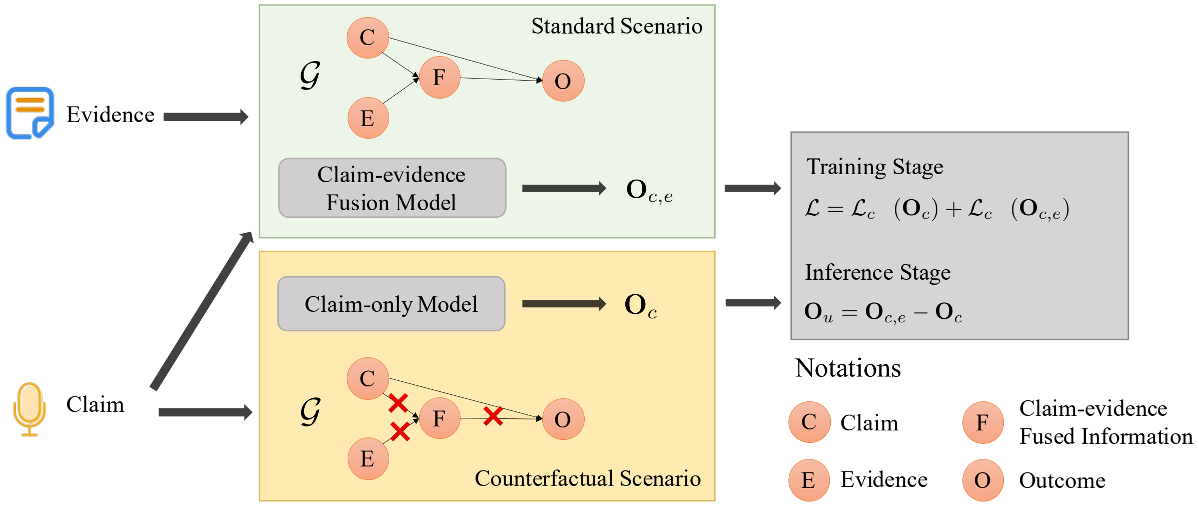

Evidence
Claim
Standard Scenario
$\mathcal{G}$
C
F
E
O
Claim-evidence Fusion Model
$\mathbf{O}_{c,e}$
Claim-only Model
$\mathbf{O}_c$
$\mathcal{G}$
C
F
E
O
Counterfactual Scenario
Training Stage
$\mathcal{L} = \mathcal{L}_c \quad (\mathbf{O}_c) + \mathcal{L}_c \quad (\mathbf{O}_{c,e})$
Inference Stage
$\mathbf{O}_u = \mathbf{O}_{c,e} - \mathbf{O}_c$
Notations
C Claim
F Claim-evidence Fused Information
E Evidence
O Outcome

Figure 2: The proposed framework CLEVER. We simulate the standard and counterfactual scenarios via training a claim-evidence fusion model and a claim-only model independently. The final prediction $\mathbf{O}_u$ is obtained by subtracting the output of counterfactual scenario $\mathbf{O}_c$ from that of standard scenario $\mathbf{O}_{c,e}$.

for fact verification from a counterfactual view. Compared to the existing two pipelines, our proposed method is augmentation-free and mitigates biases on the inference stage.

3 Method

In this section, we introduce the proposed debiasing framework CLEVER in detail. Firstly, we provide some background information of fact verification. Then, we describe the method from a causal view. Finally, we elaborate the detail of training and inference. The overview of CLEVER is shown in Figure 2.

3.1 Preliminary

3.1.1 Task Formulation

Given a claim c and its corresponding evidence set $\{e_1, e_2, \ldots, e_n\}$, a fact-checking model is required to predict the veracity of claim, i.e., the evidence support, refute, or lack enough information to justify the claim.

3.1.2 Causal View of Fact Verification

The causal graph is mathematically a directed acyclic graph, where vertices denote variables and the edge represents the effect from the start vertex to the end vertex.

The causal view of fact verification is represented as a graph $\mathcal{G}_o = \{\mathcal{V}, \mathcal{E}_o\}$, where $\mathcal{V}$ contains four variables with each represents the claim (C), the evidence (E), the fusion of claim and evidence (F), and the output (O), respectively (See the standard scenario in Figure 2). In counterfactual scenario, we expect to capture biases in the claim, so we solely preserve the edge from claim to output. Then, we obtain an intervened causal graph $\mathcal{G}_i$, c.f., the counterfactual scenario in Figure 2.

3.2 The Proposed Framework: CLEVER

In this part, we specifically introduce how to obtain debiased predictions using the counterfactual inference technique.

The first step of counterfactual inference is establishing an imagined scenario different from standard settings. In our task, as shown at the top half of Figure 2, the standard setting is that the outcome is affected by the claim and its corresponding evidence simultaneously in the causal graph $\mathcal{G}_o$. In practice, we take both claim c and evidence $\{e_1, e_2, \ldots, e_n\}$ as inputs to simulate such setting, which can be formulated as:

$$\mathbf{O}_{c,e} = f_s(c, e_1, e_2, \ldots, e_n) \qquad (1)$$

where f_s denotes the claim-evidence fusion model, n is the number of evidence, and $\mathbf{O}_{c,e} \in \mathbb{R}^L$ denotes the predicted class distribution (L is the number of class).

Then, a key problem in our framework is how to design a counterfactual scenario for debiasing. Causally, if we expect to estimate the effect of a variable on the outcome, we can give the variable a specific treatment while keep other variables unchanged. Since the target of our work is to obtain the unbiased outcomes affected by both claim and evidence, the treatment is to make the claim-evidence fusion information unavailable for the fact-checking model. In other words, as shown at the bottom half of Figure 2, we create a counterfactual scenario $\mathcal{G}_i$ via intervention on the original causal graph $\mathcal{G}_o$, where the edge from the fused information of claim-evidence pair to the outcome is cut off. In practice, claims are solely fed into a fact-checking model f_b (i.e., claim-only model) to simulate the absence of claim-evidence information and require the model to produce prediction $\mathbf{O}_c \in \mathbb{R}^L$ based on claims solely,

$$\mathbf{O}_c = f_b(c) \tag{2}$$

The second step is comparing the outcomes under standard and counterfactual settings. The output of claim-only model $\mathbf{O}_c$ is biased that simply relies on the spurious correlation between claim patterns and labels. To reduce such biases, inspired by the Potential Outcomes Model (Sekhon, 2008), we subtract $\mathbf{O}_c$ from $\mathbf{O}_{c,e}$ with a hyperparameter α (named bias coefficient that controls the extent of bias) and obtain the counterfactual debiased output $\mathbf{O}_u$,

$$\mathbf{O}_u = \mathbf{O}_{c,e} - \mathbf{O}_c \tag{3}$$

In this way, the probability of false biased prediction is decreased while the predicted probability of ground truth is relatively higher.

Training and Inference At training stage, as biases are mainly involved in claims, we expect that the claim-only model captures such biases so that they can be reduced via the subtraction scheme. Motivated by this, we encourage the output of claim-only model $\mathbf{O}_c$ to represent the biased label distribution by imposing a classification loss on $\mathbf{O}_c$. Similarly, $\mathbf{O}_{c,e}$ is also supervised to mine the claim-evidence interaction. Formally, the objective function can be written as:

$$\mathcal{L} = \mathcal{L}_{clf}(\mathbf{O}_c) + \mathcal{L}_{clf}(\mathbf{O}_{c,e}) \tag{4}$$

where $\mathcal{L}_{clf}$ denotes the cross entropy loss.

At inference stage, since the outcome in counterfactual scenario $\mathbf{O}_c$ is biased after training, we intuitively reduce it via subtraction from the outcome in standard scenario $\mathbf{O}_{c,e}$, c.f., Eq. (3).

Discussion Overall, the proposed framework CLEVER consists of a claim-evidence model and a claim-only model, which are utilized to capture the interaction information and biased information, respectively. As we introduce a new pipeline for debiasing, here, we further emphasize the difference and merits of CLEVER compared with the weight-regularization-based approach, which is the most popular way for debiasing in this task. Firstly, we do not rely on the assumption that such two models produce similar outputs for biased instances as weight-regularization-based approaches do. Besides, we avoid utilizing the uncertain output of claim-only model to adjust the training loss of claim-evidence model. By contrast, we independently train the claim-evidence and claim-only model and propose a simple yet effective scheme to obtain debiased results on the inference stage.

4 Experiments

In this section, we conduct both quantitative and qualitative experiments on several public datasets to demonstrate the effectiveness of our proposed method CLEVER.

4.1 Experimental Setup

4.1.1 Dataset and Evaluation Metric

We utilize three categories of datasets to evaluate our method from different views.

Single-hop datasets. We utilize a biased training set FEVER-Train (Thorne et al., 2018) to train models and use an unbiased dataset FEVER-Symmetric (Schuster et al., 2019) and an adversarial dataset FEVER-Adversarial (Thorne et al., 2019) to test models, closely following existing works (Mahabadi et al., 2020; Lee et al., 2021; Xiong et al., 2021). Furthermore, we introduce a new unbiased subset of FEVER-Dev, namely FEVER-Hard[1], where all samples cannot be correctly classified using claims only. That is, the samples in FEVER-Hard are unbiased since there are no shortcuts in the claim misleading the model to explore. Therefore, it can be used to evaluate the model ability to perform evidence-to-claim reasoning indeed, i.e., the debiasing performance.

Multi-hop datasets. Besides, existing works only focus on the simple one-hop reasoning scenario, where each sample in the current train set and test set only involves one piece of evidence. However, in real-world applications, some complicated conditions require multi-hop reasoning capability. Thus, to further validate the debiasing performance under the multi-hop setting, we augment

[1] We omit the prefix 'FEVER' for conciseness in following paragraphs since all unbiased and adversarial datasets are derived from the original FEVER dataset.

Dataset	Symmetric	Hard	Adversarial
BERT-base	72.08 ± 0.51	78.05 ± 0.54	61.93 ± 1.31
EDA	72.93 ± 0.48	78.22 ± 0.61	62.12 ± 1.02
CrossAug	78.88 ± 0.46	82.19 ± 0.31	61.72 ± 0.45
ReW	73.39 ± 0.71	78.43 ± 0.52	64.52 ± 1.49
PoE	76.43 ± 0.64	80.51 ± 0.70	67.21 ± 1.69
PoE-TempS	76.89 ± 0.86	81.13 ± 0.33	67.05 ± 2.30
PoE-Dirichlet	78.55 ± 0.97	82.31 ± 0.82	66.98 ± 1.77
CLEVER (ours)	**84.73 ± 0.69**	**90.17 ± 0.75**	**68.34 ± 0.94**
Δ Improvement	+ 17.55%	+ 15.53%	+ 10.35%

Table 1: The performance comparison between our proposed method CLEVER and baselines. Three datasets are introduced to verify the model performance under an unbiased circumstance. The best result on each dataset is highlighted in boldface and the runner-up is underlined. The improvement in terms of percentage compared to the BERT-base is shown in the last row.

the dataset Train and Dev with instances consisting of several pieces of evidence and generate two multi-hop datasets Train-MH and Dev-MH. Then, we add the multi-hop instances that cannot be predicted correctly using claims only into Hard and form a new test set Hard-MH.

Multi-domain datasets. Moreover, we utilize a dataset namely MultiFC to evaluate the performance of debiasing methods under a multi-domain setting. MultiFC consists of claims collected from various domains on the website, e.g., politics, sports, and entertainment. The claim in FEVER-derived datasets under single-hop and multi-hop settings is manually-created based on Wikipedia, which is usually limited to commonsense fact such as a celebrity's nationality. Thus, we introduce the mentioned real-world dataset MultiFC to examine whether the proposed method works facing claims with varied forms. Note that we train all models without using 'NOT ENOUGH INFO' samples to keep a similar data distribution with the test set, since these test sets only involve 'SUPPORTS' and 'REFUTES' samples. Following previous works (Lee et al., 2021), we use label classification accuracy as the metric.

4.1.2 Baselines

We compare our proposed method with several baselines from both two existing pipelines:

Data-augmentation-based methods: 1) EDA (Wei and Zou, 2019). They swap words and replace synonym to generate new training samples. 2) CrossAug (Lee et al., 2021). They design a cross contrastive strategy to augment data, where original claims are modified to be negative and the evidence is changed to support such negative claims and refute the original claims.

Weight-regularization-based methods: 1) ReW (Schuster et al., 2019). They downweigh the samples which involve n-grams highly correlated to labels. 2) PoE (Mahabadi et al., 2020). They downweigh samples with spurious class distribution outputed from the bias-only model. 3) MoCaD (Xiong et al., 2021). They propose a calibration method to adjust the inaccurate predicted class distribution from bias-only models. Specifically, two calibrators (i.e., temperature scaling and Dirichlet calibrator) are employed in this work. We utilize such methods to further optimize the model PoE, forming two variants namely PoE-TempS and PoR-Dirichlet.

4.2 Performance Comparison

The overall performance of our proposed method CLEVER and several strong baselines is shown in Table 1. We can see that CLEVER outperforms all existing methods from different pipelines by a significant margin on all datasets. More specifically, we have the following observations:

Firstly, the performance gain of CLEVER is more consistent on all datasets than that of previous methods. We can observe that the runner-up on each dataset is different while CLEVER achieves the best performance on all datasets. More specifically, compared to the vanilla BERT model (i.e., BERT-base) without any debiasing method, CLEVER advances by 17.55% and 15.53% on two unbiased datasets Symmetric and Hard, re-

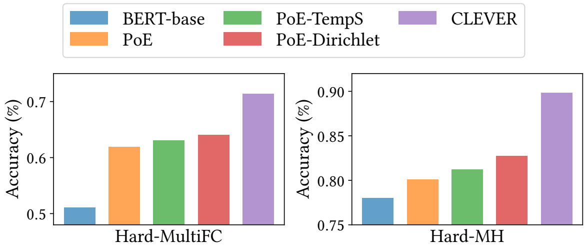


Figure 3: **Left**: The performance comparison between our proposed method CLEVER and several baselines on the real-life dataset MultiFC. **Right**: The performance comparison between our proposed method CLEVER and baselines under the complicated multi-hop reasoning circumstance.

spectively. Furthermore, most baselines, especially CrossAug, perform relatively worse on the dataset Adversarial, since debiasing methods are always specially designed for avoiding learning biases in claim while do not explicitly consider adversarial attacks. By contrast, our proposed method still achieves a promising result on it (about 10% performance improvement upon the BERT-base), which demonstrates the generalization ability of our method to handle both adversarial and biased data. This is probably because our proposed method utilize a claim-only model to adaptively capture the shortcuts the model may be prone to fall into, instead of heuristically defining the biased phrases or relying on the inaccurate output of the bias-only model in existing methods.

Secondly, it is worth noting that the methods EDA and ReW always perform much worse than the other approaches. This is mainly due to the different ways of capturing biases. EDA and ReW are similar that they both consider biases at a specific word- or phrase-level. EDA replaces some specific words with synonyms and ReW predefines biased n-grams that co-occur frequently with a specific label, which may be inflexible since it is hard to cover all biases in this way. By contrast, rest of methods, including ours, all train models to automatically augment samples and capture biases, which are of better generalization ability to learn different patterns of biases.

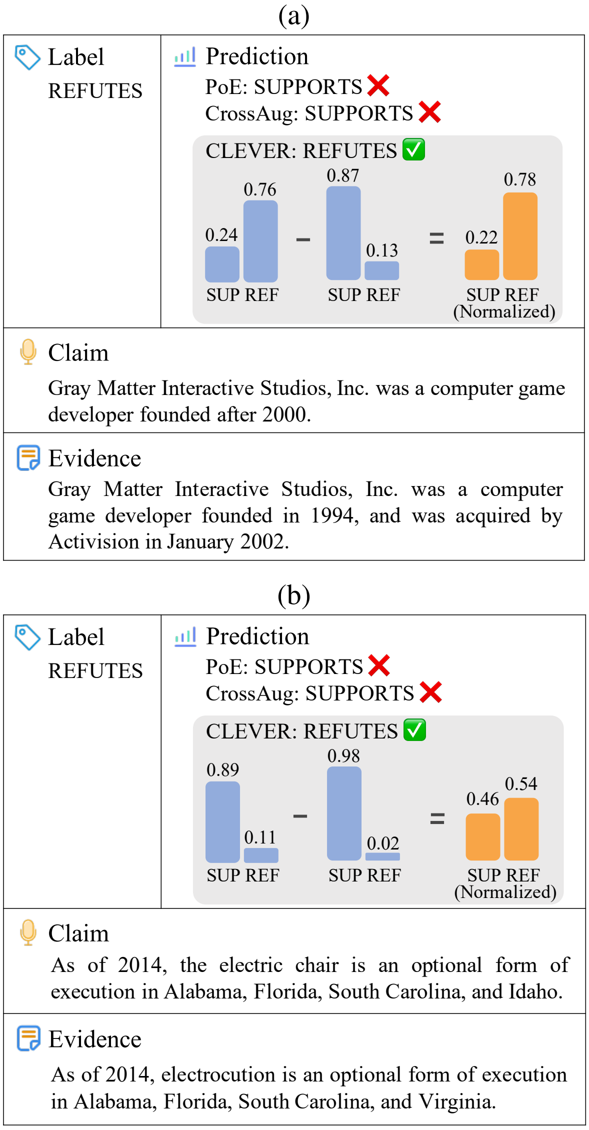
4.3 Study of Multi-hop Circumstance

Existing methods only utilize samples with single evidence to evaluate the debiasing performance, however, we argue that more complicated reasoning circumstance should be considered since a claim may be verified via several pieces of evidence in the realistic scenario. Therefore, we further validate debiasing methods under a multi-hop reasoning setting, where instances with more than one piece of evidence are involved in both biased validation set Dev-MH and unbiased set Hard-MH. Similar to the Hard dataset in the single-hop scenario, Hard-MH also involves all samples model makes wrong prediction based on the claim only. Since data-augmentation methods are hard to be adapted to such complicated scenario, we compare our method CLEVER with baselines from the weight regularization based pipeline.

As shown in the right part of Figure 3, CLEVER consistently outperforms its competitors by a significant margin (about 7% absolute improvement compared with the runner-up PoE-Dirichlet), which demonstrates its effectiveness of handling complicated data.

4.4 Performance on the Real-life Multi-domain Dataset MultiFC

We further validate the debiasing performance of our proposed method CLEVER on the dataset MultiFC, which contains plenty of claims collected from the several websites. To fit the output of our model, we merge the ‘true’, ‘mostly true’, and ‘half true’ to one class, and similarly merge the ‘pants on fire’, ‘false’, and ‘mostly false’ into one class. We train the model on training set of MultiFC and obtain the performance on the unbiased subset of MultiFC (Hard-MultiFC), on which the model cannot predict correctly using the claim solely. The results are shown in the left part of Figure 3, which demonstrates the effectiveness of our method on the real-life dataset. Furthermore, it is worth noting that the performance gap between BERT-base and debiasing methods is much larger than that on manually-created datasets in Table 1. The reason is probably that the bias in real scenario is more severe than that in handcrafted datasets, which only involving textual biases. For example, claims in real website involve entity biases in addition to textual biases. Entity may refer to a celebrity, such as Donald Trump, which is usually spuriously correlated with the fake claim, i.e., an entity bias. Thus, it is significant and urgent to develop debiasing methods to resist the negative impact of biases to fact checking models.



(a) Claim: Gray Matter Interactive Studios, Inc. was a computer game developer founded after 2000.

(a) Evidence: Gray Matter Interactive Studios, Inc. was a computer game developer founded in 1994, and was acquired by Activision in January 2002.

(b) Claim: As of 2014, the electric chair is an optional form of execution in Alabama, Florida, South Carolina, and Idaho.

(b) Evidence: As of 2014, electrocution is an optional form of execution in Alabama, Florida, South Carolina, and Virginia.

Figure 4: Two representative instances where our proposed method CLEVER outputs correct veracity prediction while baselines make mistakes. The bars denote the outputed label distribution, i.e., $\mathbf{O}_u = \mathbf{O}_{c,e} - \mathbf{O}_c$ (Eq. (3)).

4.5 Case Study

In this section, we design some case studies to further analyze the advantages of our proposed method CLEVER on a qualitative aspect. We aim to compare the performance of different models at an instance level. We choose the best debiasing method from each pipeline (i.e., CrossAug and PoE) to carry out the analysis. Specifically, we select representative examples from the dataset Hard that are correctly classified using our method while mistakenly predicted by baselines.

From Figure 4, the top instance shows that the output of claim-evidence fusion model **correctly** inclines to the ground-truth 'REFUTES' while the output of claim-only model is **mistakenly** biased towards 'SUPPORTS'. That is, the claim-evidence fusion model deals with biased instances in a different way from the claim-only model, which echoes the discovery in the previous work (Amirkhani and Pilehvar, 2021). Therefore, PoE downweighs such instance in training objective according to the biased extent of claim-only model would result in performance degradation. However, our method CLEVER separates such outputs of two models in training and the predicted probability of ground-truth label is further enlarged via subtraction on inference stage.

The bias in the bottom instance is mainly induced by the word 'is', which is highly correlated with the label 'SUPPORTS'. Data-augmentation based methods simply insert negations or antonyms, such as transforming 'is' to 'is not', are hard to capture the intrinsic conflict between the claim and the evidence. In this instance, the conflict lies between 'Idaho' and 'Virginia', not the word 'is'. Therefore, augmenting training instances via inserting negations or antonyms contribute little to such complex reasoning circumstance. However, our approach CLEVER directly captures both claim-evidence interactions and claim biases which is augmentation-free. Note that the biased label distribution is alleviated in the claim-evidence fusion model, i.e., the probability of wrong prediction 'SUPPORTS' is decreased to 0.89 from 0.98 (See Figure 4(b)), since it partly pays attention to the evidential information. Though the distribution is still biased towards the falsity due to the strong bias between 'is' and the label 'SUPPORTS', CLEVER can eliminates such bias in both models via subtraction so as to highlight the intrinsic evidential segment, thus providing the correct prediction.

5 Conclusion

In this paper, we have proposed a novel counterfactual framework CLEVER for debiasing fact-checking models. Existing works mainly follow the data augmentation pipeline and the weight regularization pipeline. Unlike them, CLEVER is augmentation-free and mitigates biases on inference stage. In CLEVER, the claim-evidence fusion model and the claim-only model are independently trained to capture the corresponding information.

On the inference stage, based on the potential outcome model in the domain of causal inference, a simple subtraction scheme is proposed to mitigate biases. Comprehensive quantified and qualified experiments have demonstrated the superiority of CLEVER.

Limitations

In this part, we show limitations of our work by categorizing wrong predictions outputed by our method CLEVER into two groups.

The first type of error is induced by the unconspicuous biased features of claims. For example, the claim *Scandinavia includes the remote Norwegian islands of Svalbard and Jan Mayen.* does not contain obvious biases so that the output of claim-only model cannot represent the biased distribution. Therefore, subtracting such output fails to mitigate biases but reduces the beneficial claim information instead. These errors may be avoided by employing different strategies for instances with distinct bias extents, which we leave as future work.

The second type of error occurs when high-level reasoning is required, e.g., mathematical computation and multi-hop reasoning, which drops into the scope of model reasoning ability. This work mainly focuses on debiasing fact-checking models that make them concentrate on the intrinsic evidential information. After debiasing, how to enhance the reasoning ability over such information is a promising future direction.

Acknowledgement

This work is supported by National Natural Science Foundation of China (U19B2038, 62141608, 62206291).

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *CSN: Politics (Topic)*.

Hossein Amirkhani and Mohammad Taher Pilehvar. 2021. Don't discard all the biased instances: Investigating a core assumption in dataset bias mitigation techniques. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4720–4728, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qian Chen, Xiao-Dan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *ACL*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *EMNLP*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2020. Learning to model and ignore dataset bias with mixed capacity ensembles. In *EMNLP Findings*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Reza Ghaeini, Sadid A. Hasan, Vivek Datla, Joey Liu, Kathy Lee, Ashequl Qadir, Yuan Ling, Aaditya Prakash, Xiaoli Z. Fern, and Oladimeji Farri. 2018. Dr-bilstm: Dependent reading bidirectional lstm for natural language inference. In *NAACL*.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *EMNLP*.

Yiqiao Jin, Xiting Wang, Ruichao Yang, Yizhou Sun, Wei Wang, Hao Liao, and Xing Xie. 2022. Towards fine-grained reasoning for fake news detection. In *AAAI*, volume 36, pages 5746–5754.

Shimon Kogan, Shimon Kogan, Tobias J. Moskowitz, Tobias J. Moskowitz, and Marina Niessner. 2019. Fake news: Evidence from financial markets.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking: A survey. In *COLING*.

Neema Kotonya and Francesca Toni. 2020b. Explainable automated fact-checking for public health claims. In *EMNLP*.

Minwoo Lee, Seungpil Won, Juae Kim, Hwanhee Lee, Cheoneum Park, and Kyomin Jung. 2021. Crossaug: A contrastive data augmentation method for debiasing fact verification models. In *CIKM*.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? *ArXiv*, abs/2006.04102.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *ACL*.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *ACL*.

Salman Bin Naeem and Rubina Bhatti. 2020. The covid-19 'infodemic': a new front for information professionals. *Health Information and Libraries Journal*.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xiansheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *CVPR*.

Tal Schuster, Darsh J. Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *EMNLP*.

Jasjeet S Sekhon. 2008. The neyman-rubin model of causal inference and estimation via matching methods. *The Oxford handbook of political methodology*, 2:1–32.

Shyam Subramanian and Kyumin Lee. 2020. Hierarchical evidence set modeling for automated fact extraction and verification. In *EMNLP*.

Pengzhan Sun, Bo Wu, Xunsong Li, Wen Li, Lixin Duan, and Chuang Gan. 2021. Counterfactual debiasing inference for compositional action recognition. *Proceedings of the 29th ACM International Conference on Multimedia*.

Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. 2021. Counterfactual explainable recommendation. In *CIKM*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *NAACL*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Evaluating adversarial attacks against multiple fact verification systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2944–2953, Hong Kong, China. Association for Computational Linguistics.

Nguyen Vo and Kyumin Lee. 2021. Hierarchical multi-head attentive network for evidence-aware fake news detection. In *EACL*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP*.

Ruibin Xiong, Yimeng Chen, Liang Pang, Xueqi Chen, and Yanyan Lan. 2021. Uncertainty calibration for ensemble-based debiasing methods. In *NIPS*.

Ruichao Yang, Xiting Wang, Yiqiao Jin, Chaozhuo Li, Jianxun Lian, and Xing Xie. 2022. Reinforcement subgraph reasoning for fake news detection. In *KDD*, pages 2253–2262.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, M. Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *ACL*.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. Gear: Graph-based evidence aggregating and reasoning for fact verification. In *ACL*.

A Dataset Statistics

We show the dataset statistics in Table 2.

Circumstance	Dataset	# SUP	# REF	SUM
Single-hop	Train	100,570	41,850	142,420
	Dev	7,983	8,681	16,664
	Symmetric	379	338	717
	Adversarial	364	402	766
	Hard	679	2,638	3,317
Multi-hop	Train-MH	120,081	41,850	168,424
	Dev-MH	9,214	9,796	19,010
	Hard-MH	855	3,027	3,882
Multi-domain	Train-MultiFC	5,634	4,938	10,572
	Dev-MultiFC	811	708	1,519
	Hard-MultiFC	195	337	532

Table 2: The statistics of datasets that is divided into three groups. These datasets are introduced to evaluate the performance of debiasing methods under different circumstances. 'SUP' and 'REF' is the abbreviation of the label 'SUPPORTS' and 'REFUTES', respectively. '#' stands for the number of.

B Implementation Detail

Following the aforementioned baselines, we employ BERT-base (Devlin et al., 2019) as the backbone model for a fair comparison, i.e., claim-evidence fusion model and claim-only model are two independent BERT models. We finetune BERT with a fully-connected forward layer over the special token [CLS] to obtain the final prediction. The maximum input length is 128, batch size is 32, and the optimizer is Adam with a learning rate of 2e-5; we train the model for 3 epochs and repeat 5 times under different random seed settings, which are all the same as previous works. We conduct all experiments using PyTorch 1.8.0 on a single GeForce RTX 662 3090 GPU with 24GB memory. The training and inference process cost about 1 hour and less than 5 minutes, respectively.

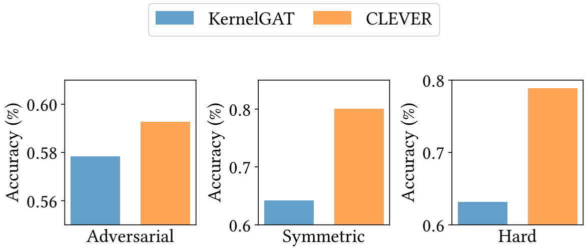


Figure 5: The performance of our proposed debiasing method CLEVER on the graph-based fact-checking model KernelGAT on three unbiased test sets.

C Validating CLEVER on Graph-based Fact-checking Model

Fact verification models can be categorized into two groups, i.e., transformer-based approaches (e.g., BERT-base we utilize in the main experiment) and graph-based approaches. To demonstrate the scalability of our proposed method CLEVER, we further validate it with another fact-checking backbone model, namely KernelGAT, which is a representative graph-based approach. All parameter settings are the same as the original paper reports. As shown in Figure 5, CLEVER obtains the consistent performance gain on all of three test sets when equipping with a graph-based fact-checking model, indicating the scalability of our method, i.e., our proposed method CLEVER can achieve satisfactory debiasing performance on two main groups of fact checking models.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
The limitations section

☑ A2. Did you discuss any potential risks of your work?
there is no risk of our work.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
section 1 and the abstract

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

section 4.1 and appendix A

☑ B1. Did you cite the creators of artifacts you used?
section 4.1

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
section 4.1

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
section 4.1

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
we use all publicly available datasets. The details of these datasets can be seen in their original paper, which has been cited properly in our paper.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
All datasets are English and the details of these datasets can be seen in their original paper, which has been cited properly in our paper.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
appendix A

C ☑ Did you run computational experiments?

section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
appendix B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
appendix B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
table 1 and section 4.2

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
appendix B

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.