Bias Mitigation for Evidence-aware Fake News Detection by Causal Intervention

Junfei Wu
Center for Research on Intelligent Perception and Computing (CRIPAC),
National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences
junfei.wu@cripac.ia.ac.cn

Qiang Liu
Center for Research on Intelligent Perception and Computing (CRIPAC),
National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences
qiang.liu@nlpr.ia.ac.cn

Weizhi Xu
Center for Research on Intelligent Perception and Computing (CRIPAC),
National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences
weizhi.xu@cripac.ia.ac.cn

Shu Wu*
Center for Research on Intelligent Perception and Computing (CRIPAC),
National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences
shu.wu@nlpr.ia.ac.cn

ABSTRACT

Evidence-based fake news detection is to judge the veracity of news against relevant evidences. However, models tend to memorize the dataset biases within spurious correlations between news patterns and veracity labels as shortcuts, rather than learning how to integrate the information behind them to reason. As a consequence, models may suffer from a serious failure when facing real-life conditions where most news has different patterns. Inspired by the success of causal inference, we propose a novel framework for debiasing evidence-based fake news detection[1] by causal intervention. Under this framework, the model is first trained on the original biased dataset like ordinary work, then it makes conventional predictions and counterfactual predictions simultaneously in the testing stage, where counterfactual predictions are based on the intervened evidence. Relatively unbiased predictions are obtained by subtracting intervened outputs from the conventional ones. Extensive experiments conducted on several datasets demonstrate our method’s effectiveness and generality on debiased datasets.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; • **Information systems → Data mining**.

KEYWORDS

fake news detection, model debiasing, causal inference

ACM Reference Format:
Junfei Wu, Qiang Liu, Weizhi Xu, and Shu Wu*. 2022. Bias Mitigation for Evidence-aware Fake News Detection by Causal Intervention. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3477495.3531850

1 INTRODUCTION

The widespread misinformation including fake news on social media has influenced several major events, such as the 2016 U.S. presidential elections [1], COVID-19 infodemic [10]. The subtly manipulated content and large scale make it difficult and laborious for fact-checkers to debunk false claims in time. In consequence, pattern-based [3, 8, 14, 19, 23, 24] and evidence-based [9, 15, 17, 18, 21, 22] fake news detection has been proposed to deal with these phenomena. In this work, we focus on the robustness of evidence-based approach, which integrates external evidences to reason whether the news is consistent with evidences and predicts its veracity.

Although there has already been steady progress made in the research of evidence-based fake news detection in recent years, state-of-the-art methods are vulnerable to unexpected dataset biases. These biases are always introduced during data collection. It has recently been demonstrated that detecting fake news by only using the claim can even achieve an approximate performance to that of using both claim and evidence [7]. These results contradict our assumption that veracity can only be determined over evidence, and confirm the existence of some signal patterns and biases within news which dominate the veracity judgement. Several research has been made to investigate the biases introduced when collecting dataset, owing to annotation artifacts [6, 13]. In addition, some biases are formed in the process of sample selection. Specifically for news detection, fact-checkers always select the news according to the dissemination scope, potential influence, and related fields. As such, some give-away words are closely related to certain veracity

*To whom correspondence should be addressed.
[1]Code available at https://github.com/CRIPAC-DIG/CF-FEND

SIGIR ’22, July 11–15, 2022, Madrid, Spain.

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531850

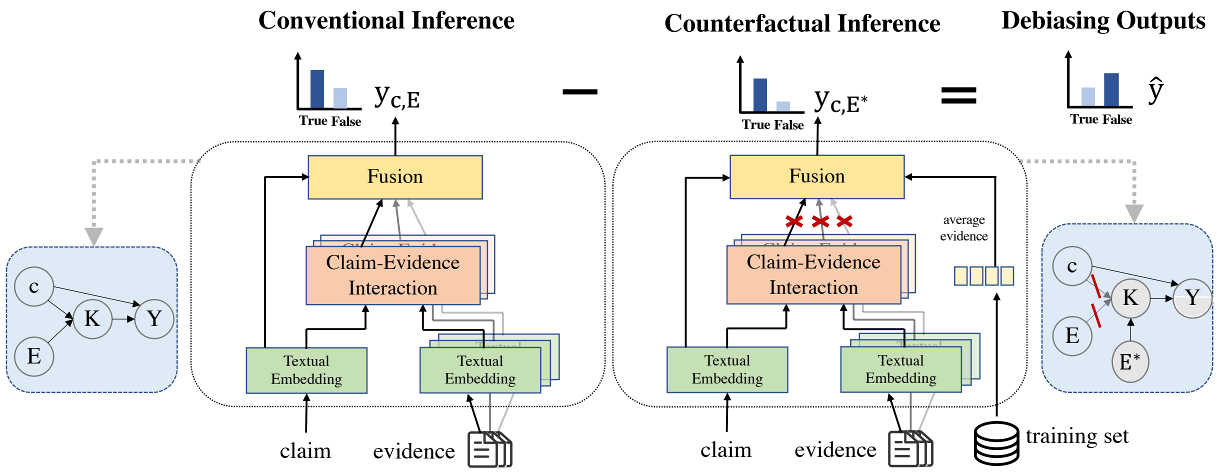


Figure 1: The architecture of the counterfactual framework. In the inference stage, the claim and evidence are first input and passed through modules to generate conventional outputs. Next, our framework replaces relevant evidences with the masked evidence to make a counterfactual inference. Finally, the debiased outputs are obtained by the subtraction of the above two outputs. Corresponding causal graphs are depicted next to the framework, including the variables of claim c, evidence E, mediator K and prediction Y.

labels, as news of the same event may contain similar writing styles and text patterns. For example, phrases like "golden medal" and "fight" often occur in news related to the sports events, forming specific patterns unintentionally. As a consequence, deep models tend to utilize these spurious correlations between keywords in news and its labels to predict, and thus fail to perform reasoning over the evidence. This issue seriously affects models' robustness and generalization, especially when transferring models to datasets that may have different patterns.

To deal with the issue above, we propose a novel method to mitigate bias by causal intervention [5, 12]. Recently some researchers have introduced causal inference to successfully solve the bias problem in vision question answering [11] and text classification [16]. Their debiasing methods can better analyze the existence of biases and remove them. Inspired by their work, we decompose the problem of evidence-based fake news detection from the perspective of causal inference. In conventional inference, the task is formulated as a question: "what will the veracity be if seeing both news and evidence?" Under this circumstance, the information of news, evidence, and the knowledge behind them have combined altogether, making up the total causal effects on predictions where the biases in news patterns are hard to detach clearly. Alternatively, we can review this problem in a counterfactual way and ask another question: "what will it be if only see the news without access to evidence?" The output distribution to this question captures biases in the form of direct causal effects of news on the prediction. Because the information of evidence is intervened, and the strong correlation between news patterns and labels is the only basis of the decision. Then we can subtract direct effects from total effects to counteract the contribution of biases to the final prediction. Such inference process is the same as that when humans make predictions to some extent, as humans can distinguish what is important or deceptive and make an unbiased inference in a complex scenario.

Therefore, we incorporate causal thought to alleviate biases in fake news detection and improve models' capability of reasoning. This debiasing framework consists of two phases, including biased training and unbiased testing. Specifically, the evidence-based model is first trained with both news and evidence as inputs on the biased dataset just like common work. While in the testing stage, the model is asked to perform extra counterfactual reasoning. This process is under the circumstance of causal intervention of blocked evidence information. And it aims to identify how biases affect the distribution of outputs. At last, the debiased prediction is obtained by subtracting biased outputs from conventional outputs. It is worth noting that our framework doesn't require a customized training process, and it is convenient to implement.

Our main contributions of this work are as follows:

- To mitigate the bias problem, we propose a novel framework for debiasing fake news detection models by causal intervention.
- We conduct extensive experiments on several real-life datasets with advanced base models. The results demonstrate that our counterfactual framework shows strong performance in mitigating biases and is model-agnostic.

2 METHOD

Problem formulation We formulate evidence-based fake news detection as a classification problem, where the object of the base model is to predict the veracity class y of a claim c given its content and relevant evidences $E = (e_1, e_2, ..., e_n)$. And our debiasing framework aims to help the model learn a set of the optimized parameter Θ_M, and make it more robust to biases within datasets.

Motivated by the effectiveness of existing debiasing work by causal inference, we leverage causal theory to analyze evidence-based fake news detection essentially. And we propose a novel counterfactual framework for debiasing fake news detection. The overall structure of our framework to intervene base model is depicted in Figure 1. In this section, we detail the biased training and bias removal over a causal graph to show how our framework generates debiased outputs.

2.1 Biased Training

Our framework first decomposes and modifies the base model to a unified structure, including a textual embedding module, claim-evidence interaction module, and fusion module. Then base models can be embedded into our framework.

During the training stage, the framework treats the base model like traditional work, and trains it on the original biased dataset. The claim and evidences are taken as inputs to the model simultaneously and are passed through each layer hierarchically. First, the claim and evidences are mapped to token embeddings by the embedding layer. Then the claim-evidence interaction module generates the claim-aware representation of evidences $T_E = i(c, E)$ after receiving the token embeddings, where $i(\cdot)$ represents the interaction function. The textual embedding layer also transforms the claim into a dense feature T_c to encode its semantics. Finally, the intermediate representations of claim T_c and evidences T_E are both input to the fusion module to aggregate information. This module outputs predicted distribution for each category, which can be denoted as:

$$y_{c,E} = h(T_c, T_E) \tag{1}$$

where $y_{c,E}$ is the output probability distribution, and $h(\cdot)$ represents the fusion function. For brevity, we represent the overall model as a function $y_{c,E} = f(c, E)$. Under this circumstance, the total causal effect of both claim and evidence on the answer is estimated by the model. To optimize the trainable parameters Θ of the model, we utilize cross-entropy loss as the training objective and minimize it, which is written as:

$$\mathcal{L}_\Theta(y, y_{c,E}) = -\sum_j y_j \log y_{E,j} \tag{2}$$

where y is the ground-truth label in binary vector.

2.2 Bias Removal over Causal Graph

After biased training, the model unintentionally memorizes and relies on some spurious correlations to predict in addition to learning reasoning. Therefore, we introduce a causal graph to view the reasoning operation of the model, and show how to capture and remove bias.

2.2.1 Causal Graph. We first take a look at evidence-based fake news detection from a causal perspective. The complete causal graph is illustrated in the conventional inference in Figure 1. It is in the form of a directed acyclic graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is a set of variables including the news c, evidences E, the mediator $K_{c,E}$ of news and evidence, as well as the prediction Y. $\mathcal{E}$ is a set of causal links denoting causal relations between these variables. For instance, $(c, E, K_{c,E}) \rightarrow Y$ is the total causal effect (TE) as all variables are involved, indicating how the claim and evidences work together to affect the mediator and the final prediction. While the causal link of $c \rightarrow Y$ is the natural direct effect (NDE) of news patterns on the answer which introduces biases, as illustrated in the causal graph in Figure 1. And it is difficult to be mitigated straightforwardly. The relatively unbiased inference should be based on the total indirect effect (TIE) which is the knowledge behind the claim and evidence, and it can be achieved by comparing TE and NDE.

2.2.2 Bias Distillation. Counterfactual inference means inferring possible results by assuming conditions that some variables do not work. Then we can distinguish how many contributions the remaining variables make to the results. In the context of fake news detection, the corresponding counterfactual inference is to ask the model to predict veracity with only access to the claim and the information of evidence being blocked. Specifically, we conduct the causal intervention to wipe out all the in-coming links of variable K and assign a certain value representing blocked evidence. Generally speaking, when only the news is provided, the veracity is hard to predict due to the lack of sufficient information for reasoning. Consequently, the model has to depend on the obvious correlations between news patterns and its label which introduces biases. We represent the evidence with the treatment of being blocked as E^*, and the counterfactual output of the model can be denoted as:

$$y_{c,E^*} = f(c, E^*) \tag{3}$$

However, in practice, the evidence-based neural model can't simply disable the component of processing evidence and receive the void value of evidences as input. To represent the blocked evidence representation T_{E^*} after the causal intervention, our framework uses the average claim-aware evidence feature which is obtained from the interaction module on the whole training set, following previous work [20]. We assume that the average feature can represent the inference that people can make only based on the existing knowledge without access to additional evidence information. Then the average feature is applied to replace the original evidence feature $T_{c,E}$ and is input to the fusion module with T_c. Thereafter, the counterfactual output can also be written as $y_{c,E^*} = h(T_c, T_{E^*})$.

Under this intervened circumstance, original causal links other than $c \rightarrow Y$ are interfered with, and thus the signals of relevant evidences are blocked.

2.2.3 Bias Removal. Finally, to mitigate bias in the outputs, the model is asked to carry out both a traditional prediction $y_{c,E}$ and a counterfactual prediction y_{c,E^*} simultaneously. As mentioned before, $y_{c,E}$ represents all effects of variables, and y_{c,E^*} captures the extent of bias within the dataset. Then we perform element-wise subtraction following [16] to obtain a relatively unbiased output $\hat{y}$:

$$\hat{y} = y_{c,E} - \lambda \cdot y_{c,E^*} \tag{4}$$

where λ is a parameter introduced to avoid insufficient or excessive subtraction.

3 EXPERIMENTS

3.1 Experimental Configurations

3.1.1 Datasets. We utilize two datasets [7] whose claims and corresponding labels are collected from two major fact-checking websites PolitiFact[2] and Snopes[3]. These datasets mainly focus on political

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

Table 1: Experimental results of base models in the setting without (BASE) or with the counterfactual framework (+CF). "F1-Ma" and "F1-Ma" represent the metrics F1-Macro and F1-Micro respectively. The superior results are highlighted in boldface.

Method	Snopes		Snopes Hard		PolitiFact		PolitiFact Hard	
	F1-Ma	F1-Mi	F1-Ma	F1-Mi	F1-Ma	F1-Mi	F1-Ma	F1-Mi
BERT	0.676	**0.723**	0.431	0.472	0.659	0.663	0.292	0.294
BERT+CF	**0.677**	0.722	**0.550**	**0.564**	**0.668**	**0.671**	**0.371**	**0.372**
MAC	0.678	**0.725**	0.525	0.575	0.658	0.661	0.365	0.366
MAC+CF	**0.680**	0.724	**0.587**	**0.598**	**0.660**	**0.661**	**0.441**	**0.443**

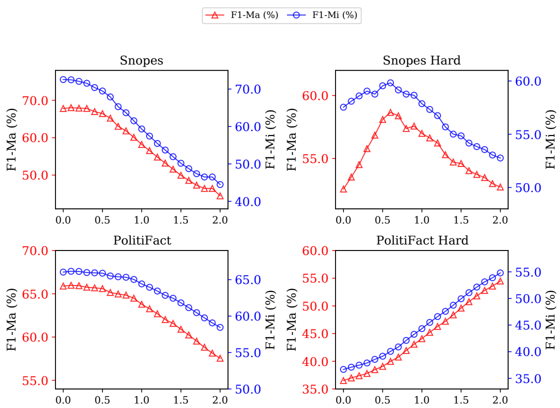


Figure 2: The performance of different coefficient λ.

events. The claim is used as a query to retrieve top-10 relevant snippets as evidence. Then the evidences from the same website origin as the claims and the claims with non-veracity labels is filtered. For Snopes claim, we merge $false$, $mostly\ false$ into $false$ claims and the rest to *true* claims, following previous work [15]. And for PolitiFact claim, we also $merge\ pants\ on\ fire$, $mostly\ false$, $false$ into $false$ claims and the rest to true claims. Note that we give up using the existing datasets processed by DeClarE [15] whose claims are also from the same websites, because its claims have been normalized and contain few patterns. In addition, we split subsets from the original testing sets like previous work [6]. These subsets contain claims which are incorrectly classified by a BERT-based claim-only model, denoted as Snopes Hard and PolitiFact Hard respectively. Such datasets are regarded as a more challenging evaluation to measure the reasoning ability of the model.

3.1.2 Baselines. To prove the effectiveness of our framework, we choose two base models of evidence-based fake news detection.

- BERT [7]. Similar to Brand et al. [2], they construct a BERT [4] based model to encode text, where the [CLS] token embedding represents corresponding semantic. And the attention mechanism is adopted to aggregate evidence information.
- MAC [18]. They introduce a hierarchical multi-head attention network to capture word-level and document-level interactions between news and evidence. Source credibility is also integrated. In this work, we implement a simplified version excluding publisher information.

3.1.3 Implementation Details. The maximum lengths of claims and evidences are set to both 100, and each claim is paired with 10 relevant articles in both datasets. We keep the same data splits following previous work [7]. The model is trained on five different seeds and the average experiment results are reported. The counterfactual parameter λ ranges from 0.1 to 1.0 with regard to the evaluation set. The pre-trained BERT is utilized to get the token embeddings. We reproduce all the base models via python 3.6.13 and Pytorch 1.5.1, and run the experiments on NVIDIA RTX 3090 GPUs.

3.2 Overall Performance

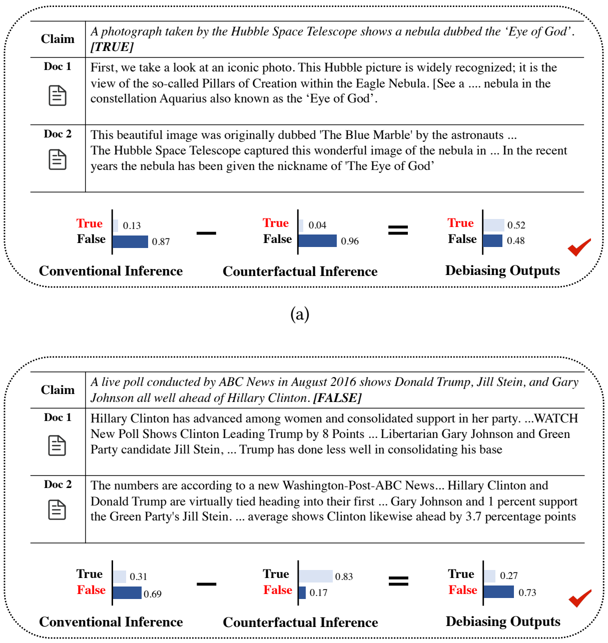
The experiment results conducted on four datasets are summarized in Table 1. By comparing the base models with and without the counterfactual framework, we can observe that our method effectively improves the performance of base models on the difficult testing sets.

Firstly, there is a significant performance decline of base models from original testing sets to hard ones. We can infer that the claim-evidence models perform better on those examples which can be easily classified to correct categories with claims only, while worse on hard examples. It is because the models tend to memorize biases of spurious correlations between news patterns and labels, instead of learning how to reason. Under the circumstance of hard examples, models can no more leverage these biases as shortcuts. Therefore, the hard testing sets pose a severe challenge to intrinsic reasoning ability.

Secondly, with the help of our counterfactual framework, base models outperform the ones without the debiasing method significantly on hard testing sets. In detail, the base models obtain an average improvement on F1 Macro up to 19.7% and 23.9% in hard scenarios respectively. It indicates that our method is effective in mitigating bias and is model-agnostic. However, it is worth noting that the framework can only maintain the performance or obtain marginal improvements on original testing sets. We attribute these different results to the bias within different testing sets. The original ones contain much bias which can be learned in training data, while the hard ones are more balanced and require complex reasoning. And our method aims to improve the robustness of models under more generalized circumstances, rather than fitting specific biases in training data.

3.3 Sensitivity Analysis

We conduct experiments with different values of counterfactual coefficient λ ranging from 0.0 to 2.0, as depicted in Figure 2. This



(b)

Figure 3: The visualization of two hard cases. Debiased outputs have been normalized. λ is set to 0.9 and omitted for brevity in the figure.

coefficient controls the proportion of counterfactual outputs to be subtracted. MAC [18] is utilized as the base model of our framework.

There is a significant fall in the original datasets when λ is increasing. It indicates that biases exist actually in these datasets and the base model tends to memorize such biases. Once the dependence on spurious correlations is reduced, the model fails to predict the veracity of news correctly. And on the Snopes Hard dataset, we can observe that the performance grows first and achieves the best when *lambda* is 0.6. This phenomenon is mainly due to different levels of bias removal. A moderate subtraction can effectively mitigate the effects of biases, while a large λ may lead to the loss of useful information. However, the performance on PolitiFact Hard dataset always increases with the growth of λ, indicating the different distributions and biases within this dataset.

3.4 Case Study

As illustrated in Figure 3, we visualize two representative instances in hard datasets. The first instance is relatively difficult since there is a long-term dependency between keywords like "Hubble Telescope", "nebula" and "Eye of God", which has been less explored by existing work [22]. Then the model can only resort to news patterns. However, our framework successfully captures the bias effects as the counterfactual prediction on the 'False' class is up to 0.96, and recovers the answer distribution by element-wise subtraction.

In the second instance, there exist some explicit contradictions between the news and the key sentence in evidences like "New Poll Shows Clinton Leading Trump by 8 Points". Then the model can easily categorize the news into the correct class based on the reasoning. And the conventional and counterfactual outputs have significantly different distributions. Our framework further consolidates the correct reasoning by enlarging the gap between them.

4 CONCLUSION

We have developed a novel counterfactual inference framework for mitigating biases in evidence-based fake news detection. Specifically, we first train the model on the original biased dataset like ordinary work. Then the bias can be distilled in the form of the direct causal effect of the news on prediction by causal intervention. In the inference stage, our framework subtracts the direct effect from the total causal effect to obtain a debiased prediction. Several experiments conducted on real-life datasets demonstrate the effectiveness and generalizability of our framework.

ACKNOWLEDGMENTS

We sincerely thank the anonymous reviewers for their insightful comments. This work is supported by National Natural Science Foundation of China (U19B2038, 62141608).

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *CSN: Politics (Topic)* (2017).
[2] Erik Brand, Kevin Roitero, Michael Soprano, and Gianluca Demartini. 2021. E-BART: Jointly Predicting and Explaining Truthfulness. *Proceedings of the 2021 Truth and Trust Online Conference* (2021), 18–27.
[3] Shantanu Chandra, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Graph-based Modeling of Online Communities for Fake News Detection. *ArXiv* abs/2008.06274 (2020).
[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*. 4171–4186.
[5] Lisa R Goldberg. 2019. The Book of Why: The New Science of Cause and Effect: by Judea Pearl and Dana Mackenzie, Basic Books (2018). ISBN: 978-0465097609.
[6] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *NAACL-HLT*. 107–112.
[7] Casper Hansen, Christian Hansen, and Lucas Chaves Lima. 2021. Automatic Fake News Detection: Are Models Learning to Reason?. In *ACL/IJCNLP*. 80–86.
[8] Qiang Liu, Feng Yu, Shu Wu, and Liang Wang. 2018. Mining significant microblogs for misinformation identification: an attention-based approach. *ACM Transactions on Intelligent Systems and Technology (TIST)* 9, 5 (2018), 1–20.
[9] Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-level evidence embedding for claim verification with hierarchical attention networks. In *ACL*. 2561–2571.
[10] Salman Bin Naeem and Rubina Bhatti. 2020. The Covid-19 'infodemic': a new front for information professionals. *Health Information and Libraries Journal* (2020).
[11] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12700–12710.
[12] Judea Pearl. 2022. Direct and indirect effects. In *Probabilistic and Causal Inference: The Works of Judea Pearl*. 373–392.
[13] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. 180–191.
[14] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility Assessment of Textual Claims on the Web. In *CIKM*. 2173–2178.
[15] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In *EMNLP*. 22–32.
[16] Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual Inference for Text Classification Debiasing. In *ACL*. 5434–5445.
[17] Qiang Sheng, Xueyao Zhang, Juan Cao, and Lei Zhong. 2021. Integrating Pattern- and Fact-based Fake News Detection via Model Preference Learning. In *CIKM*. 1640–1650.

[18] Nguyen Vo and Kyumin Lee. 2021. Hierarchical Multi-head Attentive Network for Evidence-aware Fake News Detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 965–975.

[19] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 647–653.

[20] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. 2020. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10760–10770.

[21] Lianwei Wu, Yuan Rao, Xiong Yang, Wanzhen Wang, and Ambreen Nazir. 2021. Evidence-aware hierarchical interactive attention networks for explainable claim verification. In *IJCAI*. 1388–1394.

[22] Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware Fake News Detection with Graph Neural Networks. In *WWW*.

[23] Feng Yu, Q. Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2017. A Convolutional Approach for Misinformation Identification. In *IJCAI*. 3901–3907.

[24] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining Dual Emotion for Fake News Detection *(WWW '21)*. 3465–3476.