Generalizing to the Future: Mitigating Entity Bias in Fake News Detection

Yongchun Zhu[†]
Institute of Computing Technology,
Chinese Academy of Sciences
zhuyongchun18s@ict.ac.cn

Qiang Sheng[†]
Institute of Computing Technology,
Chinese Academy of Sciences
shengqiang18z@ict.ac.cn

Juan Cao[†,*]
Institute of Computing Technology,
Chinese Academy of Sciences
caojuan@ict.ac.cn

Shuokai Li[†]
Institute of Computing Technology,
Chinese Academy of Sciences
lishuokai18z@ict.ac.cn

Danding Wang[†]
Institute of Computing Technology,
Chinese Academy of Sciences
wangdanding@ict.ac.cn

Fuzhen Zhuang[§]
Beihang University
zhuangfuzhen@buaa.edu.cn

ABSTRACT

The wide dissemination of fake news is increasingly threatening both individuals and society. Fake news detection aims to train a model on the past news and detect fake news of the future. Though great efforts have been made, existing fake news detection methods overlooked the unintended entity bias in the real-world data, which seriously influences models' generalization ability to future data. For example, 97% of news pieces in 2010-2017 containing the entity 'Donald Trump' are real in our data, but the percentage falls down to merely 33% in 2018. This would lead the model trained on the former set to hardly generalize to the latter, as it tends to predict news pieces about 'Donald Trump' as real for lower training loss. In this paper, we propose an entity debiasing framework (**ENDEF**) which generalizes fake news detection models to the future data by mitigating entity bias from a cause-effect perspective. Based on the causal graph among entities, news contents, and news veracity, we separately model the contribution of each cause (entities and contents) during training. In the inference stage, we remove the direct effect of the entities to mitigate entity bias. Extensive offline experiments on the English and Chinese datasets demonstrate that the proposed framework can largely improve the performance of base fake news detectors, and online tests verify its superiority in practice. To the best of our knowledge, this is the first work to explicitly improve the generalization ability of fake news detection models to the future data. The code is available at https://github.com/ICTMCG/ENDEF-SIGIR2022.

CCS CONCEPTS

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

KEYWORDS

Fake News Detection, Debias, Generalization

ACM Reference Format:
Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022. Generalizing to the Future: Mitigating Entity Bias in Fake News Detection. 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/10.1145/3477495.3531816

1 INTRODUCTION

In recent years, more and more people acquire news from online social media where fake news has also been widely disseminated. According to Weibo 2021 annual report on fake news refutation, 66,251 fake news pieces were detected and highlighted on Weibo.[1] The wide spread of fake news on social media has threatened both individuals and society [27, 29]. Therefore, automatic detection of fake news has been critical for promoting trust in the online news ecosystem [32].

In real-world scenarios, a fake news detector is generally trained on the existing news pieces and expected to detect fake news pieces in the future (i.e., "future data") [41]. In other words, the training and test data is unavoidably non-independent-and-identically-distributed (non-IID). However, most existing methods assume that the training and testing news pieces are sampled IID from a static news environment within the same period [3, 19, 22, 30, 35], which is unrealistic. A previous experiment [41] has showed a large performance decrease (~10%) of existing methods when changing to a more challenging temporal split from the ideal IID split.

We find that existing methods are at the risk of inadvertently capturing and even amplifying the unintended entity bias. Table 1 lists the statistics of ten typical entities in the Weibo dataset [27]. We see that the news pieces containing a certain entity have a strong correlation with the news veracity. For instance, from 2010 to 2017, 97% of news pieces containing the entity 'Donald Trump' are real.

[†] At the Key Lab of Intelligent Information Processing of Chinese Academy of Sciences. Also at the University of Chinese Academy of Sciences.
[§] At the Institute of Artificial Intelligence, and SKLSDE, School of Computer Science.
[*] Juan Cao is the corresponding author.

[1] https://weibo.com/detail/4730194303126557

SIGIR '22, July 11–15, 2022, Madrid, Spain
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531816

Table 1: Statistics of typical entities. #news indicates the number of news pieces containing the entity in the left column. %fake indicates the proportion of fake ones in all related news pieces. We see a significant difference of %fake between the 2010-2017 and 2018 subset.

	2010-2017		2018	
Entity	#news	%fake	#news	%fake
Beijing	543	51%	197	32%
Hong Kong	212	73%	59	27%
Nanjing	158	69%	51	8%
Apple	66	62%	86	74%
Samsung	54	65%	9	11%
Donald Trump	29	3%	144	67%
Jack Ma	28	57%	10	30%
McDonald	24	54%	53	100%
Huawei	21	0%	43	23%
Lionel Messi	8	0%	95	89%

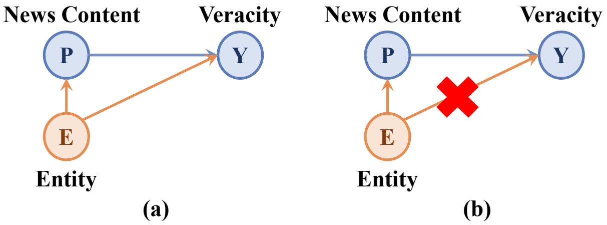

Figure 1: (a) Causal graph for existing methods, which model effects of the news content and the confounding factor (entities). (b) Our framework aims to remove the direct effect of entities.

With training using such data, models would excessively depend on the existence of certain entities for prediction. However, due to the rapid changes of the news environment [27], the correlations between a certain entity and categories vary over time. In the 2018 subset, only 33% of news pieces about ‘Donald Trump’ are real. Therefore, if a model leans to unfairly predict news pieces containing those entities to a specific veracity label according to the biased statistical information, it might hardly generalize well when applied to the future data.

In this paper, we mitigate the issue of entity bias from a cause-effect perspective. As shown in Figure 1(a), existing fake news detection methods [22, 29, 41] make predictions based on the overall contents of news pieces, which mix the direct effects of entities to the news veracity and the more generalizable non-entity signals such as writing style and emotions. With the advantage of causal learning [25, 39], we propose a novel entity debiasing framework (ENDEF) which mitigates entity bias and enhance the generalization ability of base models as shown in Figure 1(b). We specially model the direct contribution of entities to the veracity, in addition to the conventional modeling of overall contents. With explicit awareness of the entity bias during training, we remove the direct effect of related entities to perform debiased predictions. With the proposed debiasing framework, five fake news detection models in our experiment show better performance on the future data. Our contributions are as follows:

- We highlight the entity bias in fake news detection datasets and for the first time, propose to mitigate this bias for better generalization ability of fake news detectors.
- We design a debiasing framework that is convenient to be deployed along with different fake news detection models.
- We conduct both offline and online experiments to demonstrate the effectiveness of the proposed framework.

2 RELATED WORK

Fake News Detection. It aims at classifying a news piece as real or fake. Existing methods can be roughly grouped as: content-based and social-context-based fake news detection. Content-based methods mainly rely on news content features and factual resources to detect fake news [29], including text content [16, 19, 27], visual content [23, 24, 37], emotion [9, 41], and evidence bases [21, 28]. Social-context-based models exploit relevant user engagements to detect fake news [32], including propagation networks [20, 34], user profile [8, 33], and crowd feedbacks [15, 30]. Our work falls into textual content-based methods and is closely related to [37] which extracts event-invariant features and [41] which uses emotional signals for better generalization, but they do not explicitly highlight the bias from a temporal perspective. The experiments will show that our framework brings additional improvements even based on these methods that have considered the generalization issue.

Model Debiasing. The existence of the dataset bias induces models to make biased predictions, which degrades the performance on the test set [18]. Task-specific biases have been found in many areas, e.g., fact checking [26], visual question-answering [1], recommendation [2]. To mitigate such biases, some works perform data-level manipulations [7, 38], and others design model-level balancing mechanisms [10, 11]. Recently, causal learning that analyzes cause-effect relationships has been utilized for model debiasing [25, 36, 39]. Based on our analysis on task data, we propose to mitigate the entity bias for better generalization.

3 METHODOLOGY

Table 2 presents the proposed entity debiasing framework (ENDEF) for fake news detection, where we model the direct contribution of entities in addition to modeling overall contents, and then make debiased predictions by dropping the entity-based model.

3.1 Problem Formulation

Let $\mathcal{D}$ a dataset of news pieces on social media, where P is a news piece in $\mathcal{D}$ containing n tokens $P = \{w_1, \dots, w_n\}$. The tokens consist of both entities and other non-entity words. The entities in the news piece P are denoted as $E = \{e_1, \dots, e_m\}$, where m indicates the number of entities and each e represents an entity, e.g., a person, a location. To recognize the entities, we use a public tool TexSmart [14, 40][2]. Each news piece has a ground-truth label $y \in \{0, 1\}$, where 1 and 0 denote the news piece is fake and real, respectively. Given a news piece P, a fake news detection detector aims to predict whether the label of P is 1 or 0.

[2] https://ai.tencent.com/ailab/nlp/texsmart/en/inde.html. We use v0.2.0 (Large).

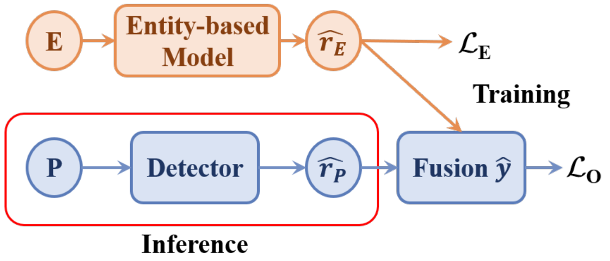

Figure 2: The proposed entity debiasing framework (ENDEF) consists of an entity-based model and a detector. The entity-based model aims to capture the entity bias, which enables the detector to learn less biased information.

3.2 Model

We propose a debiasing framework for fake news detection based on the causal graph in Figure 1(b), which improves the generalization ability of base detectors by mitigating entity bias. Our framework models two cause-effect paths $E \longrightarrow Y$ and $E \longrightarrow P \longrightarrow Y$ (Figure 2).

The distributions of entity lists E may bring spurious correlations (e.g., the spurious “Lionel Messi”-to-“real class” mapping) during model training. To explicitly model the potentially negative influence of entity bias, we train an entity-based model *only* with E as input to represent the direct effect from the entities $E \longrightarrow Y$:

$$\hat{r}_E = f_E(\{e_1, \dots, e_m\}), \tag{1}$$

where f_E is a deep network. $\hat{r}_E$ is an entity-biased logit prediction.

Generally, a fake news detector takes all tokens as input, including both entities and non-entity words:

$$\hat{r}_P = f_P(\{w_1, \dots, w_n\}), \tag{2}$$

where f_P is a fake news detector. Note that our framework is model-agnostic, and f_P can be implemented with diverse models for this task [19, 30, 37, 41]. $\hat{r}_P$ indicates its logit prediction, which represents the effect of the path $E \longrightarrow P \longrightarrow Y$.

The final probability prediction is aggregated with the two parts $\hat{r}_E$ and $\hat{r}_P$, formulated as:

$$\hat{y} = \sigma(\alpha \hat{r}_P + (1 - \alpha)\hat{r}_E), \tag{3}$$

where the $\sigma(\cdot)$ indicates the Sigmoid function and α is a hyper-parameter to balance the two terms. We train the overall framework with the cross-entropy loss:

$$\mathcal{L}_O = \sum_{(P,y)\in\mathcal{D}} -y\log(\hat{y}) - (1-y)\log(1-\hat{y}). \tag{4}$$

To achieve the effect of the entity module, we utilize an auxiliary loss, which applies additional supervision over the prediction of the entity-based model:

$$\mathcal{L}_E = \sum_{(P,y)\in\mathcal{D}} -y\log(\sigma(\hat{r}_E)) - (1-y)\log(1-\sigma(\hat{r}_E)). \tag{5}$$

The overall loss function can be formulated as:

$$\mathcal{L} = \mathcal{L}_O + \beta\mathcal{L}_E, \tag{6}$$

where the β denotes a hyper-parameter, which is set as 0.2 in this paper. This training procedure can make the entity-based model focus on learning the entity bias. Meanwhile, it enable the detector to learn less biased information.

Table 2: Statistics of the datasets.

Dataset	Weibo			GossipCop		
	Train	Val	Test	Train	Val	Test
#Fake	2,561	499	754	2,024	604	601
#Real	7,660	1,918	2,957	5,039	1,774	1,758
Total	10,221	2,417	3,711	7,063	2,378	2,359

This training procedure forces the entity-based model to focus on learning to detect fake news with *only* the entities provided and thus fit the entity bias in the training set well. Meanwhile, it enables the fake news detector to learn less biased information by encouraging the two modules to capture different signals (entity-based and non-entity-based, respectively).

3.3 Inference

To mitigate entity bias for better generalization ability, the key is to remove the direct effect via path $E \longrightarrow Y$ from the prediction $\hat{y}$. Since $\hat{y}$ can be seen as the total effect and $\hat{r}_E$ is the natural direct effect of entities to the news veracity label, the remaining $\hat{r}_P$ is actually a prediction based on less biased information. Therefore, we could mitigate the entity bias by simply using $\sigma(\hat{r}_P)$ during inference. Note that the detector in Figure 2 is not limited to specific models, making our debiasing framework be compatible with diverse base models to improve their generalization ability.

3.4 Data Augmentation

Data augmentation technique has shown its power for alleviating overfitting [13]. To further improve the generalization ability of the models, we adopt two types of token-level augmentation techniques, including drop (deleting the selected token) and mask (replacing the selected token with a special token `[MASK]`). In addition, we apply two augmentation policies: (1) randomly drop or mask words with the probability of p; and (2) randomly drop or mask entities with the probability of p. In the training stage, we randomly adopt one augmentation policy for each sample.

4 EXPERIMENTS

We experimentally answer the following research questions:

- **RQ1** Can our framework improve the generalization ability of fake news detection on future data?
- **RQ2** Can this debiasing framework bring improvement to the performance of the real-world online system?
- **RQ3** How does the mitigation of entity bias improve the performance of base models?

4.1 Experimental Settings

Datasets. A Chinese dataset and an English dataset are adopted for evaluation. For the Chinese dataset, we adopt the Weibo dataset [27] from 2010 to 2018.[3] For the English dataset, we adopt the GossipCop

[3] https://github.com/ICTMCG/News-Environment-Perception/

Table 3: Offline performance comparison of base models with and without the debiasing framework. The better result in each group using the same base model are in boldface. The marker * indicates that the improvement is statistically significant compared with the best baseline (paired t-test with p-value < 0.05).

Method	Weibo						GossipCop					
	macF1	Acc	AUC	spAUC	$F1_{real}$	$F1_{fake}$	macF1	Acc	AUC	spAUC	$F1_{real}$	$F1_{fake}$
BiGRU	0.7172	0.8214	0.8354	0.6636	0.8887	0.5456	0.7730	0.8379	0.8634	0.7358	0.8943	0.6516
w/ ENDEF	**0.7318***	**0.8286***	**0.8446***	**0.6802***	**0.8929***	**0.5707***	**0.7842***	**0.8465***	**0.8669**	**0.7472***	**0.8989***	**0.6696***
EANN	0.7162	0.8197	0.8276	0.6649	0.8875	0.5448	0.7926	0.8517	0.8765	0.7586	0.9033	0.6820
w/ ENDEF	**0.7370***	**0.8316***	**0.8398***	**0.6886***	**0.8947***	**0.5793***	**0.7937**	**0.8526**	**0.8836***	**0.7620***	**0.9039**	**0.6835**
BERT	0.7601	0.8474	0.8754	0.7102	0.9048	0.6155	0.7873	0.8439	0.8781	0.7579	0.8968	0.6778
w/ ENDEF	**0.7714***	**0.8550***	**0.8824***	**0.7257***	**0.9096***	**0.6332***	**0.7969***	**0.8496***	**0.8853***	**0.7663***	**0.8994**	**0.6944***
MDFEND	0.7051	0.7786	0.8301	0.6691	0.8519	0.5584	0.7905	**0.8518**	0.8712	0.7543	**0.9037**	0.6772
w/ ENDEF	**0.7313***	**0.8057***	**0.8490***	**0.6879***	**0.8724***	**0.5902***	**0.7970***	0.8517	**0.8824***	**0.7627***	0.9023	**0.6916***
BERT-Emo	0.7586	0.8438	0.8743	0.7061	0.9019	0.6154	0.7912	0.8455	0.8800	0.7631	0.8974	0.6849
w/ ENDEF	**0.7731***	**0.8584***	**0.8838***	**0.7278***	**0.9121***	**0.6341***	**0.8010***	**0.8520***	**0.8855***	**0.7674***	**0.9020***	**0.6987***

Table 4: Results on the online data. Each row indicates the relative improvement with our ENDEF framework over the baselines.

Method	macF1	Acc	AUC	spAUC	$F1_{real}$	$F1_{fake}$
BiGRU	2.56%	2.02%	4.02%	3.26%	1.12%	6.45%
EANN	0.57%	-0.77%	1.64%	1.02%	-0.44%	3.26%
BERT	0.60%	0.57%	0.44%	0.32%	0.33%	1.12%
MDFEND	2.57%	2.02%	1.14%	0.95%	1.19%	5.50%
BERT-Emo	0.68%	0.78%	0.22%	0.78%	0.13%	2.44%

data of FakeNewsNet [31].[4] To simulate the real-world temporal scenarios, we adopt the *temporal* split strategy. The most recent 40% news pieces are randomly included in the test and validation sets. The remaining 60% of news pieces serve as the training set. Specifically, for the Weibo dataset, the time period of the training set is from 2010 to 2017, and the samples in the test and validation sets are posted in 2018. For the GossipCop dataset, the time period of the training set is from 2000 to 2017, and all samples in the test and validation sets are posted in 2018. Table 2 shows the statistics of the datasets.

Base Models. Technically, our framework is model-agnostic, which could coordinate with various fake news detectors. Here we select five representative content-based detection methods as our base models.

- BiGRU [4] is widely used in many existing works of our task for text encoding [19, 41]. We implement a one-layer BiGRU with a hidden size of 768. Then, we utilize a mask attention layer to aggregate all the hidden states as representations of posts which are further fed into an MLP for prediction.
- EANN [37] is a model that tries to distract the fake news detection model from memorizing event-specific features. It uses TextCNN for text representation and adds an auxiliary task of event classification for adversarial learning using gradient reversal layer. The complete EANN is a multi-modal model but we here use its text-only version. For TextCNN, the window sizes are {1, 2, 3, 5, 10}. The labels for the auxiliary event classification task are derived by clustering according to the publication year.
- BERT [5, 6] is a popular pre-training model. We utilize BERT to encode tokens of news content and feed the extracted average embedding into an MLP to obtain the final prediction. For the GossipCop dataset, we adopt the original BERT model [6]. For the Weibo dataset, we adopted a modified BERT model [5].
- MDFEND [19] is the latest multi-domain text-based fake new detection model which utilizes a Domain Gate to select useful experts of MoE. We adopt the same TextCNN structure for all experts. The number of experts is set as 5. In this paper, we utilize the publication year as the domain label.
- BERT-Emo [41] combines emotional features and the BERT detector for fake news detection. As we focus on the contents rather than social contexts, we adopt a simplified version where emotions in comments are not considered.

Evaluation Metrics. Following most existing works [30, 37, 41], we report Area Under ROC (AUC), accuracy (Acc), macro F1 score (macF1) and the F1 scores of fake and real class ($F1_{fake}$ and $F1_{real}$). In addition, as the datasets are skewed (real: fake ≈ 3: 1), a fake news detector should detect fake news without misclassifying real news as possible. Formally speaking, we should improve the true positive rate (TPR) on the basis of low false positive rate (FPR). Therefore, following [17, 27, 42], we bring a metric into the evaluation of fake news detectors named standardized partial AUC ($spAUC_{FPR \leq maxfpr}$). In practice, we require FPR to be less than 10%. Hence, in this paper, we use $spAUC_{FPR \leq 10\%}$ for all experiments.

Implementation Details. We did not perform any dataset-specific tuning except early stopping on validation sets. For all methods, the initial learning rate for the Adam optimizer [12] was tuned by grid search in [1e-6, 1e-2]. We performed a grid search in [0,1] with the step of 0.1 for searching the best hyper-parameter α and we found 0.8 was the best value. The mini-batch size is 64. For the MLPs in these methods, we employed the ReLU function and set the dimension of the hidden layer as 384. The maximum

[4] https://github.com/KaiDMML/FakeNewsNet

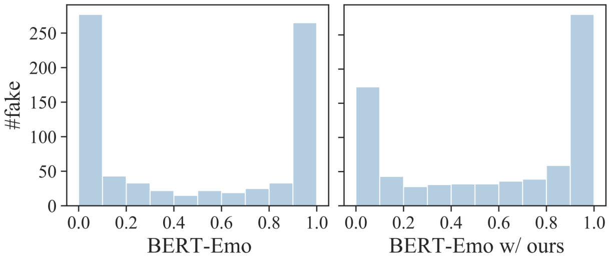

Figure 3: The distribution of prediction scores on fake news pieces in the Weibo test set. The average prediction scores of BERT-Emo and BERT-Emo w/ ours are 0.49 and 0.58, respectively.

sequence lengths of GossipCop and Weibo datasets were set as 170. We set the probability of applying data augmentation p as 0.1. We ran all methods on both two datasets for ten times and report the average scores for all metrics.

4.2 Results

4.2.1 Offline Experiments (RQ1). We conduct offline experiments with two datasets. Specifically, we apply the proposed framework upon various base models. The performance of these base models with and without the framework is shown in Table 3. From these results, we have the following observations:

- With the help of the proposed framework, most base models show a significant improvement in most metrics, which demonstrates the effectiveness of the proposed framework on future data. In addition, it also testifies ENDEF is a general framework which can be applied upon various base models.
- For most metrics, the performance improvement in the Weibo dataset is larger than that in the GossipCop dataset. We attribute such a difference between the two datasets to the length of a news piece. The average length of news pieces in the Weibo dataset is 120, while the average length of the GossipCop dataset is 606. The longer news piece would have more informative patterns, e.g., writing style, emotion, which alleviating the influence of entities.

4.2.2 Online Experiments (RQ2). We tested these models on a dump of ten-month data in 2021 from our Chinese fake news detection system. Different from the offline datasets, this online data set is much more skewed (30,977 real: 774 fake ≈ 40:1). Due to the restriction of business rules, we cannot report the absolute performance of these methods. Instead, we report the relative improvement of our proposed framework compared with different base models in Table 4, and each row indicates the relative improvement of the base model with the proposed framework over the base model. The online results show the proposed debiasing framework can improve the base models in a highly skewed scenario, which demonstrates the importance of alleviating the influence of entity bias in real-world detection systems. The best-performing model *BERT-Emo w/ ours* has been deployed in our online system which handles thousands of suspicious news pieces every day.

4.2.3 Analysis (RQ3). Figure 3 presents the distribution of prediction scores of BERT-Emo and BERT-Emo w/ ours on fake news in the Weibo test set, which demonstrates the base model with our framework can detect fake news more accurately (higher average scores of fake news). In addition, we show two fake news cases in Figure 4. *Debiased* indicates the final debiased prediction $\sigma(\hat{r}_P)$, and *entity-based* represents the prediction of the entity-based model. The two cases show that the entity-based model can capture the biased information, which enables the detector trained with our framework to make debiased predictions.

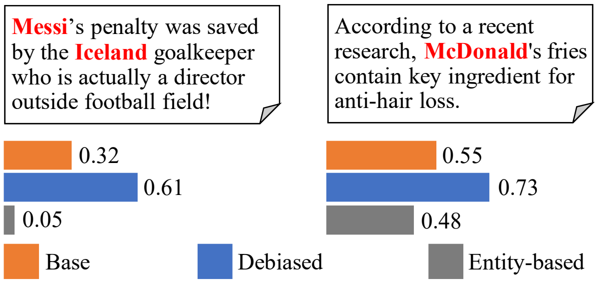

Figure 4: Two fake news cases. The entities are boldfaced. The lengths of bars represent the probability predictions of the base model (BERT-Emo), our debiased model, and the entity-based model. Our debiased model shows higher confidence in predicting the two samples as fake.

5 CONCLUSION AND FUTURE WORK

We proposed a novel entity debiasing framework (ENDEF) which improves the generalization ability of base fake news detectors on future data by mitigating the largely overlooked entity bias in existing works. Specifically, we designed a debiasing framework based on a causal graph of entities, news contents, and news veracity, where the direct effect of entities to the news veracity is explicitly modeled during training the base fake news detector and finally removed during inference to perform debiased prediction. Both offline and online experiments demonstrated the effectiveness of our proposed framework.

To the best of our knowledge, this is the first work to explicitly focus on improving the generalization ability of fake news detection models to the future data, which is a practical problem in the real-world detection system. We believe further exploration is required for a deeper understanding of the gap between news pieces in different periods. In the future, we plan to explore: (1) adapting unbiased model to the future news environment; (2) investigating the common features between news pieces of different periods; (3) extending the proposed framework ENDEF from the text-only detection to multi-modal and social graph-based detection.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their insightful and valuable comments. The research work is supported by the National Key Research and Development Program of China (2021AAA0140203), the Zhejiang Provincial Key Research and Development Program of China (2021C01164), and the National Natural Science Foundation of China (62176014).

REFERENCES

[1] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the Behavior of Visual Question Answering Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. ACL, 1955–1960.

[2] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2020. Bias and Debias in Recommender System: A Survey and Future Directions. *arXiv preprint arXiv:2010.03240* (2020).

[3] Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. 2021. Causal Understanding of Fake News Dissemination on Social Media. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 148–157.

[4] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. 103–111.

[5] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with Whole Word Masking for Chinese BERT. *arXiv preprint arXiv:1906.08101* (2019).

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, 4171–4186.

[7] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 67–73.

[8] Yingtong Dou, Kai Shu, Congying Xia, Philip S. Yu, and Lichao Sun. 2021. User Preference-Aware Fake News Detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2051–2055.

[9] Anastasia Giachanou, Paolo Rosso, and Fabio Crestani. 2019. Leveraging Emotional Signals for Credibility Detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 877–880.

[10] Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving Debiasing for Pre-trained Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL, 1641–1650.

[11] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2019. Decoupling Representation and Classifier for Long-Tailed Recognition. In *International Conference on Learning Representations*.

[12] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of International Conference on Learning Representations*, Vol. 5.

[13] Shuokai Li, Xiang Ao, Feiyang Pan, and Qing He. 2022. Learning Policy Scheduling for Text Augmentation. *Neural Networks* 145 (2022), 121–127.

[14] Lemao Liu, Haisong Zhang, Haiyun Jiang, Yangming Li, Enbo Zhao, Kun Xu, Linfeng Song, Suncong Zheng, Botong Zhou, Dick Zhu, et al. 2021. TexSmart: A System for Enhanced Natural Language Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. ACL, 1–10.

[15] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Detect rumor and stance jointly by neural multi-task learning. In *Companion Proceedings of the Web Conference 2018*. ACM, 585–593.

[16] Jing Ma, Wei Gao, and Kam-Fai Wong. 2019. Detect Rumors on Twitter by Promoting Information Campaigns with Generative Adversarial Learning. In *Proceedings of the Web Conference*. 3049–3055.

[17] Donna Katzman McClish. 1989. Analyzing a Portion of the ROC Curve. *Medical Decision Making* 9, 3 (1989), 190–195.

[18] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[19] Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. MDFEND: Multi-domain Fake News Detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3343–3347.

[20] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. FANG: Leveraging Social Context for Fake News Detection Using Graph Representation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. ACM, 1165–1174.

[21] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. ACL, 22–32.

[22] Piotr Przybyla. 2020. Capturing the Style of Fake News. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 490–497.

[23] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving Fake News Detection by Using an Entity-enhanced Framework to Fuse Diverse Multimodal Clues. In *Proceedings of the 29th ACM International Conference on Multimedia*. ACM, 1212–1220.

[24] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE International Conference on Data Mining*. IEEE, 518–527.

[25] Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual Inference for Text Classification Debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5434–5445.

[26] Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards Debiasing Fact Verification Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. ACL, 3419–3425.

[27] Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom Out and Observe: News Environment Perception for Fake News Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. ACL.

[28] Qiang Sheng, Juan Cao, Xueyao Zhang, Xirong Li, and Lei Zhong. 2021. Article Reranking by Memory-Enhanced Key Sentence Matching for Detecting Previously Fact-Checked Claims. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. ACL, 5468–5481.

[29] Qiang Sheng, Xueyao Zhang, Juan Cao, and Lei Zhong. 2021. Integrating Pattern- and Fact-based Fake News Detection via Model Preference Learning. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1640–1650.

[30] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. dEFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 395–405.

[31] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big data* 8, 3 (2020), 171–188.

[32] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

[33] Kai Shu, Suhang Wang, and Huan Liu. 2018. Understanding User Profiles on Social Media for Fake News Detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval*. IEEE, 430–435.

[34] Amila Silva, Yi Han, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. Propagation2Vec: Embedding Partial Propagation Networks for Explainable Fake News Early Detection. *Information Processing & Management* 58, 5 (2021), 102618.

[35] Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. Embracing Domain Differences in Fake News: Cross-domain Fake News Detection using Multi-modal Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 557–565.

[36] Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Clicks Can Be Cheating: Counterfactual Recommendation for Mitigating Clickbait Issue. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1288–1297.

[37] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 849–857.

[38] Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. ACL, 6382–6388.

[39] Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. 2021. Model-Agnostic Counterfactual Reasoning for Eliminating Popularity Bias in Recommender System. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1791–1800.

[40] Haisong Zhang, Lemao Liu, Haiyun Jiang, Yangming Li, Enbo Zhao, Kun Xu, Linfeng Song, Suncong Zheng, Botong Zhou, Jianchen Zhu, et al. 2020. TexSmart: A Text Understanding System for Fine-Grained NER and Enhanced Semantic Analysis. *arXiv preprint arXiv:2012.15639* (2020).

[41] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining Dual Emotion for Fake News Detection. In *Proceedings of the Web Conference 2021*. 3465–3476.

[42] Yongchun Zhu, Dongbo Xi, Bowen Song, Fuzhen Zhuang, Shuai Chen, Xi Gu, and Qing He. 2020. Modeling Users' Behavior Sequences with Hierarchical Explainable Network for Cross-domain Fraud Detection. In *Proceedings of The Web Conference 2020*. ACM, 928–938.