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 address the causal graph among entities, news contents, and news veracity, and this work 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.

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. 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.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

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

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

CSC CONCEPTS
• Information systems → Data mining. • Computing methodologies → Natural language processing.

KEYWORDS
Fake News Detection, Debias, Generalization
With training using such data, models would excessively depend when applied to the future data.

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.

| Entity           | 2010-2017 | 2018  |
|------------------|-----------|-------|
|                  | #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 $P = \{w_1, \ldots, w_n\}$. The tokens consist of both entities and other non-entity words. The entities in the news piece $P$ are denoted as $E = \{e_1, \ldots, e_m\}$, 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.

2https://ai.tencent.com/ailab/tdp/texsmart/en/inde.html. We use v0.2.0 (Large).
We propose a debiasing framework for fake news detection based on where the $\beta$ with the cross-entropy loss:

$$L_E = \frac{1}{|D|} \sum_{(p, y) \in D} -y \log(\tilde{y}) - (1 - y) \log(1 - \tilde{y}).$$

(5)

The overall loss function can be formulated as:

$$L = L_O + \beta L_E,$$

(6)

where the $\beta$ 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.

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 $E \rightarrow Y$ from the prediction $\hat{y}$. Since $\hat{y}$ can be seen as the total effect and $L_E$ is the natural direct effect of entities to the news veracity label, the remaining $E_P$ is actually a prediction based on less biased information. Therefore, we could mitigate the entity bias by simply using $\sigma(E_P)$ 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. For the English dataset, we adopt the GossipCop

### 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 |

3https://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 | 0.6115 | 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]. 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 news 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 (F1fake and F1real). In addition, as the datasets are skewed (real: fake = 3:1), a fake news detector should detect fake news without misclassifying real news as fake. 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). In practice, we require FPR to be less than 10%. Hence, in this paper, we use spAUC_FPR≤0.10 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 set the dimension of the hidden layer as 384. The maximum 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.

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 (F1fake and F1real). In addition, as the datasets are skewed (real: fake = 3:1), a fake news detector should detect fake news without misclassifying real news as fake. 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). In practice, we require FPR to be less than 10%. Hence, in this paper, we use spAUC_FPR≤0.10 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
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 \( \approx 40:1 \)). 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 \( \hat{\sigma}(r_P) \), 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.

### 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).

---

**Table 3: Comparison of prediction scores with and without the proposed framework.**

| Base Model         | Before Framework | After Framework | Improvement |
|--------------------|------------------|-----------------|-------------|
| BERT-Emo           | 0.32             | 0.48            | 0.16        |
| BERT-Emo \( w/ \) ours | 0.55             | 0.61            | 0.06        |

**Table 4: Comparison of relative improvement.**

| Base Model         | Relative Improvement |
|--------------------|----------------------|
| BERT-Emo           | 0.16                 |
| BERT-Emo \( w/ \) ours | 0.06                 |

**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.

**Figure 4: Two fake news cases.** 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.
