University Library

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Gong, S;Sinnott, R;Qi, J;Paris, C

Title:
Unseen Fake News Detection Through Casual Debiasing

Date:
2025-05-23

Citation:
Gong, S., Sinnott, R., Qi, J. & Paris, C. (2025). Unseen Fake News Detection Through Casual Debiasing. WWW '25: Companion Proceedings of the ACM on Web Conference 2025, pp.981-985. Association for Computing Machinery. https://doi.org/10.1145/3701716.3715517.

Persistent Link:
https://hdl.handle.net/11343/361932

License:
CC-BY

Unseen Fake News Detection Through Casual Debiasing

Shuzhi Gong
The University of Melbourne
Melbourne, VIC, Australia
shuzhig@student.unimelb.edu.au

Richard Sinnott
The University of Melbourne
Melbourne, VIC, Australia
rsinnott@unimelb.edu.au

Jianzhong Qi
The University of Melbourne
Melbourne, VIC, Australia
jianzhong.qi@unimelb.edu.au

Cecile Paris
Data61, CSIRO
Sydney, NSW, Australia
Cecile.Paris@data61.csiro.au

Abstract

The widespread dissemination of fake news on social media poses significant risks, necessitating timely and accurate detection. However, existing methods struggle with unseen news due to their reliance on training data from past events and domains, leaving the challenge of detecting novel fake news largely unresolved.

To address this, we identify biases in training data tied to specific domains and propose a debiasing solution FNDCD. Originating from causal analysis, FNDCD employs a reweighting strategy based on classification confidence and propagation structure regularization to reduce the influence of domain-specific biases, enhancing the detection of unseen fake news. Experiments on real-world datasets with non-overlapping news domains demonstrate FNDCD's effectiveness in improving generalization across domains.

CCS Concepts

• **Security and privacy** → **Social aspects of security and privacy**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

Fake News Detection, Graph Neural Network, Generalization

ACM Reference Format:
Shuzhi Gong, Richard Sinnott, Jianzhong Qi, and Cecile Paris. 2025. Unseen Fake News Detection Through Casual Debiasing. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3701716.3715517

1 Introduction

The proliferation of social media has accelerated the spread of both accurate and misleading information. Early and reliable detection of fake news is thus crucial to minimizing its harmful societal impact. Recent advances in fake news detection have utilized graph-based techniques, especially Graph Neural Networks (GNNs), to model news propagation patterns and extract critical insights that might be used for identifying misinformation [3]. However, these approaches typically assume that both training and testing data share the same underlying distribution (the "i.i.d." assumption). The trained models often embed biases and noise present in the training data, leading to mis-classifications during model inference [10]. This means that the model has not been trained on representative data of the fake news to be detected. In reality, fake news has often never been seen before and originates from new domains, which poses significant challenges in generalizing models trained on known distributions (so called *in-distribution*) to novel and unseen distributions (so called *out-of-distribution*, OOD).

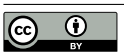
Some studies have tried to identify content-independent propagation patterns to detect fake news across different news domains [1, 4, 14], however, more recent work suggests that both content and propagation structures may differ significantly between news domains [17]. When considering the training and testing data from a source domain and target domain, domain adaptation approaches are often used [7, 9]. They attempt to address this issue by fine-tuning models using a small amount of labeled target domain data. However, such labeled data is not always available in real-world scenarios where the fake news has not been seen before, e.g. the breaking COVID-19 event. Furthermore, target domains can evolve quickly and there may not be enough time to label such new data before fake news spreads. Real fake news detection can thus be regarded as an out-of-distribution generalization task.

To address these challenges, we propose **F**ake **N**ews **D**etection by **C**ausal **D**ebiasing (FNDCD), a novel approach designed for zero-shot unseen domain fake news detection. Through causal analysis (see next section), we identify that the existence of biased training samples restricts the trained models' generalization performance on unseen data from new domains. A self-supervised weighting strategy is designed according to the news content, news propagation pattern, and existing labels. Through re-weighting, the contribution from the biased data during model training can be reduced to improve the trained model's cross-domain generalization and overall ability to tackle nascent fake news challenges.

Extensive experiments are conducted with four datasets used for unseen fake news detection. We show that FNDCD outperforms state-of-the-art models. FNDCD also provides interpretability and insights about the data and its potential for bias, allowing improvements for future news data collection and processing.

2 Problem Formulation

Unseen fake news detection. We consider graph-based fake news detection using propagation graphs (trees), comprising source news posts, comments and reposts. Given a training dataset D^{tr} =



This work is licensed under a Creative Commons Attribution 4.0 International License.
WWW Companion '25, Sydney, NSW, Australia
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1331-6/25/04
https://doi.org/10.1145/3701716.3715517

$\{\mathcal{G}_k^{tr}, y_k^{tr}\}_{k=1}^{\mathcal{N}_{tr}}$, where $\mathcal{G}_k^{tr}$ is the k-th training propagation graph, y_k^{tr} is its label, and $\mathcal{N}^{tr}$ is the number of training samples, the aim is to train a model using D^{tr} for optimal performance on unseen testing data $D^{te} = \{\mathcal{G}_k^{te}, y_k^{te}\}_{k=1}^{\mathcal{N}^{te}}$. Distribution shifts typically exist between D^{tr} and D^{te} when they are collected from different news domains at different times. In the unseen fake news detection setting, the features and labels of D^{te} are unavailable during training.

For any data sample $\mathcal{G} = \langle \mathbf{X}, \mathbf{A} \rangle$ from D^{tr} or D^{te}, the propagation graph $\mathcal{G}$ is composed of news posts, comments and reposts represented by node features $\mathbf{X}$, and user interactions (comments and reposts) represented by graph edges $\mathbf{A}$. To convert the raw news text into graph features, we use a pre-trained RoBERTa model [11].

Distribution shifts in fake news detection.

It has been established that distribution shifts often exist between training data D^{tr} and testing data D^{te} [17, 22]. In the context of misinformation and its propagation, these shifts can be characterised from three perspectives: shifts in textual content $\mathbf{X}$, shifts in propagation structure $\mathbf{A}$, and shifts in the label-feature correlations $p(y|\mathbf{X}, \mathbf{A})$. The shifts in textual content occur when the news and associated comments pertain to different news topics potentially across different news domains. For example, political news often involves vocabulary related to countries and politicians, whereas COVID-19 posts focus more on medical information. The shifts in propagation structure reflect the variation in propagation graphs between different news domains. Shifts in label-feature correlations arise when similar embeddings from different domains that might be extracted by traditional graph encoders, give rise to contrasting labels. This correlation shift presents a significant challenge for fake news detection for unseen news domains.

To address these shifts, a causal analysis is considered.

Causal analysis. We hypothesize that generalization to unseen news domains is hindered by biases in the training set, such as biases toward specific events like the US presidential election (*environment-bias*). To model this, we use a structural causal model (SCM) [21], shown in Fig.1, where nodes represent variables and edges represent causal effects. The observed propagation graph G is generated from two latent variables: the causal variable C and the environment-biased variable E. The label y is predicted based on G, influenced by both C and E.

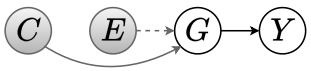


Figure 1: Structure of the causal model used for training cross-domain fake news detection. C: Causal information that supports the correct classification; E: Spurious environment-biased information harming the classification; G: Observed graph features; Y: Associated veracity label. The grey and white variables represent the degree of observability (unobserved is grey and observed is white).

Training on environment-biased samples embeds spurious correlations ($E \rightarrow G \rightarrow Y$), leading to sub-optimal performance on OOD data. These spurious correlations interfere with the causal relationship ($C \rightarrow G \rightarrow Y$). To address this, environment-biased samples must be identified and down-weighted, ensuring that only causal effects are preserved for accurate OOD classification.

Rescue of probability. Inspired by a recent work [8], the news data generation can be described by the joint probability of several variables: the textual content $\mathbf{X}$; the structure $\mathbf{A}$ of the propagation graph $\mathcal{G}$, the veracity label $\mathbf{y}$ and the environment variable $\mathbf{e}$.

Here, the environment variable $\mathbf{e}$ is treated as independent because the other variables originate from it (i.e., the variety of domains/topics causes the distribution shifts). The news content $\mathbf{X}$ should depend on $\mathbf{e}$, yet, for simplicity, we use a domain-adaptive pre-trained language model (DA-PLM) to extract the news content features such that $\mathbf{X}$ is disentangled from $\mathbf{e}$. Considering the homophily principle theory [20] linking the probability depending on some inherent similarity between nodes, we assume that users sharing similar interests are more likely to interact. Variable $\mathbf{A}$ is hence defined as depending on $\mathbf{X}$ and $\mathbf{e}$. Finally, the news veracity label $\mathbf{y}$ is generated from both the graph $\mathcal{G} = \langle \mathbf{X}, \mathbf{A} \rangle$ and the environment $\mathbf{e}$. According to the dependence between variables $\mathbf{X}$, $\mathbf{A}$, $\mathbf{y}$, and $\mathbf{e}$, the joint probability $p(\mathbf{X}, \mathbf{A}, \mathbf{y}, \mathbf{e})$ can be given as:

$$p(\mathbf{X}, \mathbf{y}, \mathbf{A}, \mathbf{e}) = p(\mathbf{e})p(\mathbf{X})p(\mathbf{A}|\mathbf{X}, \mathbf{e})p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{e}), \tag{1}$$

where the generative models $p(\mathbf{A}|\mathbf{X}, \mathbf{e})$ and $p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{e})$ can be instantiated by flexible parametric distributions $p_\theta(\mathbf{A}|\mathbf{X}, \mathbf{e})$ and $p_\theta(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{e})$ with parameter θ. Most existing works aim to maximize the likelihood $\mathcal{P}_\theta(y|X, A)$, which is unsuitable for OOD prediction where X and A are causally affected by environment biases.

We propose to filter the environment-biased information through data debiasing as follows. An environment variable $\mathbf{e}$ is defined as 1 or 0 for every training sample, indicating whether it is environment-independent ($\mathbf{e} = 1$) or environment-biased ($\mathbf{e} = 0$). Our training object is to optimize $\mathcal{P}_\theta(y|X_{e=1}, A_{e=1})$, to focus on environment-independent samples.

To infer variable $\mathbf{e}$, posterior probability is utilized as follows:

$$\begin{aligned} p_\theta(\mathbf{e}|\mathbf{A}, \mathbf{X}, \mathbf{y}) &= \frac{p_\theta(\mathbf{X}, \mathbf{y}, \mathbf{A}, \mathbf{e})}{\sum_{\mathbf{e'} \in \{0,1\}} p_\theta(\mathbf{X}, \mathbf{y}, \mathbf{A}, \mathbf{e'})} \\ &= \frac{p(\mathbf{e})p(\mathbf{X})p_\theta(\mathbf{A}|\mathbf{X}, \mathbf{e})p_\theta(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{e})}{\sum_{\mathbf{e'} \in \{0,1\}} p(\mathbf{e'})p(\mathbf{X})p_\theta(\mathbf{A}|\mathbf{X}, \mathbf{e'})p_\theta(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{e'})}, \end{aligned} \tag{2}$$

where $p(\mathbf{e})$ is the prior probability, which is set as a hyperparameter in the experiments; $p_\theta(\mathbf{A}|\mathbf{X}, \mathbf{e})$ is instantiated as the structure estimator (predicting edge connection $\mathbf{A}$ based on features $\mathbf{X}$ and environment $\mathbf{e}$); and $p_\theta(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{e})$ is instantiated as the classification module to predict the news veracity label, as detailed next.

3 Methodology

Model overview. The model structure (training phase) is shown in Fig.2. Raw text data (news content and posts) are encoded using RoBERTa[11]. The resulting propagation graph, with node feature embeddings, is input into a Classification Model and a Structure Estimator to generate the label prediction $p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{e})$ and connection likelihood $p(\mathbf{A}|\mathbf{X}, \mathbf{e})$. These outputs are used in posterior inference to estimate the environment variable $\mathbf{e}$, which weights the loss during training. The model is optimized via Expectation Maximization [18]. During testing, $\mathbf{e}$ is set to 1 for all samples, and $p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{e} = 1)$ provides the final prediction.

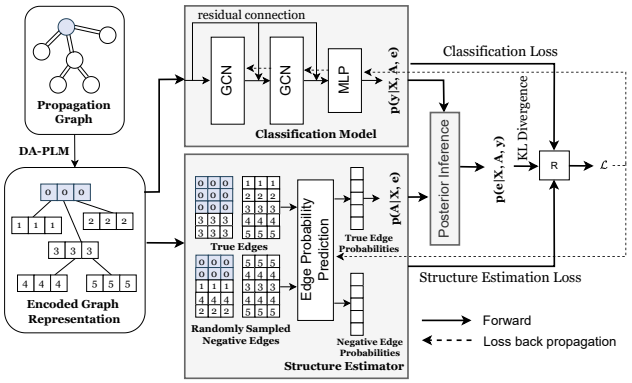


Figure 2: The structure of FNDCD. R is the loss reweight module according to the inferred environment variable e.

Classification model. To instantiate the distribution $p_\theta(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{e})$ we follow existing graph-based fake news detection models [1], combining a two-layer Graph Convolutional Network (GCN) and a multi-layer perception (MLP). Given a graph's node features $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and its adjacency matrix $\mathbf{A}$, the propagation graph's embeddings are computed through GCNs with residual connections:

$$\mathcal{Z}^{(l+1)} = \sigma\left(\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\mathcal{Z}^{(l)}\mathbf{W}^{(l)}\right) + \mathcal{Z}^{(l)}, \tag{3}$$

where $l = 0$ or 1; $\mathcal{Z}^{(0)}$ is the initial node features $\mathbf{X}$; $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the adjacent matrix of the graph with self-loops; $\mathbf{I}$ is the identity matrix; $\tilde{\mathbf{D}}$ is the degree matrix of $\tilde{\mathbf{A}}$; $\mathbf{W}^{(l)}$ is the learnable parameter matrix; and σ is the activation function. After two layers of GCNs, the outputs $\mathcal{Z}^{(2)}$ is fed into the MLP to generate the prediction $\hat{y}$, where:

$$\hat{y} = \text{softmax}(\text{MLP}(\mathcal{Z}^{(2)})) \tag{4}$$

According to Equation 2, there are two probabilities for the classification model $p_\theta(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{e} = 1)$ and $p_\theta(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{e} = 0)$ representing the classification of the environment-independent and environment-biased samples, respectively. When the data is biased to environment, the correlation between the features and the veracity labels will be agnostic since the labels are more likely to be correlated to the environment biases (e.g., specific news events). Therefore, the above classification model only describes the probability $p_\theta(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{e} = 1)$. The instantiation of $e = 0$ is introduced using Posterior Inference (see later).

Structure estimator. Next, we instantiate $p_\theta(\mathbf{A}|\mathbf{X}, \mathbf{e})$. Edge connections are estimated based on news contents $\mathbf{X}$ and environment variable $\mathbf{e}$. This is in line with reality where posts are connected by interactions (comments/reposts), which can be inferred from the post contents (node features). For simplicity and following common practice [15], we assume that the edges in the graph are conditionally independent. Then, the conditional probability of $\mathbf{A}$ can be factorized as $p_\theta(\mathbf{A}|\mathbf{X}, \mathbf{e}) = \prod_{i,j\in V} p_\theta(a_{ij}|\mathbf{X}, \mathbf{e})$, where $p_\theta(a_{ij}|\mathbf{X}, \mathbf{e})$ represents the probability of an edge existence between nodes i and j given node features $\mathbf{X}$.

As with the classification model, when $\mathbf{e} = 1$, the edge probability is inferred from the news content:

$$p_\theta(a_{ij} = 1|x_i, x_j, \mathbf{e} = 1) = \sigma([\mathbf{U}x_i, \mathbf{U}x_j]^\top \omega), \tag{5}$$

where $\mathbf{U}$ and ω are learnable parameters, and $\sigma(\cdot)$ is the activation function. Since all edges already exist, to avoid model making trivial predictions as all edge probabilities being 1, we sample random edges from the propagation graph with the same number of positive edges for model training.

Same as with the classification model, the structure estimator only instantiates the scenarios for $e = 1$, since for environment-biased samples, the propagation could express unstable or unknown patterns that should not be considered for classification. The instantiation of the structure estimator when $e = 0$ is detailed next.

Posterior inference. Using the instantiations of the classification model, the structure estimator $p_\theta(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{e} = 1)$ and $p_\theta(\mathbf{A}|\mathbf{X}, \mathbf{e} = 1)$ can be inferred. When $e = 0$, to model handle the distribution of environment-biased samples, the classification model prediction is set to Gaussian distribution $\mathcal{N}(0, 1)$, and the edge probability in the structure estimator $p_\theta(a_{ij} = 1|x_i, x_j, \mathbf{e} = 0)$ is set to $\mathcal{N}(0, 1)$. Then, the probability of a sample being environment-independent can be inferred from Equation 2 with the preset prior $p(e)$, which is provided as a hyperparameter.

Training objectives. Based on the instantiation of the classification model and the structure estimator, the environment variable can be inferred according to Equation 2 as shown by the posterior inference part of Fig. 2.

In the training process, we are given the graph node features $\mathbf{X}$, edge connections $\mathbf{A}$ and labels $\mathbf{y}$. Parameter θ is used to perform news classification, structure estimation and environment variable inference (data debiasing). The model is trained by optimizing the Evidence Lower BOund (ELBO) of observed data tuple $(\mathbf{A}, \mathbf{X}, \mathbf{y})$ based on Equation 6.

$$\begin{aligned}\log p_\theta(A, y|X) &\geq \log p_\theta(A, y|X) - D_{KL}[p_\theta(\mathbf{e}|A, X, y)||p(e)] \\ &= E_{p_\theta(e|A,X,y)}[\log p_\theta(A|X, e)p_\theta(y|X, A, e)] \\ &\quad -D_{KL}[p_\theta(\mathbf{e}|A, X, y)||p(e)] = \mathcal{L}_{ELBO}\end{aligned} \tag{6}$$

The final learning objective is the sum of three terms: (1) the classification loss $\mathcal{L}_{cl}$ shown in Equation 7; (2) the structure regularization loss $\mathcal{L}_{reg}$ shown in Equation 8; and (3) the KL divergence loss $\mathcal{L}_{KL}$ between the estimated environment variable and the prior shown in Equation 9.

$$\begin{aligned}\mathcal{L}_{cl} = -\frac{1}{N^{tr}}\sum_{i=1}^{N_{tr}} E_{p_\theta(e_i|A_i,X_i,y_i)}[e_i \log p_\theta(y_i|X_i, A_i, e_i = 1) + \\ (1 - e_i)\log p_\theta(y_i|X_i, A_i, e_i = 0)]\end{aligned} \tag{7}$$

$$\begin{aligned}\mathcal{L}_{reg} = -\frac{1}{N^{tr}}\sum_{i=1}^{N_{tr}} E_{p_\theta(e_i|A_i,X_i,y_i)}[e_i \log p_\theta(A_i|X_i, e_i = 1) + \\ (1 - e_i)\log p_\theta(A_i|X_i, e_i = 0)]\end{aligned} \tag{8}$$

$$\mathcal{L}_{KL} = -\frac{1}{N^{tr}}\sum_{i=1}^{N_{tr}} D_{KL}[p_\theta(\mathbf{e}_i|A_i, X_i, y_i)||p(e)] \tag{9}$$

4 Experiment

To evaluate our model performance on unseen fake news detection, an OOD fake news detection benchmark is used following the approach given in [22]. The datasets used to train the models and the datasets used to test the models are from non-overlapping news

domains. At training, no knowledge of test data is leaked, including news content, propagation graphs, and veracity labels, except for the UCD-RD [22] model, which will use features of the test data to adjust model parameters through contrastive learning.

Datasets. Four public datasets collected from Twitter (now called X) and Weibo (a Chinese social media platform like Twitter) are used. They are `Twitter` [13], `Weibo` [14], `Twitter-COVID19` [9] and `Weibo-COVID19` [9]. `Twitter` and `Weibo` comprise news/posts from general domains (named open-domain), and are treated as the training set. `Twitter-COVID19` and `Weibo-COVID19` only contain news/posts related to COVID-19. They represent data from an emerging/unseen topic. The statistics of the datasets are shown in Table 1.

Table 1: Dataset Statistics ('#' means "number of", 'Avg.' means average).

Statistics	Twitter	T-COVID	Weibo	W-COVID
# events	1,154	400	4,649	399
# tree nodes	60,409	406,185	1,956,449	26,687
# true news	579	148	2,336	146
# fake news	575	252	2313	253
Avg. lifetime	389 Hrs	2,497 Hrs	1,007 Hrs	248Hrs
Avg. depth/tree	11.67	143.03	49.85	4.31
Avg. # posts	52	1,015	420	67
Domain	Open	COVID-19	Open	COVID-19
Language	English	English	Chinese	Chinese

Detection is performed in both cross-domain and cross-language settings to evaluate the models' generalization capability. When testing performance on `Weibo-COVID19` in Chinese, the models are trained on `Twitter` in English, and similarly for `Twitter-COVID19`.

Baselines. To evaluate the model, baselines include sequence-based models **LSTM** [12], **RvNN** [14], Transformer-based models **PLAN** [6], **RoBERTa** [11], graph-based models **BiGCN** [1], **GACL** [24], **SEAGEN** [4] and domain-adaptive model **UCD-RD** [22] are experimented.

Implementation. All baselines and the `FNDCD` model are implemented using PyTorch[1] and trained with an NVIDIA A100 80 GB GPU. The baseline models use default hyperparameter settings from their papers. Hyperparameter $p(e)$ of `FNDCD` indicates the proportion of environment-independent samples. This is set to 0.7 and 0.6 in the source-Twitter and source-Weibo experiments, respectively. Our source code will be released upon paper publication.

Results. The experiment results are shown in Table 2. As seen, the sequence-based methods have the worst performance in both accuracy and F1 score due to their limited feature extraction capability. The graph-based models generally perform better than the sequence-based ones, highlighting the effectiveness of propagation graphs, with the exception of SEAGEN where the performance drops compared to the reported results in its original paper. This may be due to the temporal features that it uses also suffering the distribution shift, i.e., the news originating from different domains has different temporal features. Leveraging data augmentation,

[1]https://pytorch.org/

Table 2: Zero-Shot Fake News Detection on `Twitter-COVID19` and `Weibo-COVID19` (Acc: Accuracy; F-F1: F1 score on fake news detection; T-F1: F1 score on true news detection).

Source	Twitter			Weibo		
Target	Weibo-COVID19			Twitter-COVID19		
Method	Acc	T-F1	F-F1	Acc	T-F1	F-F1
LSTM	0.463	0.329	0.498	0.510	0.243	0.533
RvNN	0.514	0.426	0.538	0.540	0.247	0.534
PLAN	0.532	0.414	0.578	0.573	0.298	0.549
RoBERTa	0.623	0.459	0.711	0.603	**0.585**	0.619
BiGCN	0.569	0.429	0.586	0.616	0.252	0.577
SEAGEN	0.555	0.406	0.583	0.578	0.320	0.650
GACL	0.601	0.410	0.616	0.621	0.345	0.666
UCD-RD	0.631	0.510	0.621	0.591	0.371	0.583
FNDCD	**0.754**	**0.620**	**0.819**	**0.693**	0.513	**0.775**
↑ (%)	+19.49	+21.57	+15.19	+11.59	-12.31	+16.37

contrastive learning and testing features, GACL and UCD-RD's performances are among the best in the baselines. Our `FNDCD`'s superior performance demonstrates the effectiveness of causal debiasing, even though we only use a simple two-layer GCN encoder to encode the propagation graphs.

Case study. The environment inference results are shown in Fig. 3. The distribution of inferred environment variable e is plotted. As can be seen, the majority of training samples are assigned weights around the prior ratio. We can treat the samples with weights far away from the prior ratio as environment-biased samples. From the analysis of the model function, these samples are either difficult to classify or their propagation is hard to estimate. We also find that debiasing using the test samples is actually drawing the latent distributions of the training and testing samples closer together in an unsupervised way.

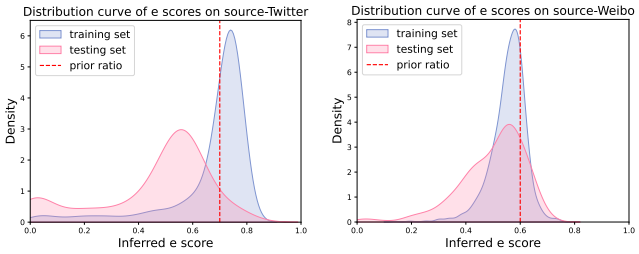


Figure 3: Distribution of inferred environment variable (left: source Twitter dataset, right: source Weibo dataset).

Parameter study. The hyperparameters of the prior distribution $p(e)$ are selected through grid search, as shown in Fig. 4. The results highlight the importance of a reasonable prior: a value set too high assumes all training data is environment-independent, reducing the debiasing effect, while a value set too low treats all training data as biased, leading to model underfitting.

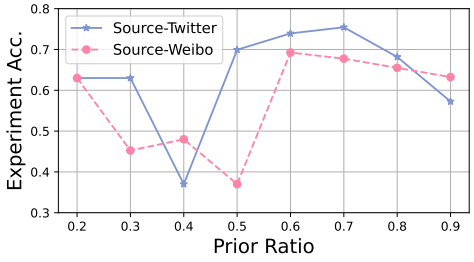


Figure 4: Parameter $p(e)$ sensitivity.

5 Related Work

The detection of fake news has been extensively studied with existing solutions typically relying on news content analysis, propagation structures, or user credibility assessment. However, such solutions often struggle when faced with distributional shifts between training and testing data, leading to performance degradation. To address this issue, researchers have explored cross-domain fake news detection, which focuses on training a model in one domain (the *source domain*) and applying it to another (the *target domain*). Broadly, these solutions can be classified into *sample-level* and *feature-level* approaches.

Sample-level approaches aim to identify training samples that exhibit domain-invariant characteristics, assigning them greater importance during model training [23, 25]. Some studies [22, 25] enhance target domain data by employing clustering algorithms to generate augmented samples, which are then incorporated into training alongside source domain data. This strategy strengthens model robustness when handling unseen domains. Feature-level approaches, on the other hand, focus on identifying and emphasizing domain-independent attributes. For example, reinforcement learning has been applied to select features that remain stable across domains [19]. Inspired by domain-adaptive learning techniques [2], some works [7, 17] train a domain discriminator adversarially to encourage the model to generate news embeddings that obscure domain-specific characteristics, thereby improving generalization.

Our approach takes a step further by leveraging causal analysis on the propagation structure, capturing more informative patterns that contribute to cross-domain fake news detection.

6 Conclusions and Limitations

We demonstrated that FNDCD achieves state-of-the-art performance in detecting unseen fake news by addressing domain-specific biases in training data through causal analysis and reweighting strategies. Besides, the reweighing strategy is only applied during the training process, leaving the trained graph encoders and classifier for the testing (veracity inference) process, to improve the scalability of real-application.

Our work has certain limitations, such as the need to pre-define the prior ratio, which could be enhanced by dynamically estimating a pseudo-environment variable. Additionally, the dependency between textual content and the environment could be more effectively modeled. Leveraging the capabilities of LLMs could address this challenge, as they have been utilized in fake news detection both as supportive agents [5] and as advanced news content processors [16]. Future research could also explore scalable, real-time fake news detection, ideally in collaboration with social media platforms like Twitter/X.

Acknowledgments

This study is supported by Melbourne Research Scholarship and CSIRO Data61 Top-up Scholarship.

References

[1] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *AAAI*.
[2] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*.
[3] Shuzhi Gong, Richard O Sinnott, Jianzhong Qi, and Cecile Paris. 2023. Fake news detection through graph-based neural networks: A survey. *arXiv preprint arXiv:2307.12639* (2023).
[4] Shuzhi Gong, Richard O Sinnott, Jianzhong Qi, and Cecile Paris. 2023. Fake News Detection Through Temporally Evolving User Interactions. In *PAKDD*.
[5] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *AAAI*.
[6] Ling Min Serena Khoo, Hai Leong Chieu, Zhong Qian, and Jing Jiang. 2020. Interpretable rumor detection in microblogs by attending to user interactions. In *AAAI*.
[7] Jingqiu Li, Lanjun Wang, Jianlin He, Yongdong Zhang, and Anan Liu. 2023. Improving rumor detection by class-based adversarial domain adaptation. In *ACM-MM*.
[8] Zenan Li, Qitian Wu, Fan Nie, and Junchi Yan. 2022. GraphDE: A generative framework for debiased learning and out-of-distribution detection on graphs. In *NeurIPS*.
[9] Hongzhan Lin, Jing Ma, Liangliang Chen, Zhiwei Yang, Mingfei Cheng, and Chen Guang. 2022. Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning. In *Findings of the ACL: NAACL*.
[10] Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624* (2021).
[11] Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* 364 (2019).
[12] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *IJCAI*.
[13] Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *ACL*.
[14] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on Twitter with tree-structured recursive neural networks. In *ACL*.
[15] Jiaqi Ma, Weijing Tang, Ji Zhu, and Qiaozhu Mei. 2019. A flexible generative framework for graph-based semi-supervised learning. In *NeurIPS*.
[16] Xiaoxiao Ma, Yuchen Zhang, Kaize Ding, Jian Yang, Jia Wu, and Hao Fan. 2024. On Fake News Detection with LLM Enhanced Semantics Mining. In *EMNLP*.
[17] Erxue Min, Yu Rong, Yatao Bian, Tingyang Xu, Peilin Zhao, Junzhou Huang, and Sophia Ananiadou. 2022. Divide-and-Conquer: Post-user interaction network for fake news detection on social media. In *WWW*.
[18] Todd K Moon. 1996. The expectation-maximization algorithm. *IEEE Signal Processing Magazine* (1996).
[19] Ahmadreza Mosallanezhad, Mansooreh Karami, Kai Shu, Michelle V. Mancenido, and Huan Liu. 2022. Domain adaptive fake news detection via reinforcement learning. In *WWW*.
[20] Berndt Müller, Joachim Reinhardt, and Michael T Strickland. 2012. *Neural networks: an introduction*. Springer Science & Business Media.
[21] Judea Pearl. 2009. *Causality*. Cambridge university press.
[22] Hongyan Ran and Caiyan Jia. 2023. Unsupervised cross-domain rumor detection with contrastive learning and cross-attention. In *AAAI*.
[23] 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 *AAAI*.
[24] Tiening Sun, Zhong Qian, Sujun Dong, Peifeng Li, and Qiaoming Zhu. 2022. Rumor Detection on Social Media with Graph Adversarial Contrastive Learning. In *WWW*.
[25] Zhenrui Yue, Huimin Zeng, Ziyi Kou, Lanyu Shang, and Dong Wang. 2022. Contrastive domain adaptation for early misinformation detection: A case study on COVID-19. In *CIKM*.