Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Invariant Subgraphs for Cross-Domain Fake News Detection via Causal Disentanglement

Shuzhi Gong[1], **Richard O. Sinnott**[1], **Jianzhong Qi**[1], **Cecile Paris**[2]
[1]The University of Melbourne
[2]Data61, CSIRO
{shuzhi, rsinnott, jianzhong.qi}@unimelb.edu.au, Cecile.Paris@data61.csiro.au

Abstract

The spread of misinformation through social media poses significant threats. Recent models using text and graph features have shown promising results in specific misinformation detection scenarios. However, these data-driven models heavily rely on training data that share similar distribution with inference data, limiting their applicability to misinformation from emerging or previously unseen domains, known as *out-of-distribution* (OOD) data. Tackling OOD misinformation is a challenging yet critical task. To address the challenge, we propose the **C**ausal **S**ubgraph-oriented **D**omain **A**daptive misinformation Detection (CSDA) model. Based on a causal analysis, CSDA extracts invariant substructures from news propagation graphs that generalise to OOD data, using a graph neural network-based mask generation process. It uses refined training objectives to ensure high-quality subgraphs. It is further powered by contrastive learning for few-shot scenarios, where a limited amount of OOD data is available for training. Extensive experiments on public social media datasets demonstrate that CSDA effectively handles OOD misinformation detection, achieving a 1.23%∼12.23% accuracy improvement over other state-of-the-art models, covering OOD news domains in politics, entertainments, health, etc.

Introduction

The rapid spread of misinformation on social media has been empirically shown to cause significant societal harm, ranging from public health risks to political polarisation and erosion of institutional trust. Large-scale observational studies demonstrate that false information spreads faster, deeper, and more broadly than true information on social platforms, largely due to social reinforcement rather than content quality alone (Vosoughi, Roy, and Aral 2018). During the COVID-19 pandemic, misinformation about vaccines, treatments, and mortality rates was found to undermine public health responses and intensify the so-called "infodemic," overwhelming society's ability to identify reliable information (Cinelli et al. 2020; Zarocostas 2020). Similar dynamics have been observed in political contexts, where misleading claims and rumours propagate through coordinated user interactions and contribute to misinformation cascades during elections and breaking events (Allcott and Gentzkow 2017; Shu et al. 2017). Crucially, such misinformation often arises in response to novel or rapidly evolving events, where historical data provides limited coverage. As a result, detection models trained on past data domains frequently fail to generalize to emerging topics, highlighting the need for robust approaches that can handle distributional shifts inherent to real-world social media environments.

While early studies (Ma and Gao 2020; Gong et al. 2023b) frequently used the term fake news, or rumours, that term has since become recognized as imprecise and politically charged, prompting a shift, in academic research, toward more neutral terminology such as misinformation. In line with prior work, we define misinformation as "verifiably false or misleading information that is presented as factual, irrespective of whether the intent behind its creation or dissemination is malicious" (Shu, Wang, and Liu 2019; Nakov and Da San Martino 2021). This definition encompasses rumors, false claims, and misleading narratives that circulate widely on social media platforms.

Graph-based misinformation detection methods, which leverage Graph Neural Networks (GNNs) to model news propagation patterns (Gong et al. 2023a), have recently attracted significant attention. Despite their success, existing GNN-based approaches generally assume that training and testing data are drawn from the same data distribution (i.i.d.), an assumption that rarely holds true in practice (Li et al. 2022). New events often give rise to novel propagation behaviours, causing substantial data distribution shifts. From an empirical perspective, most existing methods minimise average training errors and exploit correlations present in the training set (the *in-distribution* data) (Liu et al. 2021). However, such data often contains domain-specific biases. For example, (Zhang et al. 2024) observed that the veracity of political news could be spuriously correlated with specific keywords (e.g., news mentioning the "White House" or "rainbow" being disproportionately classified as true). Models trained on such biased correlations fail to generalise and hence perform poorly on out-of-distribution (OOD) news (Li et al. 2022).

To detect misinformation across different news domains (e.g., politics and sports), early studies (Ma, Gao, and Wong 2018; Bian et al. 2020) attempted to capture content-independent propagation patterns. However, later work (Min et al. 2022) revealed that propagation dynamics them-

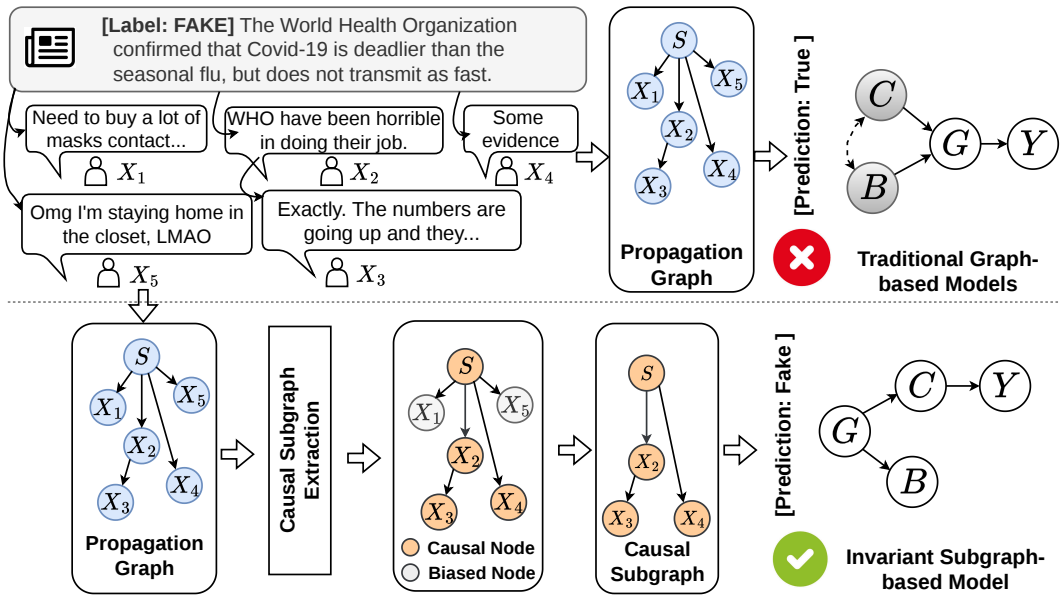

Figure 1: Illustration of invariant subgraphs and our invariant subgraph-based model (bottom). The white (visible) and grey (invisible) colors indicates the visibility of the variables.

selves may vary by domain, limiting the effectiveness of such strategies. More recently, domain adaptation approaches (Lin et al. 2022; Li et al. 2023) have been proposed to transfer trained models to new domains with limited labelled data. While promising, these methods require supervision from the target domain, which is often unavailable or costly to obtain.

In this paper, we adopt a causality-based invariant learning approach. Each propagation graph can be regarded as a mixture of two components: an *invariant subgraph* C, which captures domain-invariant and label-aware structures, and a *domain-biased subgraph* B (abbreviated as biased subgraph in the rest of the paper), which encodes domain-specific and other spurious information. These two components are initially entangled. Our key insight is that not all nodes and edges contribute equally to generalisation: only invariant subgraphs provide stable and sufficient information that can be used for identifying misinformation across unseen domains (Figure 1). By explicitly disentangling invariant subgraphs from biased subgraphs, we can achieve robust generalisation without requiring extensive labelled OOD data.

Based on this intuition, we propose the **C**ausal **S**ubgraph-oriented **D**omain **A**daptive misinformation Detection model (CSDA). CSDA introduces a mask generator that partitions each propagation graph into invariant and biased subgraphs, employing dual encoders for representation learning. To ensure that predictions rely primarily on the invariant signals, we design disentangling training objectives that enforce *invariance* and *sufficiency* representations, while constraining the influence of biased subgraphs. At inference, only the invariant branch is used to produce domain-robust predictions. When a small number of labelled OOD samples are available, CSDA further enhances cross-domain alignment through supervised contrastive learning.

In summary, our contributions are as follows:

- we propose CSDA, a zero-shot misinformation detection model that disentangles domain-invariant subgraphs from domain-specific biased subgraphs in news propagation graphs leveraging causal analysis;
- we extend CSDA to a few-shot setting, introducing a supervised contrastive objective to align causal representations across domains when limited OOD samples are available, and we
- undertake extensive experiments on six real-world datasets (including two used as training data and three as OOD testing data to simulate the real-world unseen misinformation detection). We demonstrate that CSDA consistently outperforms state-of-the-art baselines, achieving absolute accuracy gains of 1.23%–12.23%.

Related Work

Misinformation Detection. Traditional misinformation detection methods have explored *news content*, *social context*, and *social environment* aspects. Content-based methods learn content or stylistic features from textual or multimodal news content (Feng, Banerjee, and Choi 2012), and may leverage external knowledge sources for fact-checking (Samarinas, Hsu, and Lee 2021). While effective in controlled settings, such approaches are often sensitive to topic and domain shifts, as linguistic cues and factual references vary substantially across events.

Social context-based methods exploit user features (Shu, Wang, and Liu 2019) and user interactions observed during news propagation. These include both sequence-based models (Ma et al. 2016; Khoo et al. 2020) and graph-based approaches (Bian et al. 2020; Gong et al. 2023b) that encode

propagation structures using neural networks. These methods assume that propagation patterns learned from historical data remain predictive at inference time. However, later studies show that propagation dynamics themselves can vary across domains and events, limiting cross-domain generalisation.

Environment-based methods such as FANG (Nguyen et al. 2020) explicitly model associations across multiple news domains to extract broader contextual signals. While these approaches capture domain-level regularities, they typically rely on correlations present in observed environments and do not explicitly distinguish between domain-invariant and domain-specific propagation structures.

Cross-domain misinformation detection aims to train a model in one domain (the *source domain*) and apply it to another (the *target domain*). Existing methods can be broadly categorised into *sample-level* and *feature-level* approaches. Sample-level methods identify domain-invariant samples and assign them larger training weights (Silva et al. 2021; Yue et al. 2022), while feature-level methods focus on extracting or reweighting domain-independent representations. For example, (Mosallanezhad et al. 2022) utilise reinforcement learning to select invariant attributes, and adversarial domain adaptation methods (Ganin and Lempitsky 2015; Min et al. 2022; Li et al. 2023) train domain discriminators to encourage indistinguishable representations across domains.

In contrast to these approaches, our CSDA model does not assume access to target-domain data or labels, nor does it attempt to align entire representations across domains. Instead, it explicitly disentangles propagation graphs into invariant invariant subgraphs and domain-specific biased subgraphs, enabling robust zero-shot generalisation to address challenges inherent in previously unseen news domains.

Graph Out-of-Distribution (OOD) Generalisation. Despite the success of graph machine learning, most graph-based methods rely on the assumption that training and testing data are drawn from the same distribution. When this assumption is violated, performance degradation is commonly observed. Recent graph OOD generalisation methods address this issue through data-centric strategies (Feng et al. 2020; Park et al. 2022; Wu et al. 2022a; Zhao et al. 2022; Li et al. 2024), which modify training graphs to improve robustness, and invariant learning approaches (Chen et al. 2022; Miao, Liu, and Li 2022; Wu et al. 2022b; Liu et al. 2023; Yu, Liang, and He 2023; Gui et al. 2024), which aim to identify stable feature–label relationships across environments.

These methods have demonstrated effectiveness in domains such as molecular graphs and computer vision, where the label is assumed to be causally determined by specific graph substructures (e.g., a molecular motif or an image region) (Fan et al. 2022; Chen et al. 2022; Han et al. 2022). Under this assumption, isolating invariant substructures leads to improved generalisation.

However, this causal assumption does not hold in misinformation detection. In social media, the veracity of a news item is fixed by real-world facts and instead influences how users react and propagate the content. As a result, directly applying existing graph OOD methods such as CIGA (Chen et al. 2022) or G-mixup (Han et al. 2022), which are included as baselines in our experiments, fails to capture the inverted causal structure of news propagation and leads to suboptimal performance.

CSDA explicitly models this inverted causality by treating the propagation graph as a consequence of the news veracity and disentangling invariant invariant subgraphs from environment-dependent biased subgraphs. This distinction explains why CSDA differs fundamentally from prior graph OOD methods and achieves superior performance in cross-domain misinformation detection.

Preliminaries

Problem Formulation. Unseen domain misinformation detection seeks to transfer a model trained on a labelled in-distribution dataset to an out-of-distribution (OOD) dataset, which may be unlabelled or contain only a limited number of labelled samples.

Formally, given an in-distribution dataset $D_{in} = \{(G_k^{in}, y_k^{in})\}_{k=1}^{n_{in}}$ drawn from distribution P, the goal is to detect misinformation in an OOD dataset $D_{out} = \{G_k^{out}\}_{k=1}^{n_{out}}$ sampled from a different distribution $P' \neq P$. Both datasets share the same label space, i.e., $y \in \{\text{True}, \text{Fake}\}$. The task is to learn a classifier f from D_{in} that generalises effectively to D_{out}.

Each news item is represented as a propagation graph $G = \langle X, A \rangle$, where the node set X contains the source post and all associated replies or reposts, and the adjacency matrix A encodes the propagation relations. Node features are initialised using text embeddings derived from a pre-trained language model such as BERT.

Data Preparation. For each news item from both D_{in} and D_{out}, the propagation graph $G_k = \langle \mathbf{X}_k, \mathbf{A}_k \rangle$ is extracted and modelled as an undirected acyclic graph. The node set $\mathbf{X}_k = \{x_0, x_1, x_2, \ldots, x_{|\mathbf{X}_k|}\}$ contains all posts including the source news post x_0 and all associated comments/reposts $\{x_1, x_2, x_{|\mathbf{X}|_k}\}$ which can be used to provide supportive information regarding the post veracity. Each post's embedding is initialised using a pre-trained BERT model (Devlin et al. 2019) to compute the text embeddings.

The adjacency matrix $\mathbf{A}_k = \{\alpha_{ij}, i, j \in [1, |C_k|]\}$ is the set of propagation behaviours where an edge exists (i.e., $\alpha_{ij} = 1$) between node i and node j if there is a reply/repost relationship.

Causal Analysis

Structural Causal Model

It is known that OOD generalisation is impossible without assumptions on the environments $\mathcal{E}_{all}$ (Chen et al. 2022). Thus, inspired by (Chen et al. 2022; Fan et al. 2022), we first formulate the data generation process with a structural causal model and latent variable model (Chen et al. 2022), to characterise the distribution shifts that could happen on the misinformation propagation graphs.

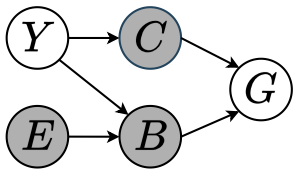

Figure 2: SCM representing the joint data generation in real news. The white (visible) and grey (invisible) colors indicates the visibility of the variables.

We employ Structural Causal Models (SCMs) to formalise the key aspects of the misinformation detection problem and to clarify the interactions among its underlying factors. Our analysis focuses on the role of propagation substructures and their potential to support robust generalisation.

As shown in Fig. 2, in the context of news propagation generation, we consider five variables: the causal invariant subgraph C, the biased subgraph B, the observed whole propagation graph G, the news label Y and the latent environment variable E (e.g. news topics, events and propagation platforms). From a data generation perspective, the veracity label Y influences the propagation process, giving rise to both domain-invariant substructures (C) and domain-specific substructures (B). This causal relationship differs from that assumed in prior works such as (Chen et al. 2022; Fan et al. 2022), where the invariant subgraph is assumed to determine the label, i.e., $C \rightarrow Y$. This discrepancy arises from differences in application scenarios. In those settings, graphs represent objects such as molecular structures or MNIST superpixel graphs, where intrinsic properties of the graph (e.g. molecular attributes) can be attributed to specific subgraphs (e.g. sub-molecules). In contrast, in the social media news context, the veracity of a news item is fixed by real-world facts, and users are assumed to behave differently depending on that veracity, resulting in propagation patterns conditioned on Y ($Y \rightarrow C, Y \rightarrow B$).

Moreover, in addition to the news veracity label Y, parts of the propagation graph are also influenced by an additional latent environment variable E. This phenomenon is common in practice, as certain misinformation-related patterns are only present in specific domains, such as political debates in political contexts or propaganda narratives during the COVID-19 pandemic (Nakov and Da San Martino 2021). We refer to these environment-dependent structures as biased subgraphs B.

Together, C and B constitute the observed graph G. Although the biased subgraphs B are also correlated with the label Y, they are strongly confounded by the environment variable E. The complete structural causal model is illustrated in Fig. 2. Traditional GNN-based misinformation detection approaches (Bian et al. 2020; Gong et al. 2023b) typically learn correlations between Y and G, which inevitably leads to reliance on the biased subgraphs B and consequently poor generalisation to unseen domains.

Formalisation

In this section, two assumptions are derived from existing work (Chen et al. 2022; Fan et al. 2022) to serve as the theoretical foundation of our proposed model. We also elaborate the SCM for the news propagation graph generation process in two assumptions.

Assumption 1 (Graph Generation Structural Causal Model).

$$G_c := f_{gen}^{G_c}(C), G_b := f_{gen}^{G_b}(B), G := f_{gen}^{G}(G_c, G_b),$$

where $f_{gen}^{G_i}$ is the graph generation function of graph $G_i \in \{G_c, G_b, G\}$. The assumption is derived from (Chen et al. 2022) and assumes that the invariant subgraph G_c and biased subgraph G_b causally originate from causal invariant variable C and domain-biased variable B individually, and are seperate subset of the whole graph G.

The Assumption 1 is the foundation to split the graph into two non-overlapping subsets for domain invariant learning.

Assumption 2 (Our Structural Causal Model).

$$C := f_{inv}(Y), B := f_{bias}(Y, E), G := f_{gen}^{G}(C, B),$$

where the f_{inv} is the domain-invariant graph generation, f_{bias} is the domain-biased graph generation, and f_{gen} corresponds the whole graph generation in Assumption 1. The Assumption 2 is the formulation of the SCM in Fig.2.

To enable a GNN to learn and extract the information about C from G, we propose a framework `CSDA` with masking mechanism exactly aligned with Assumption 1, and training objectives aligned with Assumption 2. The idea is that a neural network can learn a reasoning process better if its computation structure aligns with the process better (Xu et al. 2019, 2020).

Specifically, the alignment can be achieved by designing a masking mechanism attempting to split the original graph into invariant and biased, two non-overlapping subgraphs. The training objectives can be designed to train the masking by optimising the classification on both two subgraphs but keep the invariant branch dominant. A detailed instantiation is given under Disentangling Training Objectives of Proposed Model.

Proposed Model

In this section, we present `CSDA`, an invariant subgraph-oriented framework for unseen misinformation detection. The core idea is to disentangle news propagation graphs into causally domain *invariant* subgraphs and domain-specific *biased* subgraphs, and to base predictions primarily on the invariant component. The architecture of `CSDA` is illustrated in Figure 3.

Given a batch of propagation graphs, a mask generator assigns a binary mask to each node and edge, thereby partitioning the graph into an invariant subgraph C and a biased subgraph B. The two subgraphs are encoded separately using individual graph encoders, producing embeddings $\mathbf{z}_c$ and $\mathbf{z}_b$.

This design directly reflects the causal analysis in previous Causal Analysis. First, by predicting labels only from C,

the model encourages *domain invariance*, i.e., predictions that remain stable across environments. Second, the biased branch and hinge loss act as constraints so that C alone is sufficient for prediction, approximating the property of *conditional sufficiency*. Finally, robustness is achieved through data augmentation on B, ensuring that predictions remain unchanged when B is permuted, which corresponds to the idea of *counterfactual stability*.

The overall training objective combines cross-entropy, contrastive, and hinge losses to disentangle C and B while enforcing these properties. During inference, only the invariant branch is used for final prediction, ensuring that outputs are based on domain-invariant and sufficient substructures. CSDA is trained on in-distribution data D_{in} and evaluated on unseen out-of-distribution data D_{out} in a zero-shot manner. When a few labelled OOD samples are available, these can be incorporated with supervised contrastive learning to further enhance robustness.

Mask Generator

Our mask generator learns a mask to help split each propagation graph $\mathcal{G}$ (i.e., $\mathcal{G}_k$ – now we further drop the subscript 'k' as long as the context is clear) into an invariant subgraph $\mathcal{G}_c$ and a biased subgraph $\mathcal{G}_b$. This is achieved by computing node importance scores (denoted as α_i for node i) and edge importance scores (denoted as β_{ij} for the edge between nodes i and j) in the propagation graph $\mathcal{G}$. The aim is to measure the probability of a node or an edge belonging to an invariant subgraph.

The mask generator takes graph $\mathcal{G}$ (i.e., its features) as input and outputs the importance of its nodes and edges. A Graph Isomorphism Network (GIN) (Xu et al. 2018) is utilised to encode the graph and map the node features $\mathbf{X}$ to node embeddings $\mathcal{H}$ for the model's graph structure learning capability, as defined below:

$$\mathcal{Z}^{(l+1)} = \mathrm{MLP}^{(l)}\Big(\Big(\mathbf{A}\mathcal{Z}^{(l)} + (1+\epsilon^{(l)})\mathcal{Z}^{(l)}\Big)\Big), \quad (1)$$

where $l = 0$ or 1, $\mathcal{Z}^{(0)}$ is the initial node features $\mathbf{X}$, $\mathbf{A}$ is the adjacent matrix of the graph without normalization, $\epsilon^{(l)}$ is a learnable scalar controlling the contribution of the central node representation, $\mathrm{MLP}^{(l)}(\cdot)$ denotes a multi-layer perceptron at layer l, and σ is the activation function inside $\mathrm{MLP}^{(l)}$. $\mathcal{Z}^{(0)}$ is initiated as the node feature input $\mathbf{X}$.

After obtaining the graph's node features from GIN output $\mathcal{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N\}$, where N is the size of the node set and $\mathbf{h}_i$ represents the embedding for the i-th node, the node and edge importance scores are computed using an MLP:

$$\alpha_i = \sigma(\mathrm{MLP}([\mathbf{h}_i])), \beta_{ij} = \sigma(\mathrm{MLP}([\mathbf{h}_i, \mathbf{h}_j])), \quad (2)$$

where σ is the activation function.

Since the invariant and the biased subgraphs are defined as two non-overlapping substructures of $\mathcal{G}$, the probability of a node and an edge belonging to a biased subgraph can be established by $(1-\alpha_i)$ and $(1-\beta_{ij})$, respectively.

Using the importance scores, we construct an invariant graph mask $\mathbf{M}_c = [\alpha, \beta]$ and a biased graph mask $\mathbf{M}_b = [(1-\alpha), (1-\beta)]$. Finally, the input propagation graph $\mathcal{G}$ is decomposed into an invariant subgraph $\mathcal{G}_c = \{\mathbf{M}_c \odot \mathcal{G}\}$ and a biased subgraph $\mathcal{G}_b = \{\mathbf{M}_b \odot \mathcal{G}\}$, where $\odot$ is the filtering operation on graph $\mathcal{G}$ with the corresponding masks. The masks emphasise distinct regions of the propagation graphs, enabling subsequent GNN-based graph encoders to concentrate on different segments of the graphs.

Graph Encoder

Two subgraph encoders realised as a 2-layer of stacked GCNII (Chen et al. 2020) are used to encode the invariant and the biased subgraphs. 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 graph embeddings are computed through GCNII by:

$$\mathcal{Z}^{(l+1)} = \sigma\left((\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\mathcal{Z}^{(l)} + \mathcal{Z}^{(0)})(\mathbf{I}_n + \mathbf{W}^{(l)})\right), \quad (3)$$

where $l = 0$ or 1, $\mathcal{Z}^{(0)}$ is the initial node features $\mathbf{X}$, $\tilde{\mathbf{A}}$ is the adjacent matrix of the graph with self-loops, $\tilde{\mathbf{D}}$ is the degree matrix of $\tilde{\mathbf{A}}$, $\mathbf{I_n}$ is the identity mappig from GCNII, $\mathbf{W}^{(l)}$ is the learnable parameter matrix, and σ is the activation function. $\mathcal{Z}^{(0)}$ is initiated as the node feature input $\mathbf{X}$.

As shown in Figure 3, two parallel subgraph encoders are used to encode the invariant subgraph $\mathcal{G}_c$ and the biased subgraph $\mathcal{G}_b$ into an invariant embedding $\mathbf{z}_c$ and a biased embedding $\mathbf{z}_b$. These embeddings will subsequently be fed into news classifiers for loss calculation and misinformation prediction.

Classification Module

The classification module (CM) is responsible for predicting the news veracity based on the extracted graph embeddings. It is composed of an MLP that uses a softmax function. Given the graph embedding $\mathcal{Z}$, e.g. the invariant graph embedding $\mathbf{z}_c$, the CM computes the prediction through:

$$pred = softmax(\mathrm{MLP}(\mathcal{Z})). \quad (4)$$

Since CSDA focuses on classifying news according to causal features, we design an invariant CM and a biased CM in the model. They do not share the parameters and have different input dimensions according to model design. During model training, these two CMs are jointly trained to optimise CSDA to capture invariant information accurately. For model inference, only the prediction results from the invariant CM are used to detect misinformation. More details about the use of the outputs of these two CMs are presented in the next subsection.

Disentangling Training Objectives

The invariant and biased subgraphs are initially entangled in the observed propagation graph. If optimisation is performed solely with prediction loss, the model may converge to trivial solutions (e.g., treating the entire graph as causal), which results in sub-optimal generalisation. In particular, training on the full graph allows domain-specific biases to dominate, leading to overfitting on in-distribution (ID) data and degraded performance on OOD data. To disentangle

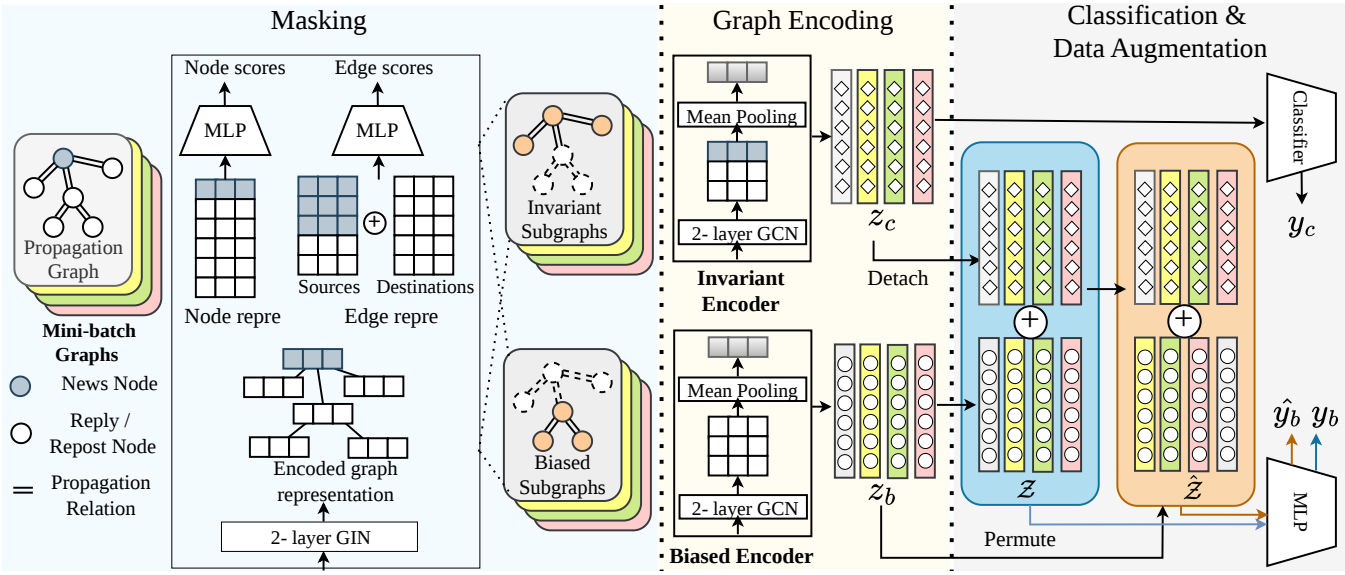

Figure 3: Architecture of CSDA, which is trained with batches of news propagation graphs. A mini-batch of propagation graphs are masked by the Mask Generator and divided into invariant and biased subgraphs. Then, the two branches of subgraphs are encoded using two independent graph encoders to produce invariant and biased embeddings. Afterwards, the invariant embedding is forwarded to an MLP classifier for veracity prediction. Meanwhile, the biased embedding is utilised as part of model optimisation for accurate subgraph extraction.

the invariant and biased subgraphs while optimising prediction accuracy, we integrate three complementary objectives in addition to the standard cross-entropy loss: contrastive learning, data augmentation, and hinge regularisation.

Loss of the invariant branch. The invariant classifier is trained with a cross-entropy objective:

$$\mathcal{L}_{ce}^{c} = CE(y_c, y), \tag{5}$$

where y_c denotes the causal prediction and y is the ground-truth label. However, $\mathcal{L}_{ce}^{c}$ alone cannot ensure faithful separation of invariant and biased subgraphs.

To encourage high-quality invariant subgraph extraction, we introduce a contrastive loss to exploit distributional structure. The key intuition is that invariant subgraphs are label-dependent and should yield similar embeddings for samples with the same label. For two in-distribution samples n and m with identical labels, the contrastive loss is defined as:

$$\mathcal{L}_{CL}^{in} = -\frac{1}{N^{in}} \sum_{n=1}^{N^{in}} \frac{1}{N_{y_n^{in}}} \sum_{m=1}^{N^{in}} \mathbb{1}_{[n \neq m]} \mathbb{1}_{[y_n^{in} = y_m^{in}]} \log \frac{\exp\left(sim(o_n^{in}, o_m^{in})/\tau\right)}{\sum_{k=1}^{N^{in}} \mathbb{1}_{[n \neq k]} \exp(sim(o_n^{in}, o_k^{in})/\tau)}, \tag{6}$$

where N^{in} is the number of ID samples in a batch, $N_{y_n^{in}}$ is the number of samples sharing label y_n^{in}, o_n^{in} is the causal representation of sample n, $sim(\cdot)$ denotes cosine similarity, and τ is a temperature parameter. This loss promotes clustering of invariant embeddings within the same class.

Loss of the Biased Branch. Relying solely on contrastive learning may still encourage the model to treat the entire graph as causal. To address this, we explicitly regulate the biased subgraph. The motivation of the biased branch is to optimise the masking mechanism. As shown in Fig.1, there are causal links $Y \rightarrow B \rightarrow G$ ① and $Y \rightarrow C \rightarrow G$ ②. To train the masking weights to split the whole graph into invariant and biased subgraphs (i.e. $C \leftarrow G \rightarrow B$), we need to optimise both the invariant and biased branch classification, as links ①, ② indicating label Y influences both C and B. The link ① has been optimised through loss (4) and (5), whilst link ② is to be optimised in this branch.

In practice, we expect the trained model to infer label Y through C but not B. The B only helps to improve the masking. Following prior work such as (Lee et al. 2021; Fan et al. 2022) to train the masking mechanism efficiently, we adapt a data augmentation approach to form two concatenated embeddings:

$$\mathcal{Z} = [\text{detach}(\mathbf{z}_c), \mathbf{z}_b], \quad \hat{\mathcal{Z}} = [\text{detach}(\mathbf{z}_c), \text{permute}(\mathbf{z}_b)],$$

where $\mathbf{z}_c$ and $\mathbf{z}_b$ are the invariant and biased embeddings, respectively. Gradient detachment *detach()* prevents biased-branch optimisation from affecting the invariant branch's gradient calculation. $\mathcal{Z}$ is forwarded to the biased classifier to produce prediction y_b, while $\hat{\mathcal{Z}}$ with permuted biased embeddings yields $\hat{y}_b$ with permuted labels $\hat{y}$. This design encourages the biased branch to focus exclusively on biased features.

Next, to prevent the model from being dominated by the biased subgraphs (e.g. the whole graph is classified as biased), we adopt a restricted hinge loss (Chen et al. 2022):

$$\mathcal{L}_{ce}^{b} = \frac{1}{N} CE(\hat{y}_b, \hat{y})\, \mathbb{1}[CE(y_b, y) \leq CE(\hat{y}_b, \hat{y})], \quad (7)$$

which back-propagates only when the joint (causal+biased) prediction is no worse than the biased-only prediction. This ensures that biased information does not subsume the invariant subgraph.

Overall Objective. The total loss function of CSDA is expressed as:

$$\mathcal{L} = \mathcal{L}_{ce}^{c} + \mathcal{L}_{ce}^{b} + \gamma \cdot \mathcal{L}_{CL}^{in}, \quad (8)$$

where γ controls the relative weight of the contrastive objective. This composite loss enforces invariance and sufficiency of the invariant subgraph while constraining the influence of domain-specific biases.

Model Fine-tuning with OOD Data

CSDA can be trained solely on in-distribution (ID) data D_{in}. In practice, however, a small number of labelled OOD samples may also be available. In such cases, CSDA can be further fine-tuned to better adapt to the target domain. This subsection describes how fine-tuning is performed through an additional supervised contrastive objective.

When OOD samples are provided, the goal is to enhance cross-domain alignment of the causal representation space. Specifically, we encourage embeddings of ID and OOD samples with the same veracity label to be close, while ensuring separation across different labels. This design reinforces the *invariance* of causal representations, ensuring that the predictive relationship $P(Y \mid C)$ remains stable across domains.

Formally, we introduce an additional supervised contrastive loss. For an OOD sample n and an ID sample m sharing the same label, the loss is defined as:

$$\mathcal{L}_{CL}^{out} = -\frac{1}{N^{out}} \sum_{n=1}^{N^{out}} \frac{1}{N_{y_n^{out}}} \sum_{m=1}^{N^{in}} \mathbb{1}_{[y_n^{out}=y_m^{in}]} \cdot \log \frac{\exp\left(sim(o_n^{out}, o_m^{in})/\tau\right)}{\sum_{k=1}^{N^{in}} \exp\left(sim(o_n^{out}, o_k^{in})/\tau\right)}, \quad (9)$$

where N^{out} and N^{in} are the numbers of OOD and ID samples in the batch, respectively; $N_{y_n^{out}}$ is the number of ID samples sharing the same label as OOD sample n; o_n^{out}, o_m^{in}, and o_k^{in} denote the corresponding causal representations extracted by CSDA; and τ is the temperature hyperparameter. This loss explicitly draws together embeddings of samples with the same veracity across domains, reinforcing causal sufficiency and discouraging reliance on domain-specific biases.

The overall fine-tuning loss of CSDA is thus updated to:

$$\mathcal{L}' = \mathcal{L}_{ce}^{c} + \mathcal{L}_{ce}^{b} + \gamma \cdot \mathcal{L}_{CL}^{out}, \quad (10)$$

where γ is the same weighting coefficient as in Equation 8. This formulation enforces that invariant embeddings are both *invariant across domains* and *sufficient for prediction*, thereby improving generalisation to unseen environments.

Experiment

Experimental Settings

Datasets Five public datasets collected from Twitter (now called X) and Weibo (a Chinese social media platform like Twitter) are utilised in the experiments: (1) Twitter (Ma, Gao, and Wong 2017), (2) Weibo (Ma et al. 2016), (3) Twitter-COVID19 (Lin et al. 2022), (4) Weibo-COVID19 (Lin et al. 2022) and (5) PHEME (Kochkina, Liakata, and Zubiaga 2018). Twitter and Weibo are open-domain datasets. They cover a variety of topics except COVID-19 and are used as the main training set. Twitter-COVID19 and Weibo-COVID19 only contain news related to COVID-19, which represent the OOD data. PHEME contains hybrid topics news collected from Twitter (now X).

To showcase the effectiveness of CSDA, two set of experiments are designed. In the first set of experiments, the models are trained on in-distribution data (e.g., Twitter) and tested on OOD data (e.g., Twitter-COVID19), to simulate the scenario where no prior knowledge about the OOD data is available. In the second set of experiments, a few OOD samples (i.e., 80 data samples of Twitter-COVID19 and Weibo-COVID19) are utilised to help optimising the models together with in-distribution data (e.g., Twitter), to simulate the scenario where we have a small number of manually labelled OOD samples. The remaining OOD data (e.g, 80% of Twitter-COVID19) are used for model testing.

Baselines We compare with 14 models including recent models DELL (Wan et al. 2024), UCD-RD (Ran and Jia 2023), CADA (Li et al. 2023), and graph OOD generalisation methods G-mixup (Han et al. 2022), CIGA (Chen et al. 2022).

Baseline models that are trained with in-distribution data only include: **LSTM** (Ma et al. 2016) which uses an LSTM-based model to learn feature representations of relevant posts over time;**CNN** (Yu et al. 2017) uses a CNN model for misinformation identification by modelling the relevant posts as a fixed-length sequence; **RvNN** (Ma, Gao, and Wong 2018) which learns the propagation of news by exploiting a tree structured recursive neural network; **PLAN** (Khoo et al. 2020) which uses a Transformer (Vaswani et al. 2017)-based model for misinformation detection by capturing long-distance interactions between tweets (source posts and associated comments); **RoBERTa** (Liu et al. 2019) which encodes the text information of a news item and classifies the news based on the text classification; **BiGCN** (Bian et al. 2020) which models news propagation by representing social media posts as nodes in a graph, and then it utilises a GCN-based model to encode the graph and classifies whether a given news item is true or fake; **GACL** (Sun et al. 2022) which enhances BiGCN by generating adversarial training samples and training with contrastive learning; **SEAGEN** (Gong et al. 2023b) which models the news propagation process by encoding the temporal propagation graph with a temporal graph network (TGN) and a neural Hawkes process; **UCD-RD** (Ran and Jia 2023) which uses prototype-based contrastive learning to initialise pro-

Source	Twitter						Weibo						Inf-Time	Train-Time	GPU-Memory
	Twitter-COVID19		Weibo-COVID19		PHEME		Twitter-COVID19		Weibo-COVID19		PHEME				
Method	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	(ms)	(s/epoch)	(bach-128)
LSTM	0.412	0.383	0.463	0.414	0.560	0.565	0.510	0.388	0.416	0.422	0.550	0.320	2-4	7	3MB
CNN	0.406	0.368	0.445	0.402	0.425	0.512	0.498	0.389	0.421	0.410	0.568	0.280	2-4	7-8	526MB
RvNN	0.436	0.430	0.514	0.482	0.580	0.551	0.540	0.391	0.479	0.493	0.571	0.574	2-4	8	42MB
PLAN	0.455	0.454	0.532	0.496	0.392	0.550	0.573	0.424	0.384	0.372	0.522	0.447	1-3	6-7	67MB
RoBERTa	0.479	0.481	0.623	0.585	0.416	0.523	0.603	0.602	0.680	0.676	0.544	0.401	0.5-1	1	15MB
PLAN-IRM	0.567	0.310	0.501	0.004	0.561	0.447	0.567	0.512	0.658	0.653	0.626	0.239	1-3	8-10	15MB
RoBERTa-IRM	0.703	0.624	0.610	**0.740**	0.408	0.541	0.531	0.137	0.709	0.717	0.617	0.152	0.5-1	1	15MB
BiGCN	0.468	0.451	0.569	0.508	0.385	0.544	0.616	0.415	0.612	0.561	0.545	0.527	5-6	18	150-200MB
SEAGEN	0.494	0.471	0.555	0.495	0.499	0.521	0.578	0.485	0.586	0.519	0.566	0.574	38-40	50	450-500MB
GACL	0.541	0.541	0.601	0.513	0.417	0.550	0.621	0.505	0.688	0.681	0.475	0.370	25-30	60	180-220MB
UCD-RD	0.665	0.610	0.631	0.566	0.492	0.519	0.591	0.477	0.689	0.617	0.581	0.402	21-24	33	200-240MB
G-mixup	0.395	0.289	0.549	0.543	0.382	0.301	0.388	0.291	0.431	0.350	0.499	0.503	1-2	5	50-80MB
CIGA-GCN	0.492	0.488	0.672	0.660	0.394	0.408	0.542	0.541	0.694	0.681	0.596	0.583	2-3	5-6	220-270MB
CIGA-GIN	0.475	0.465	0.627	0.570	0.408	0.355	0.450	0.400	0.732	0.702	0.499	0.496	2-3	5-6	250-300MB
CSDA (ours)	**0.713**	**0.660**	**0.701**	0.655	**0.584**	**0.601**	**0.697**	**0.692**	**0.741**	**0.765**	**0.671**	**0.567**	7-9	20	350-450MB

Table 1: Zero-shot misinformation Detection on Multiple Targets (Acc: Accuracy; F1: F1 score on misinformation detection; Inference/Train Time measured per sample / per run; GPU Memory measured as peak usage).

totypes via in-distribution samples, and then it aligns the OOD data features with the corresponding prototypes. In addition to these traditional misinformation detection methods, we also compare with textual OOD generalisation methods (PLAN-IRM, RoBERTa-IRM (Arjovsky et al. 2019)) and graph OOD methods including **G-mixup** (Han et al. 2022), **CIGA** (Chen et al. 2022).

Baseline models trained with both in-distribution and low-resource OOD data include: **ACLR** (Lin et al. 2022) which utilises adversarial contrastive learning to transfer pre-trained BiGCN (Bian et al. 2020) models from a source domain to a target domain for misinformation detection; **CADA** (Li et al. 2023) which serves as a plug-in module that adapts pre-trained models from a source domains to a target domain based on label-aware domain adversarial neural networks (Ganin and Lempitsky 2015). In our experiments, CADA uses BiGCN, RoBERTa, SEAGEN and GACL as the pre-trained models. The web-retrieval and Large Language Model (LLM) prompt-based methods **DELL** (Wan et al. 2024) and RAG-based method FIRE (Xie et al. 2025) are also compared.

All baselines and CSDA are implemented in Pytorch[1] and trained using an A100 GPU. The baseline models use the default hyperparameter settings from their original papers. Hyperparameter γ, τ of the CSDA model are set to 0.2, 0.1 respectively in the experiments. The hyperparameters are selected empirically based on a grid search. The parameter sweeping results are shown in Fig. 4.

Results

Table 2 and Table 1 present the model performance on the four dataset settings (from Twitter, Weibo to Twitter-COVID19, Weibo-COVID19.

In Table 1, the models are categorized into two groups.

[1]https://pytorch.org/

Method	Twitter-COVID19			Weibo-COVID19		
	Acc	T-F1	F-F1	Acc	T-F1	F-F1
$\text{CADA}_{\text{BiGCN}}$	0.681	0.621	0.725	0.716	0.552	0.792
$\text{CADA}_{\text{RoBERTa}}$	0.711	0.540	0.790	0.839	0.783	0.878
$\text{CADA}_{\text{SEAGEN}}$	0.669	0.383	0.785	0.662	0.471	0.752
$\text{CADA}_{\text{GACL}}$	0.641	0.511	0.716	0.684	0.402	0.786
ACLR	0.741	0.607	**0.799**	0.897	0.847	0.917
DELL	0.446	0.384	0.497	0.800	0.743	0.852
FIRE	0.482	0.402	0.511	0.891	0.899	0.882
$\text{CSDA}_{\text{Fine-Tuned}}$	**0.772**	**0.767**	0.797	**0.922**	**0.884**	**0.940**
↑ (%)	+4.18	+26.36	-0.25	+2.79	+4.37	+2.51

Table 2: Few-shot misinformation detection on Twitter-COVID19 (trained on Twitter) and Weibo-COVID19 (trained on Weibo) (Acc: Accuracy; T-F1: F1 score on true news; F-F1: F1 score on misinformation).

The upper group consists of sequence-based models (LSTM, CNN, RvNN, PLAN, and RoBERTa), while the bottom group includes graph-based models (BiGCN, SEAGEN, GACL, UCD-RD, G-Mixup, CIGA, and CSDA). Overall, the graph-based models outperform the sequence-based ones, underscoring the effectiveness of leveraging propagation graphs for misinformation detection. Among the graph-based models, CSDA consistently achieves the best performance across both datasets in terms of accuracy and F1 scores.

Baseline models that do not account for OOD data generally exhibit poor performance. These models are trained on open-domain in-distribution datasets and are therefore biased by domain-specific information. UCD-RD seeks to align the representations of in-distribution and OOD news samples belonging to the same class. However, it fails to address domain biases, making it less effective than CSDA. The graph OOD generalization method, CIGA, demonstrates

significant improvements only on the Weibo-COVID19 dataset, whereas G-Mixup fails to deliver any notable improvements. This may be because these methods are designed for more sophisticated graph structures and are less suited to news propagation graphs, which feature simpler structures but more complex node attributes.

Our experimental setting further considers cross-platform and cross-language transfer by training models on the English Twitter data and testing on the Chinese Weibo-COVID19 data, and vice versa. This setting introduces substantial distribution shifts stemming from differences in language, user populations, and platform-specific propagation dynamics, where most baseline methods exhibit pronounced performance degradation. In contrast, CSDA shows consistently smaller performance drops, indicating that invariant propagation subgraphs generalise more robustly across platforms and languages than domain- or language-specific correlations.

As shown in Table 2, when labelled OOD data is available, the baseline models (BiGCN, RoBERTa, SEAGEN and GACL) powered by CADA can learn features from the OOD data and achieve better accuracy than their vanilla versions. ACLR, which is designed for domain adaptation, achieves even better performance. However, these models are still outperformed by CSDA using fine-tuning with a performance improvement of $2.79 \sim 4.18\%$. DELL has good performance on the Weibo-COVID19 dataset but performs poorly on the Twitter-COVID19 data set, showing both promising results and limitations of LLMs in misinformation detection, providing potential for future work.

Significance Test

Table 3 is the performance with invariance under 5 predefined random seeds (0, 42, 43, 2025, 2026). We select the transfer from Twitter dataset to Twitter-COVID19 and Weibo-COVID19 (simplified as Twitter19 and Weibo19 in Table 3) they are representative enough to simulate the cross-domain, corss-platform scenarios. The CSDA has significant superior performance on Twitter19's accuracy, F1 than CIGA-GCN ($p < 0.01$) and RoBERTa-IRM ($p < 0.05$). But CSDA has inferior performance's F1 score. This can be resulted from the imbalanced label distribution. CSDA has better precision and recall on true-label data but perform poorly on false-label data.

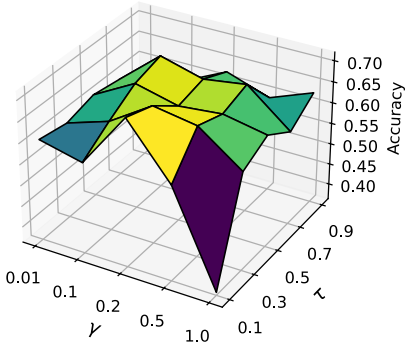

Figure 4: Parameter sensitivity analysis.

Method	Twitter19		Weibo19	
	Acc	F1	Acc	F1
CIGA-GCN	0.492±0.035	0.488±0.039	0.672±0.009	0.660±0.036
RoBERTa-IRM	0.703±0.042	0.624±0.058	0.610±0.011	0.740±0.017
CSDA	0.713±0.032	0.660±0.019	0.701±0.010	0.655±0.018

Table 3: Transfer Evaluation from Twitter to Twitter19 and Weibo19 (Acc: Accuracy; F1: F1 score).

Case Study

Successful Example. To further illustrate the interpretability and effectiveness of CSDA, we conduct a case study using the Twitter and Twitter-COVID19 datasets. Specifically, the mask generator trained on Twitter is applied to unseen Twitter-COVID19 samples to disentangle causal from biased subgraphs. As shown in Table 4, the source news item that mimics an official tone receives a near-zero node score, reflecting that its linguistic content alone is insufficient to determine veracity. In contrast, user comments that provide factual reasoning (e.g., citing rigorous testing protocols in South Korea) are assigned higher importance scores, while irrelevant or promotional content is down-weighted. A similar pattern is observed in the second example, where community responses effectively highlight the falsity of the news despite its credible surface tone.

This differentiation demonstrates that CSDA's mask generator is capable of elevating domain-invariant, causally relevant information while suppressing spurious or domain-specific cues. Consequently, the invariant encoder focuses on subgraphs that generalise across domains, thereby yielding robust OOD detection performance.

Failure Analysis Drawn from all cross-domain detection shown in Table 1, our analysis identifies two fundamental failure modes in cross-domain rumor detection. First, GNN-based models exhibit pronounced structural overfitting to propagation topology. Specifically, models trained on deep and narrow propagation trees (average depth $\sim$4, width $\sim$9 nodes per level) degrade severely when exposed to shallow yet highly expansive viral cascades (average depth $\sim$2, width $\sim$19 nodes per level), a pattern commonly observed in COVID-19 misinformation. The fixed three-layer architecture is implicitly tuned to the depth characteristics of the training distribution and consequently fails to adapt to out-of-distribution propagation structures. Second, semantic domain shifts manifest as substantial distortions in the embedding space, including a 27.5% increase in feature norms, polarity inversions across dimensions, and a 44% reduction in variance captured by the leading PCA components. These observations indicate that the learned representations lack robustness across temporal and topical domain shifts.

News, Comments, Node Scores and Edge Scores
Source news: The World Health Organization confirmed that Covid-19 is deadlier than the seasonal flu, but does not transmit as flu... [Node Score: <0.001] [Label: **FAKE**]
Comment 1: Need to buy a lot of masks contact me. [Node Score: 0.154] [Edge 0→1 Score: 0.152]
Comment 2: Because of their more rigorous testing protocols, South Korea's mortality rate of 0.6% is the most accurate. [Node Score: 0.393] [Edge 0→2 Score: 0.515]
Comment 3: Why don't you look at implementing #Covid_19 travel health cards that confirm the person has been... [Node Score: 0.514] [Edge 0→3 Score: 0.462]
Comment 4: WHO is also omitting mild cases from their stats. [Node Score: 0.556] [Edge 0→4 Score: 0.574]
Source news: Rumours are no less infectious than #coronavirus! This looks like a meticulous list, but a fake one too... URL [Node Score: 0.100] [Label: **True**]
Comment 1: Yeah, this is fake coz you guys have totally disallowed stores from delivering essent... [Node Score: 0.446] [Edge Score: 0.554]
Comment 2: Why are police personnel beating up vegetable vendors and delivery guys... [Node Score: <0.01] [Edge Score: 0.036]
Comment 3: We have forwarded your query to the xxx. You can contact them on xxx-xxxxxx. [Node Score: 0.190] [Edge Score: 0.809]

Table 4: Case study of CSDA's successful examples. from the Twitter-COVID19 dataset.

Ablation Study

Training Objective Ablation. To evaluate the contributions of different loss components and the invariant subgraph extraction module, we conduct an ablation study with four model variants, reported in Figure 5.

- **Only** $\mathcal{L}_{ce}^{c}$: The model is trained using only the causal cross-entropy loss. Without disentangling, the model tends to overfit to domain-specific propagation patterns, resulting in limited generalisation.
- **+**$\mathcal{L}_{CL}^{in}$: Adding the contrastive loss improves representation consistency across samples with the same label, leading to a clear performance gain.
- **+**$\mathcal{L}_{ce}^{b}$ **(without hinge)**: Introducing the biased branch loss without hinge constraints reduces performance. This occurs because the biased classifier is optimised on both invariant and biased embeddings indiscriminately, allowing gradients from spurious correlations in the biased branch to interfere with the invariant encoder. In the absence of hinge restrictions, biased-only predictions are weighted equally, which dilutes the causal signal and impairs OOD generalisation.
- **Full model (with hinge loss)**: Incorporating the hinge loss ensures that updates from the biased branch are propagated only when they do not conflict with the invariant branch. This stabilises training and prevents biased signals from dominating. The final CSDA achieves the best performance, confirming the necessity of the hinge mechanism for effective disentanglement.

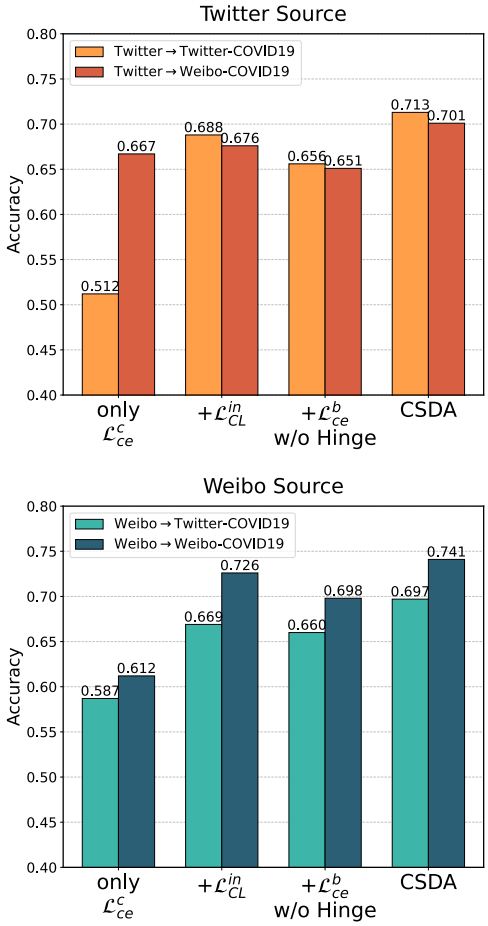

Figure 5: Ablation study under different training sources.

Overall, these results validate the design of the disentangling training objectives. In particular, the hinge loss plays a crucial role in balancing the optimisation of invariant and biased branches, safeguarding the invariant encoder from being overwhelmed by spurious domain-specific features.

Model Component Ablation We evaluate CSDA under two ablation scenarios: (i) replacing the graph encoders in both branches with NN, GCN, and GIN, and (ii) comparing soft masks with hard (binary) masks to assess the impact of mask granularity.

The results in Table 5 show two clear trends. First, CSDA consistently benefits from more expressive encoders, with GIN achieving the best performance across all settings.

Setting	Mask	NN	GCN	GIN
Twitter → Twitter-COVID19	Binary	0.580	0.535	0.656
	Soft	0.628	0.693	0.713
Twitter → Weibo-COVID19	Binary	0.645	0.626	0.677
	Soft	0.665	0.697	0.701

Table 5: Ablation study on encoder choice and mask type.

Second, soft masks uniformly outperform binary masks for every encoder and transfer scenario, demonstrating the advantage of fine-grained importance weighting.

These findings suggest that both encoder expressiveness and mask granularity are critical for effective causal–biased subgraph disentanglement, ultimately leading to improved cross-domain generalisation.

Conclusions

In this paper, we proposed CSDA, an invariant subgraph-oriented framework for cross-domain misinformation detection. Unlike prior approaches that rely heavily on domain-specific correlations, CSDA disentangles news propagation graphs into invariant and biased subgraphs. By ensuring that predictions depend on the *invariant* and *sufficient* invariant subgraphs, the model achieves robust generalisation to out-of-distribution domains. From a causal perspective, the framework implicitly addresses counterfactual queries: if biased substructures were removed or altered, the prediction would remain unchanged so long as the invariant subgraph is preserved. This robustness is validated through extensive experiments, where CSDA consistently outperforms sequence-based and graph-based baselines across zero-shot and few-shot settings. The additional fine-tuning strategy with limited OOD samples further enhances adaptability.

This work uses publicly available benchmark datasets that have been widely adopted in prior misinformation studies and are processed in accordance with their original licensing terms, without attempting to infer additional personally identifiable information. However, automated misinformation classification systems may be misused if model outputs—such as mask or node importance scores—are misinterpreted as indicators of user credibility or community value. In addition, false positives, particularly during crisis scenarios or rapid deployment, may suppress accurate information or mislead decision-making. We therefore emphasise that CSDA is intended as a decision-support tool and should be deployed with appropriate human oversight and contextual safeguards.

Acknowledgments

This study is supported by Melbourne Research Scholarship by The University of Melbourne and CSIRO Postgraduate Scholarships by Data61.

References

Allcott, H.; and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2): 211–236.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant Risk Minimization. *arXiv preprint arXiv:1907.02893*.

Bian, T.; Xiao, X.; Xu, T.; Zhao, P.; Huang, W.; Rong, Y.; and Huang, J. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *AAAI*.

Chen, M.; Wei, Z.; Huang, Z.; Ding, B.; and Li, Y. 2020. Simple and deep graph convolutional networks. In *ICML*.

Chen, Y.; Zhang, Y.; Bian, Y.; Yang, H.; Ma, K.; Xie, B.; Liu, T.; Han, B.; and Cheng, J. 2022. Learning Causally Invariant Representations for Out-of-Distribution Generalization on Graphs. In *NeurIPS*.

Cinelli, M.; Quattrociocchi, W.; Galeazzi, A.; Valensise, C. M.; Brugnoli, E.; Schmidt, A. L.; Zola, P.; Zollo, F.; and Scala, A. 2020. The COVID-19 social media infodemic. *Scientific reports*, 10(1): 16598.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Fan, S.; Wang, X.; Mo, Y.; Shi, C.; and Tang, J. 2022. Debiasing Graph Neural Networks via Learning Disentangled Causal Substructure. In *Advances in Neural Information Processing Systems*.

Feng, S.; Banerjee, R.; and Choi, Y. 2012. Syntactic stylometry for deception detection. In *ACL*.

Feng, W.; Zhang, J.; Dong, Y.; Han, Y.; Luan, H.; Xu, Q.; Yang, Q.; Kharlamov, E.; and Tang, J. 2020. Graph Random Neural Networks for Semi-Supervised Learning on Graphs. In *NeurIPS*.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gong, S.; Sinnott, R. O.; Qi, J.; and Paris, C. 2023a. Fake news detection through graph-based neural networks: A survey. *arXiv preprint arXiv:2307.12639*.

Gong, S.; Sinnott, R. O.; Qi, J.; and Paris, C. 2023b. Fake News Detection Through Temporally Evolving User Interactions. In *PAKDD*.

Gui, S.; Liu, M.; Li, X.; Luo, Y.; and Ji, S. 2024. Joint learning of label and environment causal independence for graph out-of-distribution generalization. In *NeurIPS*.

Han, X.; Jiang, Z.; Liu, N.; and Hu, X. 2022. G-Mixup: Graph data augmentation for graph classification. In *ICML*.

Khoo, L. M. S.; Chieu, H. L.; Qian, Z.; and Jiang, J. 2020. Interpretable rumor detection in microblogs by attending to user interactions. In *AAAI*.

Kochkina, E.; Liakata, M.; and Zubiaga, A. 2018. PHEME dataset for Rumour Detection and Veracity Classification.

Lee, J.; Kim, E.; Lee, J.; Lee, J.; and Choo, J. 2021. Learning debiased representation via disentangled feature augmentation. In *NeurIPS*.

Li, J.; Wang, L.; He, J.; Zhang, Y.; and Liu, A. 2023. Improving rumor detection by class-based adversarial domain adaptation. In *ACM-MM*.

Li, X.; Gui, S.; Luo, Y.; and Ji, S. 2024. Graph Structure Extrapolation for Out-of-Distribution Generalization. In *ICML*.

Li, Z.; Wu, Q.; Nie, F.; and Yan, J. 2022. GraphDE: A generative framework for debiased learning and out-of-distribution detection on graphs. In *NeurIPS*.

Lin, H.; Ma, J.; Chen, L.; Yang, Z.; Cheng, M.; and Guang, C. 2022. Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning. In *Findings of the ACL: NAACL*.

Liu, J.; Shen, Z.; He, Y.; Zhang, X.; Xu, R.; Yu, H.; and Cui, P. 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

Liu, Y.; Ao, X.; Feng, F.; Ma, Y.; Li, K.; Chua, T.-S.; and He, Q. 2023. FLOOD: A Flexible Invariant Learning Framework for Out-of-Distribution Generalization on Graphs. In *KDD*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ma, J.; and Gao, W. 2020. Debunking rumors on Twitter with tree transformer. In *COLING*. ACL.

Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B. J.; Wong, K.-F.; and Cha, M. 2016. Detecting rumors from microblogs with recurrent neural networks. In *IJCAI*.

Ma, J.; Gao, W.; and Wong, K.-F. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *ACL*.

Ma, J.; Gao, W.; and Wong, K.-F. 2018. Rumor detection on Twitter with tree-structured recursive neural networks. In *ACL*.

Miao, S.; Liu, M.; and Li, P. 2022. Interpretable and Generalizable Graph Learning via Stochastic Attention Mechanism. *ICML*.

Min, E.; Rong, Y.; Bian, Y.; Xu, T.; Zhao, P.; Huang, J.; and Ananiadou, S. 2022. Divide-and-Conquer: Post-user interaction network for fake news detection on social media. In *WWW*.

Mosallanezhad, A.; Karami, M.; Shu, K.; Mancenido, M. V.; and Liu, H. 2022. Domain adaptive fake news detection via reinforcement learning. In *WWW*.

Nakov, P.; and Da San Martino, G. 2021. Fake news, disinformation, propaganda, media bias, and flattening the curve of the COVID-19 infodemic. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 4054–4055.

Nguyen, V.-H.; Sugiyama, K.; Nakov, P.; and Kan, M.-Y. 2020. FANG: Leveraging social context for fake news detection using graph representation. In *CIKM*.

Park, H.; Lee, S.; Kim, S.; Park, J.; Jeong, J.; Kim, K.-M.; Ha, J.-W.; and Kim, H. J. 2022. Metropolis-Hastings Data Augmentation for Graph Neural Networks. In *NeurIPS*.

Ran, H.; and Jia, C. 2023. Unsupervised cross-domain rumor detection with contrastive learning and cross-attention. In *AAAI*.

Samarinas, C.; Hsu, W.; and Lee, M. L. 2021. Improving evidence retrieval for automated explainable fact-checking. In *NAACL*.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1): 22–36.

Shu, K.; Wang, S.; and Liu, H. 2019. Beyond news contents: The role of social context for fake news detection. In *WSDM*.

Silva, A.; Luo, L.; Karunasekera, S.; and Leckie, C. 2021. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *AAAI*.

Sun, T.; Qian, Z.; Dong, S.; Li, P.; and Zhu, Q. 2022. Rumor Detection on Social Media with Graph Adversarial Contrastive Learning. In *WWW*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *science*, 359(6380): 1146–1151.

Wan, H.; Feng, S.; Tan, Z.; Wang, H.; Tsvetkov, Y.; and Luo, M. 2024. DELL: Generating Reactions and Explanations for LLM-Based Misinformation Detection. In *ACL-Findings*.

Wu, L.; Lin, H.; Huang, Y.; and Li, S. Z. 2022a. Knowledge Distillation Improves Graph Structure Augmentation for Graph Neural Networks. In *NeurIPS*.

Wu, Y.; Wang, X.; Zhang, A.; He, X.; and Chua, T.-S. 2022b. Discovering Invariant Rationales for Graph Neural Networks. In *ICLR*.

Xie, Z.; Xing, R.; Wang, Y.; Geng, J.; Iqbal, H.; Sahnan, D.; Gurevych, I.; and Nakov, P. 2025. FIRE: Fact-checking with Iterative Retrieval and Verification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2901–2914.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2018. How Powerful are Graph Neural Networks? In *ICLR*.

Xu, K.; Li, J.; Zhang, M.; Du, S. S.; Kawarabayashi, K.-i.; and Jegelka, S. 2019. What can neural networks reason about? *arXiv preprint arXiv:1905.13211*.

Xu, K.; Zhang, M.; Li, J.; Du, S. S.; Kawarabayashi, K.-i.; and Jegelka, S. 2020. How neural networks extrapolate: From feedforward to graph neural networks. *arXiv preprint arXiv:2009.11848*.

Yu, F.; Liu, Q.; Wu, S.; Wang, L.; Tan, T.; et al. 2017. A Convolutional Approach for Misinformation Identification. In *IJCAI*.

Yu, J.; Liang, J.; and He, R. 2023. Mind the Label Shift of Augmentation-Based Graph OOD Generalization. In *CVPR*.

Yue, Z.; Zeng, H.; Kou, Z.; Shang, L.; and Wang, D. 2022. Contrastive domain adaptation for early misinformation detection: A case study on COVID-19. In *CIKM*.

Zarocostas, J. 2020. How to fight an infodemic. *The lancet*, 395(10225): 676.

Zhang, J.; Li, Z.; Liu, Q.; Wu, S.; Wang, Z.; and Wang, L. 2024. Evolving to the Future: Unseen Event Adaptive Fake News Detection on Social Media. In *CIKM*.

Zhao, T.; Liu, Y.; Neves, L.; Woodford, O.; Jiang, M.; and Shah, N. 2022. Data Augmentation for Graph Neural Networks. In *AAAI*.

Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see the Experiment, Dataset.
   (e) Did you describe the limitations of your work? Yes, with the future work.
   (f) Did you discuss any potential negative societal impacts of your work? NA
   (g) Did you discuss any potential misuse of your work? No, because no misuse.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, but the reported results are 5 times runs' average results with significance test.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? NA
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA