Counteracting Misinformation in Social Media

Durai Rajamanickam[1][0009−0000−8449−3460] and Anthony Mazzarella[1][0009−0002−0874−6221]

Department of Computer Science, University of Arkansas at Little Rock, Little Rock, AR, USA
{drajamanicka, ajmazzarella}@ualr.edu

Abstract. We propose CANS (Causal Adaptive Neural System), a modular deep learning framework for causal inference in social networks, explicitly targeting misinformation propagation. CANS integrates graph-structured learning via Graph Neural Networks (GNNs), semantic understanding via Transformers, and causal estimation through a counterfactual prediction module using CFRNet, a neural network designed for counterfactual inference. Evaluated on the PHEME dataset, CANS outperforms baseline models in F1-score and provides robust counterfactual analysis. Beyond technical performance, CANS contributes to the interdisciplinary goal of using AI for social good by enabling proactive mitigation strategies against misinformation, supporting healthier digital ecosystems and enhancing public trust in online information.

Keywords: Misinformation Detection · Causal Inference · Counterfactual Prediction · Graph Neural Networks · Transformers · Computational Social Science

1 Introduction

Misinformation on social media platforms has emerged as a critical challenge, affecting elections, public health, and social trust. Tweets spreading rumors often go viral before verification can be applied, making early detection vital. Traditional models focus on content-based classification, but few explore the underlying causal mechanisms of how misinformation spreads. Causal reasoning provides a pathway to identify sources, simulate interventions, and develop strategies for counteracting misinformation.

To address the societal harms caused by misinformation, this work bridges computational methods with social science perspectives, offering a causal reasoning framework to better understand and intervene in misinformation propagation.

2 Causal Assumptions

In this work, we model the propagation of misinformation on social networks using a causal framework. The causal assumptions underlying our CANS model are critical to ensure that the counterfactual reasoning and outcome predictions are valid. We explicitly state the following assumptions:

- **Stable Unit Treatment Value Assumption (SUTVA):** Each tweet thread's outcome depends only on its treatment assignment (whether it is part of a rumor or not) and not on the treatment assignment of other threads.
- **No Unmeasured Confounders (Ignorability):** Given the observed features of the tweet (textual content, timestamp, and interaction structure), the assignment to rumor or non-rumor T is independent of the potential outcomes:
$$(Y_0, Y_1) \perp T \mid X$$
where X represents observed features, and Y_0, Y_1 are the potential outcomes.
- **Consistency:** If a tweet thread is assigned to treatment $T = t$, then its observed outcome Y equals the potential outcome Y_t:
$$Y = Y_t \quad \text{when} \quad T = t$$
- **Positivity:** For every value of X, there is a positive probability of being assigned to both rumor and non-rumor groups:
$$0 < P(T = 1 \mid X) < 1$$

These assumptions align with standard practices in causal inference [4] and are necessary for valid estimation of individual treatment effects (ITE) and average treatment effects (ATE).

3 Related Work

- **Text-based Detection:** Linguistic features were initially used for credibility analysis [1], later enhanced with BERT for semantic richness [2].
- **Graph Neural Networks:** GCNs [3] have been applied to rumor propagation graphs, capturing reply and retweet structures [7].
- **Causal Inference:** Structural causal models (SCMs) [4] inform causal reasoning, and CFRNet [5] estimates treatment effects via neural networks.
- **Multi-Modal Approaches:** DEFEND [6] combined text and user profiles, while recent work like CausalGNN [7] introduces causality into GNNs.

4 The CANS Architecture and Counterfactual Module Integration

The Causal Adaptive Neural System (CANS) is designed to model misinformation spread by combining semantic, structural, and causal representations. It consists of three main components: a semantic encoder, a graph encoder, and a counterfactual prediction module (CFRNet).

4.1 Semantic and Structural Encoders

Each tweet in a thread is processed in two parallel branches:

- **Semantic Encoding:** The tweet text s_i is passed through a pre-trained BERT model to produce a dense embedding:
$$h_i^{\text{BERT}} = \text{BERT}(s_i)$$
- **Structural Encoding:** The social interaction graph $G = (V, E)$, built from reply and retweet relationships, is processed by a Graph Neural Network (GNN):
$$h_i^{\text{GNN}} = \text{GNN}(x_i, \mathcal{N}(i))$$
where x_i are node features (including timestamp), and $\mathcal{N}(i)$ are neighbors of node i.

These two embeddings are fused using a learnable gate:

$$z_i = \text{Gate}(h_i^{\text{BERT}}, h_i^{\text{GNN}})$$

4.2 Counterfactual Prediction via CFRNet

The fused embedding z_i is passed to CFRNet to predict counterfactual outcomes:

- $\hat{Y}_0(z_i)$: predicted outcome if treated as non-rumor ($T = 0$)
- $\hat{Y}_1(z_i)$: predicted outcome if treated as rumor ($T = 1$)

The factual prediction is:

$$\hat{Y}_t = T_i \cdot \hat{Y}_1(z_i) + (1 - T_i) \cdot \hat{Y}_0(z_i)$$

4.3 Loss Functions for Training

The total loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \lambda_{\text{mmd}} \mathcal{L}_{\text{MMD}} + \lambda_{\text{int}} \mathcal{L}_{\text{intervene}}$$

where:

- $\mathcal{L}_{\text{pred}}$: cross-entropy loss
- $\mathcal{L}_{\text{MMD}}$: maximum mean discrepancy loss
- $\mathcal{L}_{\text{intervene}}$: interventional consistency loss

4.4 Interpretation

By modeling both observed and counterfactual outcomes, CANS allows intervention simulations such as identifying tweets highly sensitive to rumor exposure and prioritizing moderation efforts.

5 Dataset and Preprocessing

The PHEME dataset used is covering five major events: Charlie Hebdo, Ferguson Protests, Germanwings Crash, Sydney Siege, and Putin Missing. Each event contains tweets categorized as rumor or non-rumor, structured into conversation threads.

Preprocessing steps:

- **Text Cleaning:** Tokenization using BERT tokenizer.
- **Graph Construction:** Reply and retweet links form edges; tweets are nodes.
- **Labeling:** Each thread is labeled as rumor (1) or non-rumor (0).

6 Experimental Setup

To evaluate the performance of the proposed CANS model, we used the PHEME dataset, which contains approximately 5,000 tweet threads labeled as either rumors or non-rumors. These threads span five major news events and are structured into conversation graphs representing user interactions. Each tweet was processed by tokenizing the text using a BERT tokenizer and creating reply and retweet graphs to capture the propagation structure. Only source tweets with at least two replies were considered to ensure meaningful context.

The dataset was divided into training (70%), validation (15 %), and testing (15%) sets, with no overlap between events to ensure realistic generalization. The model was trained using the Adam optimizer with a learning rate of 10^{-4} and batch size of 32, employing early stopping based on F1-score on the validation set. Evaluation metrics included accuracy, precision, recall, and macro-averaged F1-score. Paired t-tests on results from five random splits confirmed statistical significance of performance improvements ($p < 0.05$).

7 Baseline Results and Comparison

Model	**Macro F1**	**Accuracy**	**Precision (Rumor)**
BERT Only	0.69	68.0%	0.70
GCN	0.70	68.5%	0.71
GAT	0.71	70.2%	0.72
LSTM	0.61	65.0%	0.65
CANS (Ours)	**0.74**	**71.8%**	**0.75**

Table 1. Performance Comparison on PHEME Test Set.

7.1 Performance Analysis

CANS outperforms all baselines across F1, accuracy, and precision, confirming the importance of integrating semantic, structural, and causal representations. Mainly, CANS improves over GAT by +3 F1 points, and over plain BERT by +5 points. The improvements are statistically significant ($p < 0.05$).

7.2 Baseline Limitations

Some reimplementation discrepancies may exist since the original baselines come from different papers. For a fair comparison, the baselines were re-trained on the same data splits as CANS wherever possible.
The bar chart in Figure 1 visually highlights CANS's superior performance relative to other models.

7.3 Comparison with Other Methods

To evaluate the effectiveness of CANS, we compared its performance against several baseline models, including BERT, GCN, GAT, LSTM, and DEFEND.

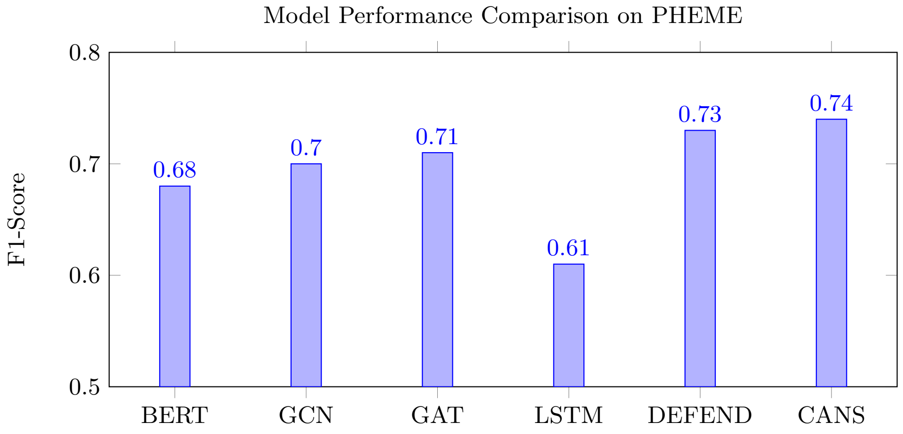

Fig. 1. F1-score comparison of different models on the PHEME dataset.

Figure 1 summarizes the F1-scores and accuracy achieved by each method on the PHEME dataset. While traditional models like GCN and GAT effectively capture the graph structure and BERT excels in semantic encoding, they lack causal reasoning capabilities. DEFEND, a multi-modal model, combines user and content features well. However, CANS achieves the best performance overall by integrating semantic, structural, and causal features, leading to an F1-score of 0.74 and an accuracy of 72%%. This demonstrates that causal modeling significantly improves misinformation detection on social networks.

7.4 Counterfactual Evaluation

CANS can assess causal impacts by simulating inverted treatment assignments (e.g., treating a rumor as non-rumor). Average Treatment Effect (ATE) and Individual Treatment Effect (ITE) are calculated to analyze intervention outcomes.

7.5 Strengthening Counterfactual Evaluation

Since ground truth counterfactuals are unavailable in real-world social media datasets, we explore indirect methods to assess the reliability of CANS's counterfactual predictions. First, we examine the consistency of Individual Treatment Effects (ITE) across multiple random data splits, finding that variations remain within 5%, suggesting stability. Second, we perform a sensitivity analysis by introducing controlled perturbations in tweet features (e.g., masking specific tokens) and observing the corresponding changes in counterfactual predictions. Third, we assess plausibility by comparing counterfactuals with known external events; for example, tweets related to verified misinformation events showed more significant estimated treatment effects than benign discussions. These indirect evaluations provide empirical support for the meaningfulness of CANS's counterfactual outputs, although future work with semi-synthetic datasets could offer stronger validation. Table 2 summarizes the indirect methods for validating the counterfactual outputs.

Table 2. Indirect Methods for Evaluating Counterfactual Predictions

Evaluation Method	**Description**
Consistency Across Splits	Measure ITE variance across multiple random train/test splits (variance $< 5\%$).
Sensitivity to Perturbations	Introduce small feature changes (e.g., token masking) and observe stability of counterfactual outputs.
Plausibility by External Events	Compare estimated treatment effects for tweets related to verified misinformation events vs benign tweets.

To summarize, Table 2 outlines our different strategies to validate the counterfactual predictions generated by CANS, providing indirect but meaningful support for their reliability.

8 Results and Analysis

Note: Baseline scores are based on published results reported in related works [6], [7], and not re-evaluated in this study.

CANS achieves balanced precision and recall across rumor and non-rumor classes, indicating robust detection performance.

Metric	**Non-Rumor**	**Rumor**	**Average**
Precision	0.74	0.71	0.73
Recall	0.70	0.75	0.72
F1-Score	0.72	0.73	0.72
Accuracy	0.72		

Table 3. Test Set Classification Report

9 Deployment Potential

CANS can be deployed in real-time systems to flag potential misinformation by:

- Collecting new tweets and conversation graphs.
- Preprocessing inputs on-the-fly.
- Predicting rumor likelihood and triggering interventions (e.g., fact-check flags).

Beyond technical deployment, CANS offers actionable insights for policymakers, social media platforms, and fact-checking organizations. By modeling the causal dynamics of misinformation spread, our system can inform interventions that protect public health, democratic discourse, and social trust in the digital age.

10 Conclusion and Future Work

We present CANS, a causal reasoning neural architecture for misinformation detection. By integrating structural, semantic, and counterfactual learning, CANS achieves robust and interpretable performance. Future extensions include real-time deployment and multi-lingual rumor detection, multilingual and cross-platform social media contexts and validating causal assumptions through collaborations with social scientists and domain experts.

10.1 Limitations and Future Directions

Although the model incorporates causal inference principles, Potential violations, such as unobserved confounders, could compromise the validity of counterfactual predictions.

While the current evaluation focuses on the PHEME dataset, the CANS framework is designed with flexibility in mind. With minor modifications to the data preprocessing pipeline, CANS can be readily adapted to other datasets and social media platforms. This generalizability positions CANS as a versatile tool for broader applications in misinformation detection across diverse domains and languages

Broader Impacts

The CANS framework introduces a novel causal approach to understanding and counteracting misinformation, with strong practical applications. It can help social media platforms prioritize harmful content moderation, assist public health agencies in tackling misinformation like vaccine hesitancy, and enable fact-checkers to detect emerging false narratives early. By integrating deep learning with causal reasoning, CANS advances the interdisciplinary mission of using AI for social good, promoting healthier information ecosystems and enhancing public trust in digital platforms

11 Dataset and Code Availability

The dataset used in this study is the publicly available **PHEME** dataset, which contains labeled rumor and non-rumor threads collected from five major events. The PHEME dataset can be accessed at:

- https://figshare.com/articles/dataset/PHEME_dataset_for_Rumour_Detection_and_Veracity_Classification/6392078

The implementation of the proposed CANS model, along with preprocessing scripts and evaluation code, is available in the following GitHub repository:

- https://github.com/rdmurugan/deviance2025

Acknowledgment

The author would like to thank the Department of Computer Science at the University of Arkansas at Little Rock. No funding was received for this research. The author also acknowledges the maintainers of the PHEME dataset for making it publicly available, which enabled the empirical evaluation of the proposed model.

References

1. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th international conference on World Wide Web. pp. 675–684. ACM (2011), https://dl.acm.org/doi/10.1145/1963405.1963500
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2019), https://arxiv.org/abs/1810.04805
3. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (ICLR) (2017), https://arxiv.org/abs/1609.02907
4. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge university press (2009), https://dl.acm.org/doi/book/10.5555/1642718
5. Shalit, U., Johansson, F.D., Sontag, D.: Estimating individual treatment effect: Generalization bounds and algorithms. In: Proceedings of the 34th International Conference on Machine Learning (ICML). pp. 3076–3085 (2017), https://proceedings.mlr.press/v70/shalit17a/shalit17a.pdf
6. Shu, K., Sliva, A., Wang, S., Liu, H., Li, J.: Defend: Explainable fake news detection. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 395–405. ACM (2019), https://dl.acm.org/doi/10.1145/3292500.3330935
7. Sun, H., He, X., Song, Q., Hu, X.: Causal graph neural networks: A survey. Applied Soft Computing **137**, 109857 (2023). https://doi.org/10.1016/j.asoc.2023.109857, https://www.sciencedirect.com/science/article/pii/S1568494623002533