Estimating Online Influence Needs Causal Modeling! Counterfactual Analysis of Misinformation Engagement on Social Media

Lin Tian and Marian-Andrei Rizoiu

Behavioral Data Science, University of Technology Sydney
15 Broadway, Ultimo, NSW 2007, Australia
{Lin.Tian-3,Marian-Andrei.Rizoiu}@uts.edu.au

Abstract

Understanding true influence in social media requires distinguishing correlation from causation—particularly when analyzing misinformation spread. While existing approaches focus on exposure metrics and network structures, they often fail to capture the causal mechanisms by which external temporal signals trigger engagement. We introduce CITRUS (Causal Influence through Treatment-Response Understanding in Social media), a novel joint treatment-outcome framework that leverages existing sequential models to understand how external signals—search trends, news coverage, influencer activity—trigger misinformation engagement. Through experiments on real-world misinformation and disinformation datasets, CITRUS outperforms existing benchmarks by 15-22% in predicting engagement across diverse counterfactual scenarios, including exposure adjustment, temporal alignment shifts, and varied intervention durations. Case studies on 492 social media users demonstrate that our causal effect measure aligns strongly with expert-based empirical influence assessments, validating CITRUS as a robust framework for understanding information spread dynamics. CITRUS also reveals that low-baseline misinformation can scale 6-fold under external promotion, showing super-linear growth, and unmasks hidden amplifiers—accounts with modest followings that double engagement rates, outperforming supposed "influencers" with 100x more followers.

Introduction

In March 2020, as COVID-19 began its global spread, a simple graphic titled "Flatten the Curve" created by Associate Professor Siouxsie Wiles and cartoonist Toby Morris ignited an unprecedented cascade of social media engagement[1]. Within just 72 hours, the visualization accumulated over 10 million impressions on X (formerly Twitter), was translated into more than 25 languages, and was adopted by government health agencies worldwide (Bavel et al. 2020). What made this particular content achieve such extraordinary reach was not merely its clear visualization of pandemic dynamics—it was the complex interplay between its timely release amid escalating global concern, amplification by influential public health accounts, and concurrent spikes in search traffic as reflected in Google Trends data (Jackson, Bailey, and Welles 2020). As the graphic's engagement metrics surged, a corresponding peak in "social distancing" searches appeared across global search indices, preceding measurable changes in mobility patterns across affected regions (Gao et al. 2020). Today, with hindsight, we say that the "Flatten the Curve" graphic was influential as it increased the public's awareness of non-pharmaceutical interventions for the pandemic.

But what is influence? It is more than exposure or virality; it is one's capacity to shape attitudes and change behaviors. And while exposure (particularly repeated exposure) and influence are intimately linked, true influence estimation requires causal reasoning. This causal perspective on influence has been established in seminal works across multiple disciplines (Aral, Muchnik, and Sundararajan 2009; Eckles, Kizilcec, and Bakshy 2016; Watts, Rothschild, and Mobius 2021), which collectively demonstrate that traditional influence measures often conflate homophily with causal effects. In online social media, information cascades equate to exposure—they are the digital equivalent of the *word-of-mouth* phenomenon that allow content to spread widely. But how do they relate to influence? This relationship requires counterfactual analysis to examine what would happen to engagement if external signals changed. This is crucial for misinformation cascades, where identifying who truly shapes public discourse on polarizing topics can inform intervention strategies.

True influence is unobserved, and difficult to estimate (Ram and Rizoiu 2024). We therefore leverage observable exogenous attention signals such as search trends, news coverage cycles, and influencer amplification, and ask how they causally impact content engagement and information diffusion. Existing approaches using content features and network structure overlook causal relationships between temporal signals and engagement (Zhou et al. 2021a). Previous influence studies focus on exposure metrics without causal perspectives (Cha et al. 2010; Bakshy et al. 2011; Kwak et al. 2010), while causal inference frameworks (Pearl 2009; Peters, Janzing, and Schölkopf 2017) lack temporal sophistication for sequential social media dynamics.

To tackle these challenges, we introduce CITRUS, a causal framework that jointly models treatment intensities (external signals driving engagement) and social engagement outcomes. CITRUS builds on advances in sequential modeling, adapting transformer architectures (Vaswani

[1]https://www.washingtonpost.com/graphics/2020/world/corona-simulator/

et al. 2017) and selective state space models (e.g., Mamba (Gu and Dao 2024)) to meet the demands of causal inference with time-varying treatments. By comparing integration mechanisms (token-based, attention-based, layer-based, and adapter-based), we present detailed insights into the architectural requirements for effective causal modeling in social media contexts, particularly in misinformation scenarios.

This work explores three key research questions: (1) Does joint modeling of treatment intensities and outcomes improve predictions under realistic policy-driven scenarios, especially those aimed at curbing misinformation? (2) How effectively do transformers and state space models capture temporal dependencies between external signals and engagement? (3) How reliably can these models predict engagement outcomes under hypothetical scenarios, such as changes in exposure rates or policy active timings?

Related Work

Our work bridges social media engagement, causal inference, and sequence modeling.

Social Media Engagement. The field evolved from content (Cheng et al. 2014) through networks (Zhao et al. 2015) to temporal dynamics (Rizoiu et al. 2017). Recent advances span attention-based cascades to cross-platform frameworks (Li et al. 2017; Ding, Wang, and Wang 2019; Wang et al. 2017a; Qiu et al. 2018; Cao et al. 2017; Kong et al. 2023; Calderon, Ram, and Rizoiu 2024). While effective at modeling cascades, these approaches lack explicit causal modeling of external drivers—our key contribution.

Causal Inference. G-methods (Robins 1986) and marginal structural models (Robins, Hernan, and Brumback 2000; Lok 2008; Schulam and Saria 2017) provide foundations. Neural approaches handle time-varying confounding (Lim 2018; Bica et al. 2020), while continuous-time methods use Gaussian processes and point processes (Soleimani, Subbaswamy, and Saria 2017; Schulam and Saria 2017; Hızlı et al. 2023). These lack the architectural flexibility needed for social media's complex dynamics.

Sequence Models. Beyond Transformers (Vaswani et al. 2017; Zhou et al. 2021b; Wu et al. 2021), State Space Models (SSMs) efficiently capture long-range dependencies (Gu et al. 2020; Dao et al. 2022; Gu and Dao 2024). Despite these advances, few studies have applied modern SSMs to social media engagement prediction in causal contexts. Existing approaches typically rely on graph-based (Lu et al. 2023), RNN (Wang et al. 2017b), or transformer methods (Zuo et al. 2020) that assume uniform sampling or discrete snapshots, overlooking the fine-grained temporal patterns essential to engagement dynamics.

Problem Formulation and Methodology

Consider a social media event $\mathcal{E}$ with associated posts $\mathcal{P} = \{p_1, \ldots, p_N\}$. Each post $p \in \mathcal{P}$ is represented by tuple (t_0, x, u, o, H), comprising posting time t_0, textual content x, user metadata u, category $o \in \mathcal{O}$, and interval-censored engagement history $H = \{(t_j, e_j)\}_{j=1}^{m}$. Here, $e_j \in \mathbb{R}^d$ captures multiple engagement types (e.g., shares, comments) at observation time t_j. External signals $G = \{(t_k, g_k)\}_{k=1}^{l}$ quantify search intensity for content-relevant keywords, serving as exogenous indicators of public attention.

Our task is to predict future engagement trajectories under counterfactual modifications to external signals. Formally, given observations within window τ_{obs}, we forecast $\{\hat{e}(t_0 + \tau_{\text{obs}} + k\Delta t)\}_{k=1}^{K}$ under policy-driven signal transformations, where $K = \lfloor T/\Delta t \rfloor$ defines the prediction horizon.

Causal Assumptions for Social Media Influence

Our framework addresses a critical question: how do external attention signals causally drive misinformation engagement on social media? We build on the potential outcomes framework (Rubin 1974) with three key assumptions tailored to social media contexts:

Assumption 1 (Consistency)**.** *The potential engagement outcome $Y[a]$ under a specific pattern of external signals a equals the observed engagement Y when those exact signals actually occur.*

This assumes observed relationships between external signals and engagement persist if patterns reoccur. In social media environments, this consistency is supported by the relative stability of engagement mechanisms—including recommendation algorithms, user interfaces, and notification systems—which function as reliable mediators between external signals and user behavior (Becker, Brackbill, and Centola 2017). Empirical evidences (Rizoiu et al. 2017; Calderon, Ram, and Rizoiu 2024) show consistent outperformance when modeling exogenous drivers, with HIP's Linear Time-Invariant property ensuring identical stimuli yield reliable responses (Muchnik, Aral, and Taylor 2013). This holds within stable timeframes, before major platform changes.

Assumption 2 (Fully-Mediated Policy Effect)**.** *External signals affect engagement outcomes through observable mechanisms rather than hidden pathways.*

The causal process follows: exogenous signals $\rightarrow$ heightened topical salience $\rightarrow$ content exposure via platform mechanisms $\rightarrow$ measurable engagement (e.g., likes, shares). Following the framework propsed by Imbens and Rubin (2015), we assume an exclusion restriction: conditional on observed mediators, external signals have no additional direct effect on engagement outcomes. This assumption is empirically plausible in our domain: external signals shift topical salience (Calderon, Ram, and Rizoiu 2024), which determines exposure (Bakshy, Messing, and Adamic 2015), and exposure governs engagement responses (Rizoiu et al. 2017). The predictive gains of exposure-based models (e.g., OMM, HIP) further indicate that unmeasured direct pathways are limited.

Assumption 3 (Temporal Precedence)**.** *Causes precede effects—current engagement can be influenced by past external signals but not by future ones.*

This assumption aligns with the natural temporal ordering of social media interactions, where content engagement follows rather than precedes external attention signals. Rizoiu et al. (2017)'s Hawkes model supports this, with external stimuli at t affecting engagement at $\tau > t$ via measurable maturity time lags, validated on YouTube data. Calderon, Ram, and Rizoiu (2024) shows similar gains treating signals as

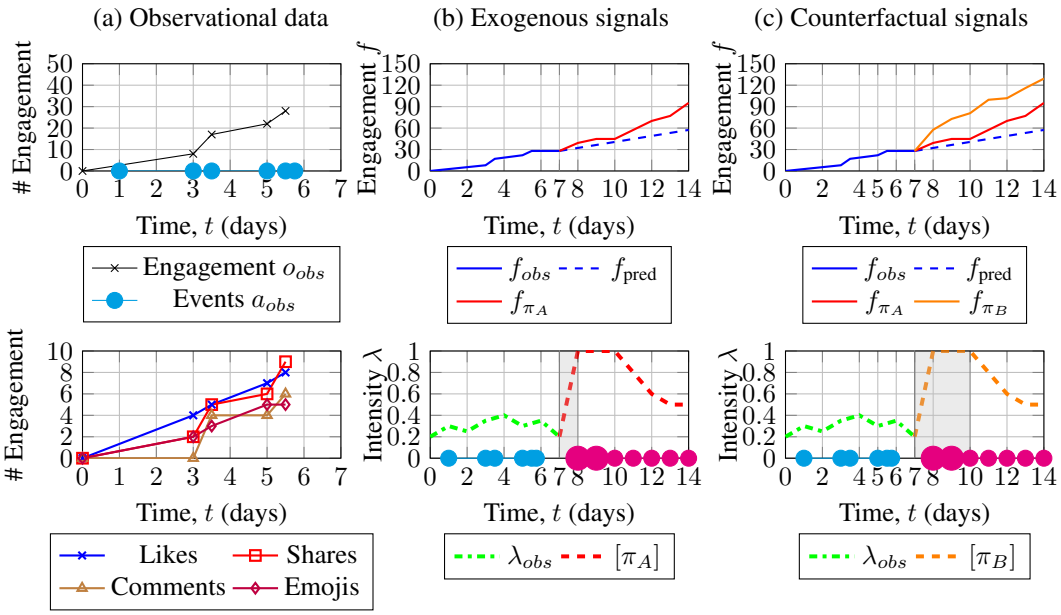

Figure 1: Visualization of engagement data and queries for social media post p. **(a) Observational data** during the period $[0, 7]$ days. The top plot shows cumulative engagement over time (black line with crosses) and observed events (cyan dots). The bottom plot displays individual engagement metrics: Likes (blue crosses), Shares (red squares), Comments (brown triangles), and Emojis (purple diamonds). **(b) Exogenous signal under policy** π_A (1-day exposure, shaded day 7–8). Top: Observed (solid blue), predicted without intervention (dashed blue), predicted under π_A (dashed red). Bottom: Normalized intensity λ_{obs} (dashed green), π_A intensity (dashed red), observed events (cyan dots), policy actions (magenta dots). Vertical line at day 7 marks intervention start. **(c) The counterfactual signal.** How the engagement trajectory of post p would have evolved if policy π_B had been applied during $[0, 7]$ with a three-day exposure time (shaded area from day 7 to 10). The top plot shows observed engagement (solid blue line), predicted engagement without intervention (dashed blue line), predicted engagement under π_A (dashed red line), and counterfactual engagement under π_B (solid orange line). The bottom plot displays normalized intensity λ_{obs}(dashed green line), counterfactual policy π_B intensity (dashed orange line), observed events (cyan dots), and policy actions (magenta dots). A vertical line at day 7 marks the intervention start.

antecedents. Platform algorithms reinforce this unidirectionality by boosting visibility after rising attention (Centola et al. 2018). These assumptions, while not exhaustively verifiable, provide a reasonable foundation for causal inference in social media settings and allow us to identify treatment effects from observational data. However, translating these theoretical foundations into actionable insights requires a framework capable of capturing the complex, sequential nature of social media dynamics—where multiple external signals interact over time to shape engagement outcomes.

CITRUS: Joint Treatment-Outcome Modeling

To operationalize these causal assumptions, CITRUS jointly models treatment intensities and engagement outcomes using sequential models (Fig. 1). By leveraging the temporal precedence assumption, CITRUS captures how past external signals accumulate and interact to influence future engagement. The joint modeling approach directly addresses the fully-mediated policy effect assumption, explicitly representing the pathways from external signals through platform mechanisms to measurable outcomes.

We model the treatment intensity $\lambda^*_\pi(t)$ as the probability of a binary treatment event occurring at discrete time step t. Let $f^*_\pi(t) = \beta_0 + g_b(t) + g^*_a(t) + g^*_o(t) + g^*_g(t)$, where β_0 is a baseline parameter, $g_b(t)$ is a time-varying baseline function, $g^*_a(t)$ and $g^*_o(t)$ capture dependence on past treatments and past engagement outcomes, respectively, and $g^*_g(t)$ encodes external signal influence. The treatment intensity is defined via a sigmoid link: $\lambda^*_\pi(t) = \sigma(f^*_\pi(t))\,, \sigma(x) = \frac{1}{1+\exp(-x)}$.

External signals (e.g., Google Trends) are aggregated using a finite causal convolution over w historical observations sampled at 10-minute intervals ($\Delta_g = 10$): $g^*_g(t) = \sum_{k=0}^{w-1} \alpha_k\, g(t - k\Delta_g)\, \mathbf{1}[t - k\Delta_g \geq 0]$, where $g(\cdot) \in [0, 100]$ denotes the external signal value and $\{\alpha_k\}$ are learnable weights (optionally normalized via a softmax). The indicator ensures that only valid past samples contribute. Given the intensity $\lambda^*_\pi(t)$, a binary treatment is drawn as $A_t \sim \text{Bernoulli}(\lambda^*_\pi(t))$.

To align external signals with engagement observations, we implement a temporal alignment mechanism. For each observation time t_j, we collect signals within a lag window:

$$G(t_j) = \{g_k \mid t_j - \tau_{\text{lag}} \leq t_k < t_j\},$$

where τ_{lag} is determined through cross-validation. The resulting feature vector $\mathbf{v}_g(t_j) = [g(t_j - \Delta_g), g(t_j -$

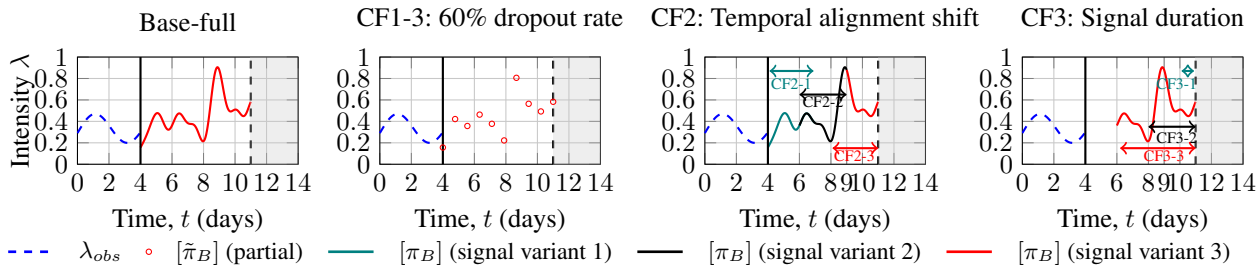

Figure 2: Counterfactual scenarios for evaluating causal effects of external signals on engagement. **Base-full**: Observed engagement trajectory (λ_{obs}, blue dashed) and full external signal pattern (red solid) over 14 days. **CF1-3**: Exposure manipulation via 60% dropout rate, simulating reduced signal visibility (red dots show sparse signals). **CF2**: Temporal alignment shifts testing intervention timing—CF2-1 (early, teal), CF2-2 (baseline, black), CF2-3 (late, red). **CF3**: Duration manipulation comparing 1-day (CF3-1), 3-day (CF3-2), and 5-day (CF3-3) sustained interventions. Vertical line at day 4 separates historical observation from treatment period; dashed line at day 11 marks prediction onset (gray region). These scenarios enable estimation of $\Delta_{\mathcal{C}} = \mathbb{E}[Y|G_{\mathcal{C}}] - \mathbb{E}[Y|G]$ under different policy interventions.

$2\Delta_g), \ldots, g(t_j - w\Delta_g)]$ captures the temporal context.

Counterfactual Analysis

To evaluate causal effects, we systematically manipulates exogenous signals along the temporal dimension to understand their causal impact on engagement dynamics (Fig. 2). Formally, a counterfactual scenario $\mathcal{C}$ as a transformation of the external signal $G = \{(t_k, g_k)\}_{k=1}^{l}$ through a temporal manipulation function Ψ_θ:

$$G_{\mathcal{C}} = \Psi_\theta(G) = \{(t_k + \delta_\theta(t_k, g_k), g_k \cdot \gamma_\theta(t_k, g_k))\}_{k=1}^{l},$$

where δ_θ shifts the timing of signals and γ_θ adjusts signal intensity, enabling three counterfactual scenarios (CF): **Exposure Manipulation** via dropout rate (CF1) adjusts intensity (γ_θ) to explore exposure effects (0%→20% as CF1-1, 20%→40% as CF1-2, 40%→60% as CF1-3); **Temporal Alignment Shifts** (CF2) the onset timing (δ_θ) of a fixed-duration signal to test sensitivity to early exposure; **Signal Duration Manipulation** (CF3) modifies persistence to evaluate sustained attention effects (1-day as CF3-1, 3-day as CF3-2, 5-day as CF3-3). For each counterfactual scenario, we estimate the expected engagement outcomes under the transformed signal and calculate the causal effect as the difference between counterfactual outcomes: $\Delta_{\mathcal{C}} = \mathbb{E}[Y|G_{\mathcal{C}}] - \mathbb{E}[Y|G]$.

Model Training and Optimization

We train CITRUS using a combined loss function $\mathcal{L}$ that explicitly optimizes both treatment intensity modeling and outcome prediction:

$$\mathcal{L} = \underbrace{\text{MSE}(Y(t), Y_{\text{pred}}(t;\theta_Y))}_{\text{Outcome Loss}} + \alpha \underbrace{\text{BCE}(A_{\text{true}}(t), \lambda_A(t;\theta_A))}_{\text{Intensity Loss}},$$

where α is a hyperparameter controlling the relative importance between accurate outcome prediction and accurate treatment intensity modeling. This joint optimization ensures that the model captures both the occurrence pattern of external signals (treatments) and their effects on engagement outcomes.

For Mamba, we add temporal coherence loss $\mathcal{L}_{\text{temp}} = \frac{1}{|\mathcal{P}|}\sum_{p,j} \|\mathbf{h}_{j+1} - \exp(\Delta t_j^+ \cdot \tilde{\mathbf{A}}_t)\mathbf{h}_j\|^2$ to enforce consistent state transitions. For Transformers, we use attention consistency loss $\mathcal{L}_{\text{att}} = \frac{1}{|\mathcal{P}|}\sum_{p,i,j,k,l} w_{ijkl}\|\mathbf{A}_{ij} - \mathbf{A}_{kl}\|^2$, where $\mathbf{A}_{ij}$ represents attention weights from position i to j and w_{ijkl} is a similarity weight based on temporal distance. These architecture-specific regularizations handle irregular temporal sampling in social media data.

Experiment and Results

This section introduces the datasets, models and baselines, and report our experimental findings on the optimal architecture choice for exogenously-driven engagement modeling.

Datasets

Social Media Data. We use two misinformation/disinformation datasets: (1) `SocialSense` (Kong et al. 2022): 1,035,302 Facebook posts (2019-2021) across four domains—Australian Bushfires (78,030), Climate Change (138,278), Vaccination (178,894), and COVID-19 (640,100)—labeled by experts as misinformation/conspiracy content; (2) `DiN` (Tian et al. 2025): 746,653 posts from 41 coordinated information operation accounts (2019-2024) classified into 9 narratives. Both datasets include engagement metrics (e.g. likes, shares) collected via CrowdTangle API[2].

External Signals. We collected Google Trends data to provide external signals aligned with both the temporal and thematic scope of the data sets. Keywords—selected via frequency analysis of post content and expert consultation—were theme-specific for SocialSense (e.g., "bushfire evacuation," "climate hoax") and narrative-specific for DiN (e.g., "election fraud," "deep state"). Search intensities were collected at 10-minute intervals, normalized to 0-100, forming signal timeline $G = \{(t_k, g_k)\}_{k=1}^{l}$ aligned with posts using lag window τ_{lag} via Google Trends API [3].

Models and Baselines

We evaluate four integration strategies for incorporating external signals, applied to both Transformer (Vaswani et al. 2017)

[2]https://www.crowdtangle.com/ before August 2024.

[3]https://serpapi.com/google-trends-api

Model	Base	Scenario 1: Exposure			Scenario 2: Timing			Scenario 3: Duration		
		CF1-1	CF1-2	CF1-3	CF2-1	CF2-2	CF2-3	CF3-1	CF3-2	CF3-3
T	0.19*	–/–	–/–	–/–	–/–	–/–	–/–	–/–	–/–	–/–
M	0.19*	–/–	–/–	–/–	–/–	–/–	–/–	–/–	–/–	–/–
MBPP	0.19	0.33/–	0.30/–	0.28/–	0.23/–	0.23/–	0.23/–	0.33/–	0.32/–	0.28/–
T+Tok	0.13	0.21/0.54	0.22/0.47	0.23/0.41	0.21/0.48	0.20/0.43	0.20/0.41	0.24/0.54	0.25/0.48	0.25/0.46
T+Att	0.12	0.20/0.51	0.19/0.42	0.18/0.38	0.19/0.45	0.19/0.39	0.18/0.35	**0.22**‡/0.53	0.20/0.47	**0.19**/0.45
T+L	0.13	0.20/0.52	0.19/0.43	0.19/0.39	0.19/0.46	0.19/0.40	0.19/0.37	0.23/**0.51**†	0.20/**0.43**‡	0.20/**0.41**‡
T+Apt	0.12	**0.19**‡/**0.49**‡	**0.18**‡/**0.40**‡	0.18/0.36	**0.18**‡/0.43	0.18/**0.37**‡	0.18/**0.34**‡	**0.22**‡/0.54	**0.19**‡/0.48	**0.19**‡/0.46
M+Tok	0.13	0.21/0.53	0.20/0.46	0.19/0.40	0.20/0.47	0.20/0.42	0.19/0.40	0.25/0.54	0.23/0.48	0.22/0.46
M+S	0.12	0.19/0.50	0.19/0.41	0.18/0.37	**0.18**‡/0.44	0.18/0.38	0.18/0.35	0.24/0.54	0.22/0.47	0.21/0.45
M+L	0.12	0.20/0.51	0.19/0.42	0.18/0.37	0.19/0.45	**0.15**‡/0.39	0.17/0.36	0.24/0.54	0.22/0.47	0.22/0.45
M+Apt	**0.11**	**0.19**‡/**0.49**‡	**0.18**‡/**0.40**‡	**0.17**‡/**0.35**‡	**0.18**‡/**0.42**‡	0.18/**0.37**‡	**0.17**‡/**0.34**‡	0.24/0.53	0.22/0.47	0.21/0.44

Table 1: Root Mean Squared Error (RMSE) and Binary Cross Entropy (BCE) of engagement prediction and treatment intensity modeling over a 7-day horizon under counterfactual scenarios. Values are means across 5 runs with different seeds on both datasets. Each cell shows RMSE/BCE format. Lower values indicate better performance. Bold values highlight the best performance per scenario for both metrics. Vanilla Transformer (T) and Mamba (M) serve as reference baselines but cannot perform counterfactual analysis as they lack external signal integration (marked with '*' for base case, '–' for scenarios). MBPP processes external signals but lacks treatment intensity modeling (BCE shown as '–'). Statistical significance tested via paired t-tests comparing each model against MBPP for RMSE and against T+Tok for BCE: † $p< 0.05$, ‡ $p< 0.01$. Model abbreviations: T = Transformer, M = Mamba, Tok = Token, Att = Attention, L = Layer, Apt = Adapter, S = Selection.

and Mamba (Gu and Dao 2024): (1) **Token** jointly embeds signals with engagement data; (2) **Attention/Selection** uses architecture-specific temporal modeling—attention heads for Transformers, selective scan conditioning for Mamba; (3) **Layer** processes signals through MLPs with cross-attention or state injection; (4) **Adapter** adds parameter-efficient modules (Houlsby et al. 2019) to condition on signal intensities. Baselines include vanilla Transformer, Mamba (Gu and Dao 2024), and MBPP (Rizoiu et al. 2022).

Results and Analysis

Evaluation Framework. Following Hızlı et al. (2023), we use a semi-synthetic evaluation framework to assess counterfactual predictions. Since true counterfactual outcomes are inherently unobservable, we use oracle models as ground truth generators. We train these oracles on held-out observational data to learn causal relationships between external signals and engagement outcomes. The oracles then serve as counterfactual generators—when we modify inputs (e.g., reduce exposure by 40%), they generate corresponding counterfactual outputs. We evaluate our model by comparing its predictions against these oracle-generated outcomes rather than factual outcomes. Robustness checks and per-dataset results are provided in the Appendix (Tian and Rizoiu 2025).

Specifically, we evaluate predictive accuracy using RMSE computed against an oracle model that provides both factual and counterfactual trajectories. For any scenario s (with $s = 0$ denoting the factual case), we compute $\mathrm{RMSE}^{(s)} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\hat{y}^{(s)}(t) - y^{\mathrm{oracle},(s)}(t)\right)^2}$. This evaluates the model's ability to recover the oracle's causal response under both observed and "what-if" interventions, without requiring true counterfactuals.

RQ1: Joint Modeling and Predictive Accuracy. Table 1 shows *Mamba+Adapter* achieves the lowest baseline RMSE (0.11). Adapter-based integration proves robust across both architectures, outperforming Token, Attention/Selection, and Layer variants, suggesting parameter-efficient adaptation best incorporates external causal signals.

Counterfactual scenarios reveal differential model robustness. As exposure increases (CF1), all models show improved performance with RMSE decreasing from 0.19-0.21 (20% exposure) to 0.17-0.19 (60% exposure), indicating that stronger external signals create more predictable engagement patterns. Temporal alignment shifts (CF2) show minimal performance variation across models, suggesting temporal displacement is well-handled by all architectures. However, signal duration extension (CF3) poses the greatest challenge, with all models showing elevated RMSE (0.22-0.25), indicating difficulty in modeling sustained interventions. Cross-dataset comparison shows disinformation (DiN) is harder to predict than misinformation, showing both higher baseline RMSE and steeper increases at 60% exposure.

RQ2: Architectural Differences in Temporal Causal Dependency Modeling. As shown in Table 1, *Mamba+Adapter* outperforms other architectures in capturing treatment dynamics, achieving the lowest BCE (0.35 at 60% exposure) with a 10.9% improvement over the average Transformer variant. An inverse relationship exists between treatment in-

Model	Exposure			Timing			Duration		
	CF1-1	CF1-2	CF1-3	CF2-1	CF2-2	CF2-3	CF3-1	CF3-2	CF3-3
T+Tok	0.17 ± 0.11	0.33 ± 0.18	0.49 ± 0.10	0.15 ± 0.08	0.19 ± 0.06	0.22 ± 0.05	0.16 ± 0.05	0.30 ± 0.09	0.43 ± 0.10
T+Att	0.18 ± 0.05	0.35 ± 0.08	0.52 ± 0.11	0.16 ± 0.08	0.20 ± 0.07	0.25 ± 0.06	0.17 ± 0.05	0.31 ± 0.09	0.46 ± 0.11
T+L	0.16 ± 0.05	0.30 ± 0.08	0.44 ± 0.09	0.13 ± 0.07	0.17 ± 0.06	0.20 ± 0.05	0.15 ± 0.05	0.27 ± 0.08	0.39 ± 0.09
T+Apt	0.19 ± 0.05	0.36 ± 0.09	0.55 ± 0.12	0.17 ± 0.09	0.22 ± 0.07	0.26 ± 0.06	0.18 ± 0.06	0.33 ± 0.10	0.48 ± 0.12
M+Tok	0.19 ± 0.05	0.38 ± 0.09	0.57 ± 0.12	0.18 ± 0.09	0.23 ± 0.08	0.27 ± 0.07	0.19 ± 0.06	0.35 ± 0.10	0.50 ± 0.12
M+S	$\mathbf{0.22 \pm 0.06}$	$\mathbf{0.43 \pm 0.10}$	$\mathbf{0.64 \pm 0.14}$	$\mathbf{0.20 \pm 0.10}$	$\mathbf{0.26 \pm 0.08}$	$\mathbf{0.30 \pm 0.07}$	$\mathbf{0.22 \pm 0.07}$	$\mathbf{0.39 \pm 0.11}$	$\mathbf{0.56 \pm 0.13}$
M+L	0.19 ± 0.05	0.37 ± 0.09	0.56 ± 0.12	0.17 ± 0.09	0.22 ± 0.07	0.26 ± 0.06	0.19 ± 0.06	0.34 ± 0.10	0.49 ± 0.12
M+Apt	0.21 ± 0.06	0.41 ± 0.10	0.61 ± 0.13	0.18 ± 0.10	0.25 ± 0.08	0.29 ± 0.07	0.21 ± 0.06	0.38 ± 0.11	0.54 ± 0.13

Table 2: Average Treatment Effect (ATE), computed via G-computation (Robins 1986), quantifies causal impacts over a 7-day horizon. Values represent means across 7 runs with different seeds on both datasets, with 95% bootstrap confidence intervals ($\pm$), normalized across four engagement metrics (likes, comments, emojis, shares). Higher ATE indicates stronger effects. Model abbreviations: T = Transformer, M = Mamba, Tok = Token, Att = Attention, L = Layer, Apt = Adapter, S = Selection.

tensity and modeling difficulty: BCE decreases from 0.49 to 0.35 as exposure rises from 20% to 60%. This suggests that high-intensity external signals may be more consistent with predictable patterns, offering potential for early detection of viral content trends. Temporal misalignment further increases architectural differences. For 5-day early interventions, *Mamba+Adapter* demonstrates a 20.2% BCE advantage over *Transformer+Token*, showing the strength of state space models in temporally shifted causal processes.

RQ3: Counterfactual Robustness and Reliability. Table 2 displays how external signals causally drive engagement under different intervention scenarios. *Mamba+Selection* shows strong causal effect estimation, achieving the highest ATE (0.64 ± 0.14) at maximum exposure (CF1-3: $40\% \rightarrow 60\%$), a 4.9% improvement over *Mamba+Adapter*.

Exposure manipulations (CF1) uncover non-linear causal dynamics. ATE increases accelerate with intensity: 94.1% growth from CF1-1$\rightarrow$CF1-2 ($0.17 \rightarrow 0.33$ average) versus 65.7% from CF1-2$\rightarrow$CF1-3 ($0.33 \rightarrow 0.55$), showing diminishing but still substantial returns with higher exposure. Temporal alignment shift interventions (CF2) show that early action increases causal effects. The average ATE ratio between 5-day early (CF2-3) and 1-day early (CF2-1) interventions is 1.47 across all models, with *Mamba+Selection* showing the strongest temporal sensitivity ($0.20 \rightarrow 0.30$, 50% increase). Signal duration analysis (CF3) confirm sustained interventions result in compounding effects: 5-day campaigns (CF3-3) generate 2.5x the impact of 1-day bursts (CF3-1) on average.

Confidence intervals indicate *Mamba+Selection* maintains the most reliable estimates despite higher ATEs, with tighter bounds relative to effect sizes. This combination of strong causal effects and estimation stability makes *Mamba+Selection* optimal for identifying influential sources, though *Mamba+Adapter* offers an alternative with marginally lower but more consistent performance across scenarios.

Case Study: Identifying Causal Influencers

The Influence Paradox: Why Followers Don't Equal Impact. Who truly drives misinformation spread on social media? Conventional wisdom equates influence with follower count, yet our causal analysis identifies a critical disconnect. Using CITRUS with Mamba integration (M+Apt), we estimate the *causal influence* of sources by treating their posting behavior as interventions and measuring engagement effects.

Causal effect is a tighter approximation of influence. We compare our causal effect measure against the gold standard—*empirical influence* derived from expert human assessments with 492 users in the anti-climate discourse on X/Twitter (Ram and Rizoiu 2024). Fig. 3 shows pairs of Spearman correlation between the gold standard empirical influence and two approximations: the follower count and our proposed causal effect influence. Visibly, the causal effect is better estimation for empirical influence ($\rho = 0.57$) than follower counts ($\rho = 0.49$), particularly visible for the top 10% most influential users. Further analysis using Kendall's W (Kendall and Smith 1939) confirms this stronger rank agreement between causal effect and empirical influence ($W = 0.70$) compared to followers ($W = 0.67$). The Concordance Correlation Coefficient reveals that while follower count fails entirely to capture the magnitude of influence ($CCC = 0.00$), causal effect maintains some concordance ($CCC = 0.21$). In particular, the two approximation measures themselves show minimal agreement ($\rho = 0.32$, $W = 0.21$, $CCC = 0.01$), suggesting they capture fundamentally different aspects of influence. These results challenge the common assumption that account popularity (follower count) is a reliable approximation for true influence, demonstrating CITRUS's causal effect as an estimate both in ranking and scale.

The Amplification Dynamics of Misinformation. Our analysis of climate change misinformation shows how different narratives respond to external amplification, in Fig. 4. For each opinion, we calculate a weighted composite engagement score: $E(o) = \hat{e}_{\text{likes}}(o) + \hat{e}_{\text{shares}}(o) + 3 \times \hat{e}_{\text{comments}}(o) + \hat{e}_{\text{emoji}}(o)$, then normalize these scores using percentile rank-

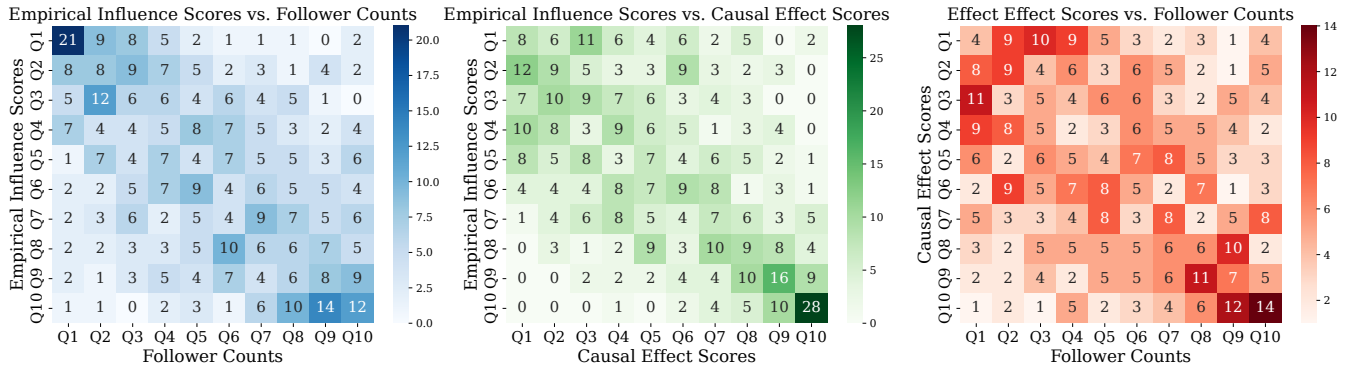

Figure 3: Decile heatmaps comparing influence measures with Spearman correlation (ρ), Kendall's rank agreement (W), and Concordance Correlation Coefficient (CCC). **Left:** Follower counts vs. empirical influence ($\rho = 0.49$, $W = 0.67$, CCC $= 0.00$). **Center:** Causal effect vs. empirical influence ($\rho = 0.57$, $W = 0.70$, CCC $= 0.21$). **Right:** Follower counts vs. causal effect ($\rho = 0.32$, $W = 0.21$, CCC $= 0.01$).

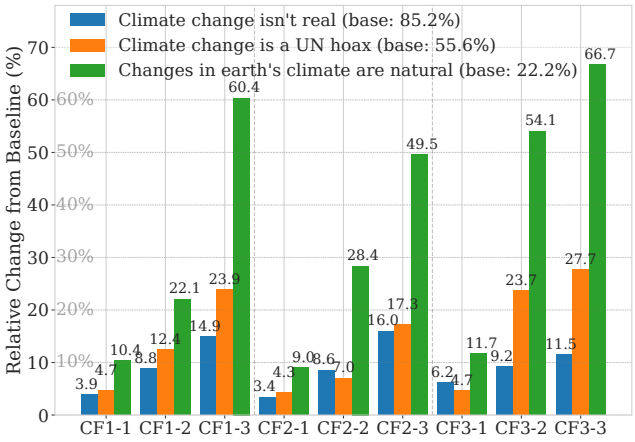

Figure 4: Relative engagement changes for climate misinformation narratives with varying baseline popularity (shown in parentheses) under counterfactual scenarios: exposure manipulation (CF1), temporal alignment shifts (CF2), and signal duration manipulation (CF3). Low-baseline narratives exhibit super-linear amplification up to 66.7%.

ing: $P(o) = \frac{\text{number of opinions with } E \leq E(o)}{\text{total number of opinions}}$. We track three narratives with varying baseline popularity: established ("Climate change isn't real", 85.2th percentile), emerging ("UN hoax", 55.6th), and fringe ("natural changes", 22.2th). Under progressive exposure manipulation (CF1), the lowest-baseline narrative shows super-linear amplification with a 6-fold scaling coefficient (10.4%→22.1%→60.4%), confirming that small external signals can trigger massive engagement increases. This behavior validates high-potential, low-baseline content theory (Rizoiu and Xie 2017), showing that emerging misinformation operates similarly to sleeping beauty content with high potential that have yet to achieve widespread attention but can rapidly increase under external promotion. The intermediate narrative shows moderate scaling effects (4.7%→12.4%→23.9%), consistent with non-linear threshold dynamics in social influence, as evidenced by tipping point experiments (Centola et al. 2018). Temporal alignment shifts (CF2) prove that initiating promotion one day earlier gets the highest impact across all narrative types (16.0%, 17.3%, and 28.4%, respectively). This finding aligns with prior work on optimal promotion timing (Rizoiu and Xie 2017) and supports the broader framework of social acceleration (Rosa 2013). Extended-duration experiments (CF3) show that a sustained 5-day exposure maximizes narrative spread, with the emerging narrative reaching a 66.7% increase. This result corroborates the mere exposure effect (Zajonc 1968) in information diffusion, indicating that temporal persistence, rather than intensity alone, drives engagement escalation.

Unmasking the Amplifiers. CITRUS further identifies influential public groups that amplified misinformation. For instance, @AustraliansforSafeTechnology increased the engagement score for "5G/smart tech is unsafe" narratives from 0.51 to 0.63 in February 2020, while @ClimateChangeBattleRoyale elevated "Climate change crisis isn't real" content from 0.11 to 0.23 in September 2019.

Conclusion

We investigate how external signals drive social media engagement in misinformation spread using CITRUS, a joint treatment-outcome framework that adapts Transformers and Mamba to improve engagement predictions under policy interventions. CITRUS faces inherent causal challenges: unobserved algorithmic confounding, selection bias, and network interference effects. Additional limitations include dependence on scarce high-quality signals, sensitivity to platform changes, and temporal stability assumptions. Despite these constraints, validation against expert-annotated data confirms CITRUS provides actionable insights where traditional metrics fail. By moving beyond correlation to causation, CITRUS enables evidence-based interventions and offers platforms a reliable metric to identify who truly spreads misinformation: not just high-follower accounts but the hidden catalysts of harmful content spread.

Acknowledgments

This research was supported by the Advanced Strategic Capabilities Accelerator (ASCA), the Defence Science and Technology Group, the Defence Innovation Network, and the Australian Academy of Science.

References

Aral, S.; Muchnik, L.; and Sundararajan, A. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51): 21544–21549.

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 65–74.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239): 1130–1132.

Bavel, J. J. V.; Baicker, K.; Boggio, P. S.; Capraro, V.; Cichocka, A.; Cikara, M.; Crockett, M. J.; Crum, A. J.; Douglas, K. M.; Druckman, J. N.; et al. 2020. Using social and behavioural science to support COVID-19 pandemic response. *Nature human behaviour*, 4(5): 460–471.

Becker, J.; Brackbill, D.; and Centola, D. 2017. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the national academy of sciences*, 114(26): E5070–E5076.

Bica, I.; Alaa, A. M.; Jordon, J.; and van der Schaar, M. 2020. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*.

Calderon, P.; Ram, R.; and Rizoiu, M.-A. 2024. Opinion market model: stemming far-right opinion spread using positive interventions. In *Proceedings of the international AAAI conference on web and social media*, volume 18, 177–190.

Cao, Q.; Shen, H.; Cen, K.; Ouyang, W.; and Cheng, X. 2017. Deephawkes: Bridging the gap between prediction and understanding of information cascades. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1149–1158.

Centola, D.; Becker, J.; Brackbill, D.; and Baronchelli, A. 2018. Experimental evidence for tipping points in social convention. *Science*, 360(6393): 1116–1119.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. 2010. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the international AAAI conference on web and social media*, volume 4, 10–17.

Cheng, J.; Adamic, L.; Dow, P. A.; Kleinberg, J. M.; and Leskovec, J. 2014. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, 925–936.

Dao, T.; Fu, D.; Ermon, S.; Rudra, A.; and Ré, C. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35: 16344–16359.

Ding, K.; Wang, R.; and Wang, S. 2019. Social media popularity prediction: A multiple feature fusion approach with deep neural networks. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2682–2686.

Eckles, D.; Kizilcec, R. F.; and Bakshy, E. 2016. Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences*, 113(27): 7316–7322.

Gao, S.; Rao, J.; Kang, Y.; Liang, Y.; Kruse, J.; Dopfer, D.; Sethi, A. K.; Reyes, J. F. M.; Yandell, B. S.; and Patz, J. A. 2020. Association of mobile phone location data indications of travel and stay-at-home mandates with COVID-19 infection rates in the US. *JAMA network open*, 3(9): e2020485–e2020485.

Gu, A.; and Dao, T. 2024. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In *First Conference on Language Modeling*.

Gu, A.; Dao, T.; Ermon, S.; Rudra, A.; and Ré, C. 2020. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems*, 33: 1474–1487.

Hızlı, Ç.; John, S.; Juuti, A. T.; Saarinen, T. T.; Pietiläinen, K. H.; and Marttinen, P. 2023. Causal modeling of policy interventions from treatment-outcome sequences. In *International Conference on Machine Learning*, 13050–13084. PMLR.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2790–2799. PMLR.

Imbens, G. W.; and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Jackson, S. J.; Bailey, M.; and Welles, B. F. 2020. *# HashtagActivism: Networks of race and gender justice*. Mit Press.

Kendall, M. G.; and Smith, B. B. 1939. The problem of m rankings. *The annals of mathematical statistics*, 10(3): 275–287.

Kong, Q.; Booth, E.; Bailo, F.; Johns, A.; and Rizoiu, M.-A. 2022. Slipping to the Extreme: A Mixed Method to Explain How Extreme Opinions Infiltrate Online Discussions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 524–535.

Kong, Q.; Calderon, P.; Ram, R.; Boichak, O.; and Rizoiu, M.-A. 2023. Interval-censored transformer hawkes: Detecting information operations using the reaction of social systems. In *Proceedings of the ACM web conference 2023*, 1813–1821.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, 591–600.

Li, C.; Ma, J.; Guo, X.; and Mei, Q. 2017. Deepcas: An end-to-end predictor of information cascades. In *Proceedings of the 26th international conference on World Wide Web*, 577–586.

Lim, B. 2018. Forecasting treatment responses over time using recurrent marginal structural networks. *Advances in neural information processing systems*, 31.

Lok, J. J. 2008. Statistical Modeling of Causal Effects in Continuous Time. *The Annals of Statistics*, 1464–1507.

Lu, X.; Ji, S.; Yu, L.; Sun, L.; Du, B.; and Zhu, T. 2023. Continuous-time graph learning for cascade popularity prediction. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2224–2232.

Muchnik, L.; Aral, S.; and Taylor, S. J. 2013. Social influence bias: A randomized experiment. *Science*, 341(6146): 647–651.

Pearl, J. 2009. *Causality*. Cambridge university press.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Qiu, J.; Tang, J.; Ma, H.; Dong, Y.; Wang, K.; and Tang, J. 2018. Deepinf: Social influence prediction with deep learning. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2110–2119.

Ram, R.; and Rizoiu, M.-A. 2024. Empirically measuring online social influence. *EPJ Data Science*, 13(1): 53.

Rizoiu, M.-A.; Soen, A.; Li, S.; Calderon, P.; Dong, L. J.; Menon, A. K.; and Xie, L. 2022. Interval-censored Hawkes processes. *Journal of Machine Learning Research*, 23(338): 1–84.

Rizoiu, M.-A.; Xie, L.; Sanner, S.; Cebrian, M.; Yu, H.; and Van Hentenryck, P. 2017. Expecting to be hip: Hawkes intensity processes for social media popularity. In *Proceedings of the 26th international conference on world wide web*, 735–744.

Rizoiu, M.-A.; and Xie, L. X. 2017. Online popularity under promotion: Viral potential, forecasting, and the economics of time. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 182–191.

Robins, J. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12): 1393–1512.

Robins, J. M.; Hernan, M. A.; and Brumback, B. 2000. Marginal structural models and causal inference in epidemiology.

Rosa, H. 2013. *Social acceleration: A new theory of modernity*. Columbia University Press.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5): 688.

Schulam, P.; and Saria, S. 2017. Reliable decision support using counterfactual models. *Advances in neural information processing systems*, 30.

Soleimani, H.; Subbaswamy, A.; and Saria, S. 2017. Treatment-response models for counterfactual reasoning with continuous-time, continuous-valued interventions. In *33rd Conference on Uncertainty in Artificial Intelligence, UAI 2017*.

Tian, L.; Booth, E.; Bailo, F.; Droogan, J.; and Rizoiu, M.-A. 2025. Before It's Too Late: A State Space Model for the Early Prediction of Misinformation and Disinformation Engagement. In *Proceedings of the ACM on Web Conference 2025*, 5244–5254.

Tian, L.; and Rizoiu, M.-A. 2025. Estimating Online Influence Needs Causal Modeling! Counterfactual Analysis of Social Media Engagement. arXiv:2505.19355.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, J.; Zheng, V. W.; Liu, Z.; and Chang, K. C.-C. 2017a. Topological recurrent neural network for diffusion prediction. In *2017 IEEE international conference on data mining (ICDM)*, 475–484. IEEE.

Wang, Y.; Shen, H.; Liu, S.; Gao, J.; and Cheng, X. 2017b. Cascade Dynamics Modeling with Attention-based Recurrent Neural Network. In *IJCAI*, volume 17, 2985–2991.

Watts, D. J.; Rothschild, D. M.; and Mobius, M. 2021. Measuring the news and its impact on democracy. *Proceedings of the National Academy of Sciences*, 118(15): e1912443118.

Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34: 22419–22430.

Zajonc, R. B. 1968. Attitudinal effects of mere exposure. *Journal of personality and social psychology*, 9(2p2): 1.

Zhao, Q.; Erdogdu, M. A.; He, H. Y.; Rajaraman, A.; and Leskovec, J. 2015. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 1513–1522.

Zhou, F.; Xu, X.; Trajcevski, G.; and Zhang, K. 2021a. A survey of information cascade analysis: Models, predictions, and recent advances. *ACM Computing Surveys (CSUR)*, 54(2): 1–36.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021b. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 11106–11115.

Zuo, S.; Jiang, H.; Li, Z.; Zhao, T.; and Zha, H. 2020. Transformer hawkes process. In *International conference on machine learning*, 11692–11702. PMLR.