Evaluating the Effectiveness of Cyber Cognitive Attacks: A Sentiment-Based Approach

Bonnie Rushing, University of Colorado Colorado Springs, United States
William Hersch, United States Air Force Academy, United States
Kora Gwartney, University of Colorado Colorado Springs, United States
Shouhuai Xu, University of Colorado Colorado Springs, United States

The Washington DC Conference on the Social Sciences 2026
Official Conference Proceedings

Abstract

Cyber cognitive attacks are a growing threat, yet their effectiveness remains difficult to measure systematically. We introduce the Cyber Cognitive Attack Effects Chain—Resonance (affective engagement), Proliferation (spread), and Influence (impact)—and propose a lightweight, data-driven method linking sentiment, topic, and engagement metrics. Resonance is operationalized via headline sentiment, while proliferation is captured through engagement percentiles, enabling tail-aware comparisons of narrative spread. Applying this framework to the FakeNewsNet–PolitiFact dataset (1,056 articles), we find that false narratives are more emotionally polarized, skew more negative, and achieve substantially higher median engagement than real news, indicating broader and more consistent virality. Topic–sentiment interactions reveal high-risk pockets (notably false–health and false–celebrity) that amplify spread, while real political content dominates aggregate engagement through a small number of highly amplified items. These patterns align with DISARM execution tactics, providing quantitative support for Maximise Exposure (TA17) and Deliver Content (TA09). The framework generalizes across datasets and offers practical implications for detection and defense, including sentiment-first triage and topic-aware monitoring of high-variance narratives. We conclude by outlining limitations and directions for extending measurement to downstream influence.

Keywords: social media, affective computing, security, cognitive attacks, disinformation

Introduction

Emotion-laden falsehoods now travel faster and farther than truth on social platforms, creating real harms in public health, politics, and online safety. Large empirical studies show that false news propagates "farther, faster, deeper, and more broadly" than true news, with affectively charged content, especially anger and fear, driving shares and attention (Cervone et al., 2024; Duch, 2021; Vosoughi et al., 2018). Adversaries exploit this psychology with coordinated amplification, bots, and recommender-system dynamics, tactics cataloged by DISARM under goals to maximise exposure and drive online harms (Bontridder & Poullet, 2021; DISARM Foundation, 2024). The consequences are concrete: health misinformation distorts risk perception and behavior (Wang et al., 2019), mainstream platforms can algorithmically surface false vaccine narratives (Juneja & Mitra, 2021), and politicized narratives reshape online discourse and mobilization (Stieglitz & Dang-Xuan, 2013). Framed as "cognitive warfare," these campaigns intentionally target attitudes, decisions, and collective cognition (Claverie & du Cluzel, 2024; Rushing & Xu, 2026).

Yet despite growing doctrine and anecdotal evidence, the field lacks standardized, data-driven measures that tie emotional framing to observed spread.

We address this gap with a measurable *effects chain* and lightweight metrics linking sentiment to virality. Our three-phase *Cyber Cognitive Attack Effects Chain* includes *Resonance* → *Proliferation* → *Influence*, and we quantify how sentiment and topical framing relate to spread. Using tweet counts as a behavioral proxy for proliferation, we test whether emotionally extreme (especially negative) language is associated with higher virality and whether certain topic–sentiment combinations are especially potent.

Specifically, we study: (i) how sentiment differs between false and real narratives; (ii) how sentiment correlates with virality; (iii) which topic–sentiment pairs yield the greatest spread; and (iv) how these observables map to DISARM Execution (PO3) tactics. In this paper, we operationalize *Resonance* and *Proliferation*; *Influence* (belief/behavior change) requires longitudinal or off-platform outcomes and is out of scope.

Contributions

We make two contributions. (1) A reusable methodology anchored in a three-phase *effects chain*, *Resonance* (sentiment), *Proliferation* (engagement), *Influence* (impact), with simple, cross-platform measurements. (2) An empirical case study on FakeNewsNet–PolitiFact[1] that quantifies sentiment and topic effects on spread, identifies high-risk topic–sentiment pockets, and proposes a concise *Resonance–Proliferation Score* (RPS) for triage.[2]

Background and Related Work

Research relevant to the effectiveness of cyber cognitive attacks generally falls into four interconnected areas.

[1]https://github.com/KaiDMML/FakeNewsNet
[2]Video summary available at https://www.youtube.com/watch?v=Xwf_GZ-BRB0

Influence Campaign Lifecycle Models

Influence campaign models, such as cyber kill chains and DISARM (DISARM Foundation, 2024), offer structured ways to categorize cyber cognitive operations through phases defined by specific tactics, techniques, and procedures (TTPs). However, these frameworks often lack quantitative support using real-world engagement data and do not formalize the influence process. To address this gap, we define a novel three-phase *Cyber Cognitive Attack Effects Chain*, composed of *Resonance*, *Proliferation*, and *Influence*. This effects chain models how emotionally charged content hooks audiences (Resonance), spreads through digital ecosystems (Proliferation), and alters perception or behavior (Influence). Our approach draws conceptually from the "ABCs of Influence" model by François (2019), which identifies core influence goals: Affecting attitudes (A), influencing Behavior (B), and shaping Cognition (C). We operationalize these goals using sentiment analysis and engagement metrics, quantifying affect through emotional polarity, behavior through retweet volume, and cognitive impact through sustained narrative engagement. By bridging conceptual frameworks like DISARM and the ABCs with empirical measures, our methodology enables scalable, cross-platform assessment of cognitive attack effectiveness.

Narrative Virality and Emotional Engagement

Studies in narrative virality and memetics emphasize the role of emotional resonance in driving the online propagation of content. Previous work is largely qualitative and highlights that narratives leveraging strong emotional cues, particularly negative emotions like fear, anger, or outrage, demonstrate increased viral potential and deeper audience engagement (Cervone et al., 2024; Duch, 2021; Hutto & Gilbert, 2014). Krishnan and Sitaraman (2013) investigated the effectiveness of disinformation through a large-scale measurement study on the effectiveness of video ads. Their findings suggest that deeper interactions, such as sustained attention or click-through behavior (e.g., retweets), are the preferred indicators of user engagement and intent, aligning with our use of tweets and retweets as a proxy for cognitive attack effectiveness.

Platform Manipulation and Exposure Maximization

The literature on platform manipulation examines how malicious actors exploit online platforms to maximize the reach and persistence of cyber cognitive attacks. Common techniques include coordinated posting, amplification via bot networks, AI use, hashtag flooding, and exploitation of algorithmic recommendation systems (Caramancion, 2023; DISARM Foundation, 2024; Juneja & Mitra, 2021). While these manipulative tactics have been well-documented conceptually, empirical studies using systematic datasets remain relatively sparse. Our work expands this literature by empirically assessing how such manipulation aligns with emotional sentiment and topic framing to enhance narrative effectiveness.

FakeNewsNet and PolitiFact

We use the *FakeNewsNet* PolitiFact dataset, which labels news as *real* or *fake* based on professional fact-checks by the nonpartisan organization PolitiFact.[3] Unlike other datasets (e.g., *BuzzFeedNews*, *FacebookHoax*, *BS Detector*), FakeNewsNet includes rich features such as social context, temporal metadata, and tweet-level diffusion patterns (Shu et al., 2018). Each

[3]https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/

article is linked to corresponding tweet IDs, enabling analysis of narrative spread on social media. Prior studies have used this dataset for false narrative detection, linguistic analysis, and modeling user credibility (Shu et al., 2018; Su et al., 2023a; Su et al., 2023b). We extend this work by applying sentiment and topic analysis to examine how emotional framing and thematic focus affect the spread, and thus the effectiveness, of real versus false narratives on Twitter.

By integrating and expanding on these research streams, our study provides a quantifiable assessment of cyber cognitive attack effectiveness, explicitly linking sentiment analysis and topical framing to measurable engagement outcomes on social media platforms.

Methodology

Our methodology is a reusable analytical framework for evaluating cognitive attack effectiveness using measurable indicators of resonance and proliferation. Later, we apply the methodology to the *FakeNewsNet* dataset, focusing on sentiment analysis, topic classification, and engagement metrics.

Terminology. Throughout, we refer to PolitiFact-labeled *fake* items as **False** and to PolitiFact-labeled *true* items as **Real**.

Research Questions (RQs)

These research questions guide our investigation.

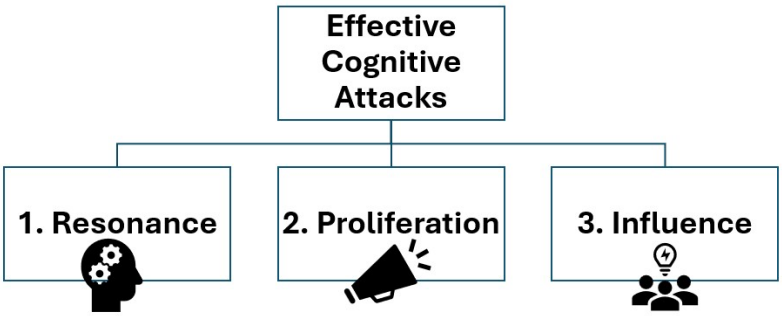
RQ1 (Resonance – Sentiment Framing): Do cyber cognitive attack narratives differ significantly in emotional sentiment from verified real news narratives?

RQ2 (Proliferation – Virality): How does emotional sentiment correlate with the virality of false versus real news narratives on social media?

RQ3 (Sentiment–Topic Interaction): Which topical categories (e.g., politics, health, celebrity) exhibit the strongest interaction between emotional sentiment and narrative virality?

RQ4 (Framework Support): How can empirical sentiment and engagement data be used to provide quantitative support for the DISARM Execution (PO3) tactics?

Figure 1
Cyber Cognitive Attack Effects Chain



This conceptual model illustrates the three critical stages of cognitive attack success: (1) Resonance, in which emotionally charged or sentiment-rich content captures attention and

triggers cognitive biases; (2) Proliferation, where the narrative is amplified through sharing behaviors such as retweets and (3) Influence, where the content ultimately alters perceptions, decisions, or behaviors in alignment with the attacker's objectives.

Metrics

We define a three-phase *Cyber Cognitive Attack Effects Chain*: **Resonance** (affective signal inferred from text sentiment), **Proliferation** (behavioral spread measured by engagement), and **Influence** (downstream impact on beliefs/behavior, typically requiring longitudinal or off-platform data). This paper operationalizes the first two phases and treats *Influence* as future work.

Definition 1 (Effective Cyber Cognitive Attack). *Content that resonates emotionally, proliferates through sharing, and ultimately influences perceptions or behavior in line with the attacker's goals.*

Definition 2 (Resonance). *The narrative's affective pull, inferred quantitatively from sentiment polarity/strength (e.g., VADER compound) as a proxy for an emotional hook.*

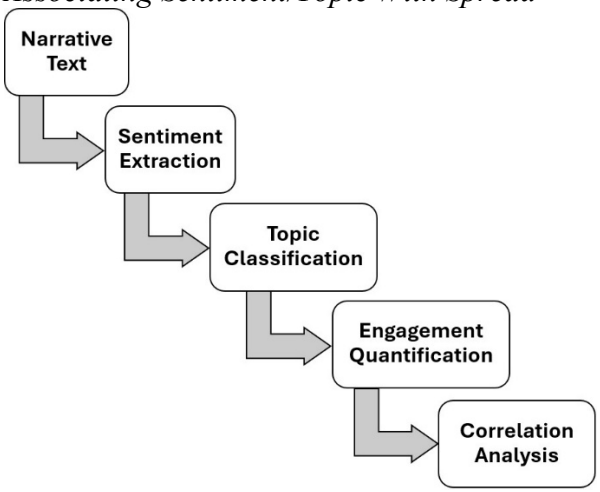
Definition 3 (Proliferation). *The extent of spread, measured by engagement (e.g., tweet/retweet counts) at the article level.*

Definition 4 (Influence). *Downstream impact on beliefs or actions; hardest to measure without surveys, field data, or natural experiments.*

Generalizable Methodology

The framework applies wherever text, a veracity/risk label, and engagement exist. It standardizes three elements: (i) **Resonance** via sentiment polarity/strength; (ii) **Proliferation** via article-level engagement; (iii) **Influence** reserved for settings with longitudinal or off-platform outcomes. Standardization enables cross-dataset comparisons, tail-aware monitoring, and topic-specific triage.

Figure 2
Four Steps: (1) Sentiment Extraction; (2) Topic Tagging; (3) Engagement Aggregation; (4) Associating Sentiment/Topic With Spread



Applicable Dataset Criteria

To apply this methodology, a dataset should contain the following attributes: (i) Labeled narratives (e.g., true/false, benign/deceptive), (ii) Text content suitable for sentiment analysis, (iii) Engagement metrics (e.g., shares, likes, retweets), and (iv) Optional: Topical context or metadata (e.g., tags, categories).

Methodology Steps

The methodology employs four steps to categorize and analyze narrative texts, as depicted in Figure 2: 1. *Sentiment Extraction*: Apply rule-based (e.g., VADER) or ML-based sentiment models to narrative text to generate polarity scores. 2. *Topic Classification*: Categorize narratives by domain (e.g., health, politics, celebrity) to analyze sentiment-topic interactions. 3. *Engagement Quantification*: Aggregate share/retweet/reaction metrics as proxies for narrative proliferation. 4. *Correlation Analysis*: Examine relationships between sentiment intensity and engagement to infer cognitive resonance and spread dynamics.

Statistical Analysis

We complement descriptive summaries with three inferential components: (i) Pearson's χ^2 tests of independence compare sentiment distributions across veracity classes (reporting Cramér's V) (Agresti, 2013); (ii) tweet counts are modeled via a negative binomial (NB) regression with log link (Cameron & Trivedi, 2013; Hilbe, 2011; McCullagh & Nelder, 1989); (iii) tail behavior is examined via quantile regressions at $\tau \in \{0.5, 0.75, 0.9\}$ (Koenker, 2005; Koenker & Bassett, 1978).

Where families of hypotheses arise, we control the false discovery rate (FDR) using the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995).

NB Regression

Let Y_i denote the tweet count for article i, and let $\text{Fake}_i \in \{0, 1\}$ indicate veracity (1 = fake). Under the NB2 parameterization, $Y_i \sim \text{NB}(\mu_i, \alpha)$ with $E[Y_i] = \mu_i$ and $\text{Var}(Y_i) = \mu_i + \alpha\mu_i^2$, where $\alpha > 0$ is the over-dispersion parameter. We model the expected tweet count using a log-link negative binomial regression:

$$\log \mu_i = \beta_0 + \begin{bmatrix} \text{Fake}_i & \mathbf{S}_i^\top & \mathbf{T}_i^\top & (\text{Fake}_i\,\mathbf{S}_i)^\top & (\text{Fake}_i\,\mathbf{T}_i)^\top \end{bmatrix} \begin{bmatrix} \beta_F \\ \boldsymbol{\beta}_S \\ \boldsymbol{\beta}_T \\ \boldsymbol{\beta}_{S\times F} \\ \boldsymbol{\beta}_{T\times F} \end{bmatrix}. \qquad (1)$$

We report coefficients as incidence rate ratios (IRR = e^{β}) with robust standard errors. In words, we model the expected tweet count $\mu_i = E[Y_i]$ with a log link. Exponentiating coefficients gives IRR: for a given category, $e^{\beta} > 1$ means higher expected tweets than the baseline, $e^{\beta} < 1$ means lower.

Sentiment Binning

We compute VADER's compound sentiment score for each text, which ranges from −1 (most negative) to +1 (most positive). Following standard practice (Hutto & Gilbert, 2014), we bin items as *positive* if compound ≥ 0.05, *negative* if compound ≤ −0.05, and *neutral* otherwise. To assess sensitivity, we perturb the positive/negative cutoffs by ±0.02 and obtain substantively unchanged results.

Sentiment Source (Titles Only)

We compute sentiment from article titles (headlines) rather than full text. This choice targets the decision surface at share time: a large, unbiased measurement study of Twitter traffic found that 59% of links shared were never clicked by the sharer, implying that, for many users, the headline is the only linguistic stimulus encountered before propagation (Gabielkov et al., 2016). Headlines are also intentionally engineered and experimentally shown to shape engagement at scale (Matias et al., 2021). Cognitive research further shows that headlines pre-frame interpretation, bias memory and inferences, and can produce durable misbelief even when article bodies are read (Ecker et al., 2014). Because emotion (especially high-arousal affect) predicts virality, we operationalize affect via headline sentiment at this stage (Berger & Milkman, 2012). We use VADER, a rule-based sentiment model validated on short, social-media–style texts and robust to punctuation, casing, boosters, and negation patterns common in headlines (Hutto & Gilbert, 2014).

Limitations (Scope of Inference)

Title-only sentiment does not capture in-article nuance or rhetorical structure; accordingly, we interpret our results as predictors of Proliferation (sharing) rather than persuasion, and report robustness checks below.

Robustness and Sensitivity

- *Subset concordance:* On a subset with body text available, correlate headline vs. lead-paragraph sentiment; report sign agreement and Pearson/Spearman r.
- *Model stability:* Refit core models using (i) lead-paragraph sentiment and (ii) the mean of headline+lead; verify qualitative conclusions unchanged.
- *Publisher style controls:* Add outlet/domain fixed effects to reduce confounding by editorial tone.
- *Threshold sensitivity:* Vary VADER bin cutoffs by ±0.02 and report unchanged directionality.

Engagement Definition and Outlier Policy. We define article-level engagement as the total number of associated tweets/retweets provided by FakeNewsNet for that article (including zeros), aggregated at the article level without normalization. Outliers are excluded in figures only, using Tukey's IQR rule (Tukey, 1977) (values < Q1 − 1.5 · IQR or > Q3 + 1.5 · IQR), computed within each {veracity, sentiment} cell. In total, 175 (16.57%) observations were flagged and removed.

Case Study: PolitiFact Dataset

To demonstrate the usefulness of our methodology, we apply it to the *FakeNewsNet* PolitiFact dataset (Shu et al., 2018), which includes 1,056 labeled news articles (624 real, 433 false) and their associated tweet engagement metrics. The articles are dated from 2010–2018 (median date is July 2, 2018). Articles are verified as *real* or *false* by nonpartisan PolitiFact fact-checkers (Shu et al., 2018; Su et al., 2023a; Su et al., 2023b) and contain associated tweet IDs and engagement (retweet) counts. We use only public tweet IDs and article metadata as distributed by FakeNewsNet; no additional user data were collected.

Sentiment Analysis Implementation

We use VADER (Hutto & Gilbert, 2014) and its compound score (−1 to +1) as our polarity measure, given its calibration for short, informal text and robust performance on social content; extending to neural models is future work.

Topic Classification Implementation

Articles are tagged into *politics*, *health*, *celebrity*, or *other* via simple keyword rules (e.g., politics: *election, congress, trump/biden*; health: *vaccine, virus, pandemic*). The goal is interpretable buckets for interaction analysis, not exhaustive taxonomy. This lightweight heuristic approach enables scalable and interpretable topical labeling suitable for analyzing large, pre-labeled datasets such as FakeNewsNet (Shu et al., 2017; Vosoughi et al., 2018; Wang et al., 2019).

Narrative Effectiveness Metrics

We operationalize the effectiveness of cyber cognitive attacks using a quantifiable metric: the number of tweets and retweets associated with each news article. Prior work in information diffusion and online influence measurement has validated retweet volume as a proxy for narrative proliferation (Stieglitz & Dang-Xuan, 2013; Vosoughi et al., 2018).

Analytical Tools

All analysis used standard Python libraries.[4] Anonymized code and scripts sufficient to reproduce all figures/tables will be provided in the supplemental artifact for review.

Analysis and Findings

Addressing RQ1: Emotional Sentiment Differences

RQ1: Do cyber cognitive attack narratives differ significantly in emotional sentiment from verified real news narratives? *(Resonance – Sentiment Framing)*

We compare VADER-binned polarity (negative/neutral/positive) by veracity to test whether false items are more emotionally extreme.

[4]Implemented with pandas, nltk, and matplotlib Python libraries.

Table 1
Sentiment Distribution by Veracity (VADER). VADER Thresholds: ≥ 0.05 Positive, ≤ −0.05 Negative, Else Neutral. Percentages Are Row-Wise

News Type	n	%	n	%	n	%	n
False	198	45.8	162	37.5	72	16.7	432
Real	94	15.1	412	66.0	118	18.9	624
Total	292	26.4	574	51.8	190	17.2	1,056
	Negative		Neutral		Positive		Total

Key insights from Table 1:

- **False narratives** have more than double the number of negative articles compared to real news.
- **Real news** tends to be more neutral in tone, with 412 out of 624 (66%) articles falling in the neutral sentiment range.
- This suggests that **negative emotional tone is more characteristic of cognitive attacks.**

Insight 1. *False narratives exhibit greater emotional polarization than real news, confirming the use of sentiment-laden language as a cognitive trigger. This supports the first stage (Resonance) in our Cognitive Attack Effects Chain, where emotional framing primes audiences for deeper engagement.*

Addressing RQ2: Sentiment and Narrative Virality

RQ2: How does emotional sentiment correlate with the virality of false versus real news narratives on social media? *(Proliferation – Virality)*

We relate sentiment bins to article-level tweet counts to assess whether emotional extremity, especially negative valence, corresponds to greater spread, and whether this differs by veracity.

Negative Sentiment Engagement Patterns. The proportion of tweet engagement tied to negatively valenced article titles differs starkly. Within the false tweets, 28.8% of tweets come from negative-sentiment articles (vs. 8.6% of real news tweets).

Insight 2. *False narratives rely on negative sentiment nearly 3× more to drive engagement. This aligns with cognitive security research showing that emotionally provocative and negatively valenced content spreads more rapidly, especially when false.*

Aggregate Virality Patterns: Total tweets (all articles)— False: 165,352; Real: 418,164. Average tweets per article— False: 265.2; Real: 464.2. Median tweets per article— False: 40.0; Real: 4.0.

Insight 3. *Although real news has higher total and average tweet volume, false narratives are shared more consistently, with a median tweet count nearly 10× higher. This suggests that false narratives drive broader mid-level engagement, while real news depends on a few high-performing stories.*

Engagement distributions are right-skewed with heavy upper tails: a small minority of articles accounts for a large share of exposure. Consequently, we report medians and upper quantiles (75th/90th) alongside means, using tail-aware models (Koenker, 2005; Koenker & Bassett, 1978).

Figure 3

Tweet Engagement by Sentiment Level for Real vs. False Narratives. False Narratives Consistently Show Higher Median Tweet Counts Across All Non-neutral Sentiment Levels

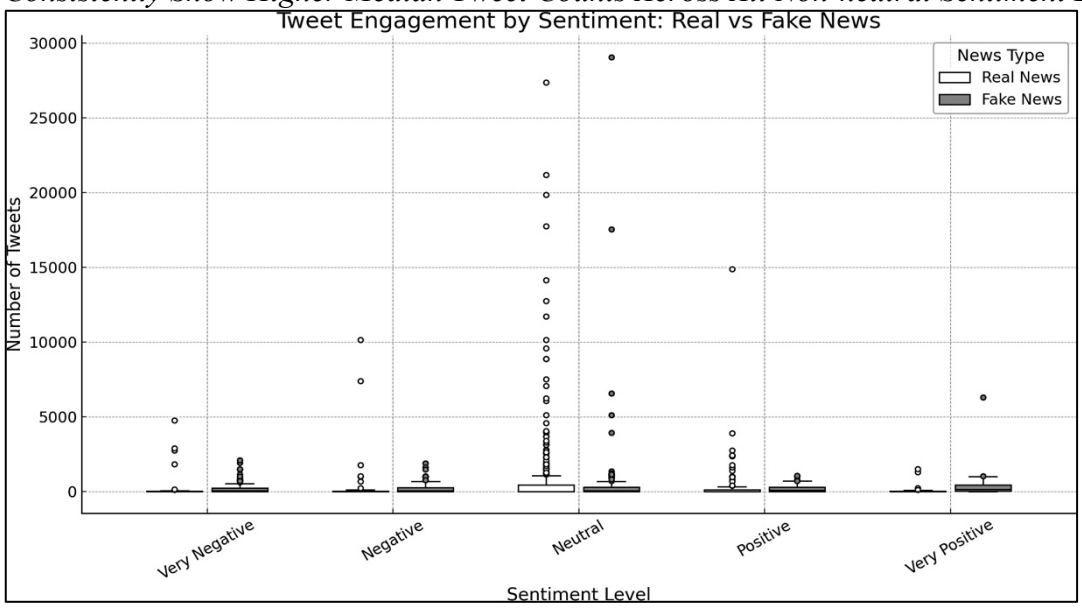
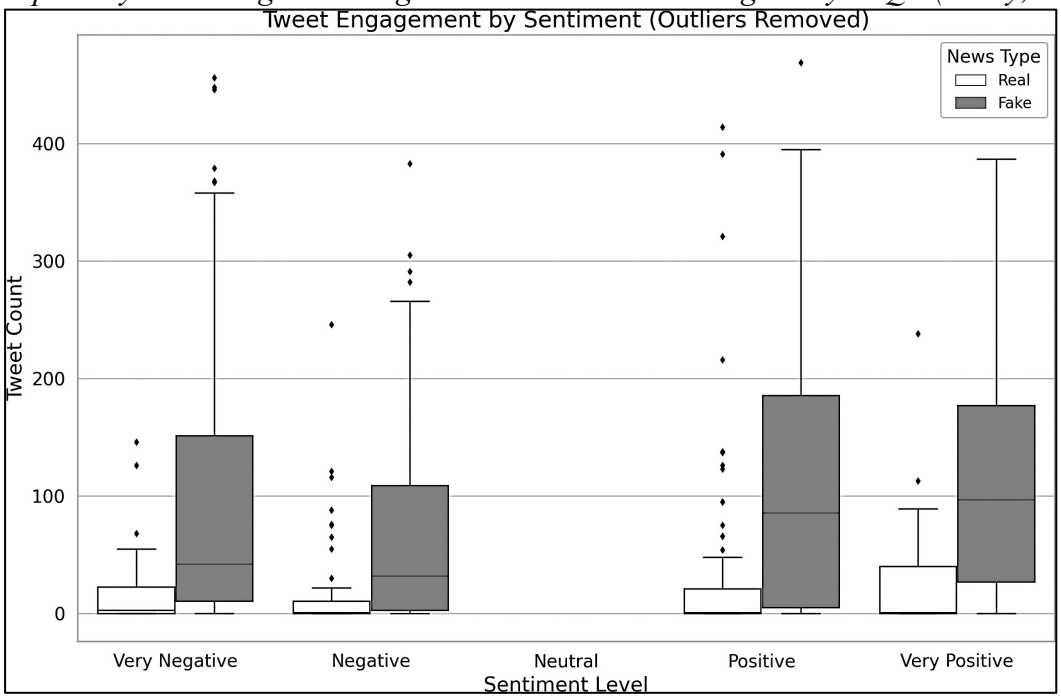


Figure 4

Tweet Engagement by Sentiment (Neutral Sentiment and Outliers Removed). Each Point Represents an Article, Comparing Tweet Counts for Real and False Narratives Across Sentiment Bins (Neutral Excluded for Clarity). Dark Gray Points (False) Cluster Higher, Especially in the Negative Range. Outliers Removed Using Tukey's IQR (Tukey, 1977)



Most Polarized Narratives. Tables 2 (real news) and 3 (false narratives) list the top five most emotionally polarized article titles, ranked by absolute compound VADER sentiment scores. Notably, false articles often feature sensational or conspiratorial themes, such as political death hoaxes or state violence, while real articles focus on contentious but fact-based reporting.

Table 2
Top 5 Most Polarized Real News Articles (PolitiFact-Labeled "Real")

Rank	Title	Sentiment Score
1	*Hannity says Obama won't even use the term "war on terror"*	−0.8591
2	*Call 'Islamic terrorism' what it is: a threat to freedom*	−0.8402
3	*Georgia crisis triggers war of words on White House campaign trail*	−0.8402
4	*Barack Obama says Mitt Romney condemned coal-fired power plants*	−0.8020
5	*Most heroin in U.S. now comes across Mexican border, not from Afghanistan*	−0.7960

Table 3
Top 5 Most Polarized False News Articles (PolitiFact-Labeled "Fake")

Rank	Title	Sentiment Score
1	*Russian source behind Trump dossier killed in mystery helicopter crash*	−0.9423
2	*Could Trump Win The Nobel Peace Prize? Peace in Korea Could Make It Happen*	0.9423
3	*Donald Trump Dead From A Fatal HEART ATTACK! — White House In Chaos*	−0.9173
4	*82-Year-Old Who Killed A Muslim In Self-Defense Cleared Of All Charges*	−0.9081
5	*Seattle Police Begin Gun Confiscations: No Law Broken, No Warrant, Just Guns Gone*	−0.9042

Addressing RQ3: Sentiment–Topic Interaction

RQ3: Which topical categories (e.g., politics, health, celebrity) exhibit the strongest interaction between emotional sentiment and narrative virality? *(Proliferation – Sentiment–Topic Interaction)*

We examine topic–sentiment interactions by averaging engagement within {topic, sentiment, veracity} cells to identify high-risk pockets that amplify spread. Topics were classified into politics, health, celebrity, or other, covering 41.3% (436/1,056) of the dataset.

Table 4

Sentiment by Topic and Veracity (VADER). † Small n; Interpret With Caution

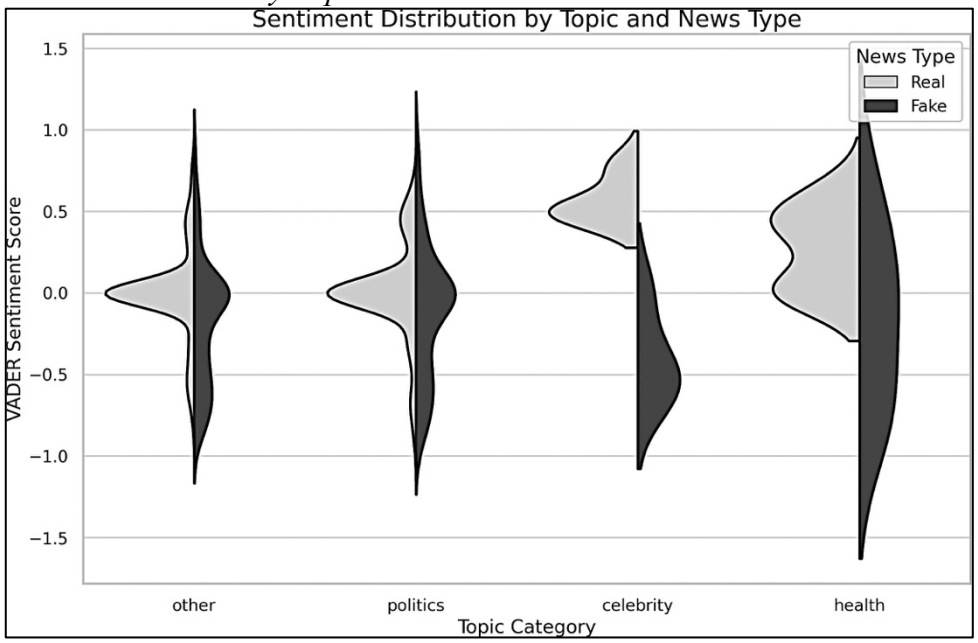
Label	Topic	Mean	Median	Std	Count
False	Celebrity†	−0.409	−0.494	0.283	4
False	Health	−0.161	0.000	0.558	7
False	Other	−0.189	0.000	0.388	251
False	Politics	−0.126	0.000	0.408	170
Real	Celebrity†	0.565	0.494	0.142	4
Real	Health	0.270	0.382	0.251	15
Real	Other	−0.001	0.000	0.268	369
Real	Politics	0.009	0.000	0.283	236

Engagement by Topic. Average tweet count per article, including zero-tweet items: Politics — Real: 845.25; False: 178.54. Health — Real: 82.00; False: 441.17. Celebrity — Real: 0.00; False: 465.00. Other — Real: 642.81; False: 457.05.

Insight 4. *False narratives outperform real news in health and celebrity topics, likely due to emotional, sensational, or conspiratorial framing. Real political news earns the highest average engagement overall, suggesting a concentration of mainstream amplification.*

Figure 5

Sentiment Distribution Across Topic Categories in Real and False Narratives Headlines. False Narratives Exhibit a Broader and More Negative Sentiment Spread, Particularly in Politics and Celebrity Topics



Key Takeaways:

- **Emotion + Topic = Virality.** Emotionally extreme disinformation, especially in health and celebrity categories, correlates with higher tweet engagement.

- **Real news dominates political engagement,** but false narratives maintain broader traction in emotionally volatile domains.
- **False narratives amplify sentiment variance,** reinforcing their potency in the Proliferation stage of the Cognitive Attack Effects Chain.

Addressing RQ4: Framework Support via Empirical Signals

RQ4: How can empirical sentiment and engagement data be used to provide quantitative support for the DISARM Execution (PO3) tactics? *(Framework Support)*

DISARM (DISARM Foundation, 2024) enumerates attacker tactics but offers limited quantitative support. Our dataset enables an operational crosswalk from observed signals (sentiment extremity, engagement/virality, topic–sentiment interactions) to DISARM's Execution (PO3) tactics.

Table 5

Crosswalk: DISARM PO3 Tactics and Associated Empirical Proxies

Tactic (ID)	Observable Proxy/Evidence
TA17: Maximise Exposure	Higher medians and heavy upper tails in tweet counts across non-neutral sentiment bins (Table 1; Figs. 3, 4).
TA09: Deliver Content	Topic–sentiment "pockets" with outsized engagement (false–health/celebrity); see Fig. 5 and topic-level averages.
TA18: Drive Online Harms	High-engagement fear/conspiracy/victimization frames among polarized false items (Tables 2, 3).
TA08: Pump Priming	Requires temporal diffusion traces, not testable here.
TA11: Persist	Requires longevity/evading moderation, not testable here.
TA10: Drive Offline Activity	Requires off-platform linkage, not testable here.

Insight 5. *Our sentiment and virality measures provide quantitative support for Maximise Exposure (TA17) and targeted Deliver Content (TA09), and are consistent with Drive Online Harms (TA18) in how polarized false narratives mobilize affective themes. Temporal and off-platform tactics (TA08, TA10, TA11) remain promising targets for future datasets with timestamps and event linkages.*

Discussion

This work advances the measurement of cyber cognitive attacks by (i) formalizing a three-phase *Cyber Cognitive Attack Effects Chain*: *Resonance → Proliferation → Influence*, and (ii) demonstrating with real data that sentiment is a scalable correlational signal for the first two phases. Across RQ1–RQ3, we find that false narratives exhibit greater emotional polarization than real news, that negative sentiment accounts for a disproportionate share of their engagement, and that topic–sentiment "pockets" (e.g., false × health/celebrity × negative) are especially virality-prone. These patterns align with the ABCs of Influence—Affect, Behavior, Cognition—and empirically support parts of DISARM's execution-phase tactics.

Resonance-Proliferation Score (RPS)

The Resonance–Proliferation Score (RPS) operationalizes the first two stages of our effects chain in a single, bounded, interpretable metric. By multiplying sentiment extremity (resonance) by a topic-normalized engagement percentile (proliferation), RPS highlights items that are both emotionally potent and unusually viral for their topic. Because proliferation is percentile-based, the score is robust to heavy-tailed engagement and comparable across topics and time.

For article i, let $s_i \in [-1, 1]$ be the VADER compound score and y_i the tweet count. Let $\tau_i \in$ {politics, health, celebrity, other} denote its topic.

Resonance

$$E_i = |s_i| \in [0,\ 1]. \qquad (2)$$

Optionally upweight negative valence via

$$E_i(\gamma) = \gamma \max\{0,\ -s_i\} + (1-\gamma)\max\{0,\ s_i\}, \quad \gamma \in [0,\ 1],$$

with $\gamma = 0.7$ by default (*chosen a priori; tuning γ is left to future work*).

Proliferation (Topic-Normalized Percentile)

Using a mid-rank empirical CDF within topic:

$$P_i = \frac{\#\{j \in G_{\tau_i} : y_j < y_i\} + 0.5\ \#\{j \in G_{\tau_i} : y_j = y_i\}}{n_{\tau_i}} \in (0,\ 1]. \qquad (3)$$

(If $n\tau_i = 1$, set $P_i = 1$.)

Score

$$\mathrm{RPS}_i = E_i \cdot P_i \in [0,\ 1]. \qquad (4)$$

High values thus require both strong sentiment (resonance) and unusually high, topic-relative spread (proliferation).

Group Comparison and Thresholding

$$\mathcal{A}_{\mathrm{RPS}} = \frac{\mathrm{median}(\mathrm{RPS}_i \mid \mathrm{False})}{\mathrm{median}(\mathrm{RPS}_i \mid \mathrm{Real})}. \qquad (5)$$

To flag high-risk items, choose a threshold θ as the 90th percentile of RPS among Real articles and alert when $\mathrm{RPS}_i \geq \theta$.

In words, RPS multiplies sentiment extremity (resonance) by a topic-normalized engagement percentile (proliferation), so high values flag emotionally charged items that spread unusually widely for their topical peer group.

Implications for Detection and Defense

Our findings suggest concrete levers for defenders seeking to prioritize monitoring and early intervention. *Sentiment-first triage:* Because false narratives lean heavily on negative valence, a low-latency filter that surfaces spikes in negative polarity variance can act as a resonance alarm. *Targeted watchlists by topic:* The strongest proliferation occurs within specific sentiment–topic pockets. Maintaining dynamic watchlists (e.g., health, celebrity) and flagging sentiment shifts inside those domains improves precision and reduces analyst load. *Proliferation diagnostics:* Use distribution-sensitive indicators (median, 75th/90th percentiles, tail share) to detect abnormal spread early. *Content response strategy:* Because resonance is affective, counter-messaging should avoid mirroring outrage. Interventions that de-escalate emotion, foreground uncertainty, and provide concise, credible alternatives are more likely to disrupt the resonance→proliferation link. *Human-in-the-loop:* Automated sentiment/topic flagging should feed analyst workflows that assess coordination signals and policy context before enforcement or labeling. The push to operationalize misinformation risks in the generative-AI era motivates measurable affective signals (e.g., sentiment polarity) as early indicators of persuasive content (Zhou et al., 2023).

Connecting Theory to Operations: ABCs and DISARM

The effects chain operationalizes the ABCs: *Affect* (Resonance) captured via sentiment polarity/variance; *Behavior* (Proliferation) captured via sharing metrics; *Cognition* (Influence) requiring longer-horizon or off-platform evidence. Mapping these observables to DISARM's PO3 tactics, our measurements provide quantitative support for Maximise Exposure (TA17) and targeted Deliver Content (TA09), and are consistent with patterns expected under Drive Online Harms (TA18). Tactics that require temporal traces or offline linkage (Pump Priming (TA08), Persist (TA11), Drive Offline Activity (TA10)) remain promising but untested in our dataset.

RPS Triage Dashboard

Purpose and Users

The dashboard prototype depicted in Figure 6 supports content-moderation analysts and threat-intel reviewers who must quickly triage emerging narratives. It ranks items with a lightweight, interpretable signal—the Resonance–Proliferation Score (RPS), defined as sentiment extremity × topic-normalized spread, motivated by evidence that emotionally charged content spreads faster and farther online (Cervone et al., 2024; Duch, 2021; Vosoughi et al., 2018). The goal is to operationalize doctrine (e.g., DISARM execution goals and cognitive-warfare aims) with transparent, data-driven cues (Claverie & du Cluzel, 2024; DISARM Foundation, 2024; François, 2019).

Interaction Model

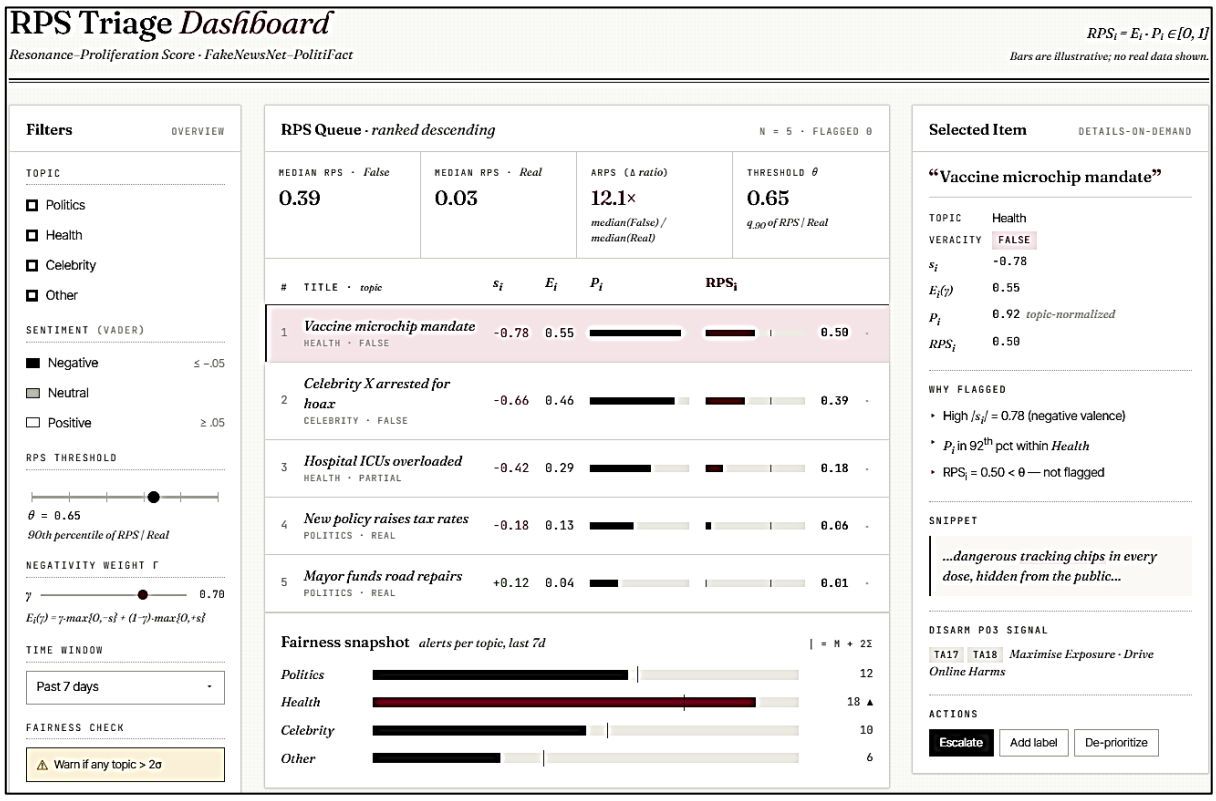
The layout follows an overview–filter–details flow to reduce cognitive load in scan-and-commit workflows: the left panel offers low-latency filters (topic, sentiment) and a threshold slider anchored to the 90th percentile of Real items. The center panel lists items with compact preattentive encodings: $|s_i|$ from VADER, topic-normalized percentile P_i, and the composite RPS. The right panel provides details-on-demand (snippet and "why flagged"), supporting accountable triage tied to observable features.

Fairness and Safety Cues

A fairness snapshot warns when alerts cluster within a topic (e.g., > 2σ from baseline). This widget serves as a pragmatic cue for auditing skew in exposure or moderation focus (Caramancion, 2023; Juneja & Mitra, 2021).

Figure 6

Prototype RPS Triage Dashboard. Left: Filters (Topic, Sentiment, RPS Threshold). Center: Ranked Queue With Per-Item Si, |Si|, Topic-Normalized Percentile Pi, and RPS. Right: Details and Explanation for the Selected Item. Bars Are Illustrative; No Real Data Are Shown



Methodological Considerations and Robustness

Heavy-tailed engagement warrants robust summaries (medians, quantiles) and tail-aware alerts. Construct validity improves by triangulating sentiment (lexicon/model, subjectivity, moral emotions) (Koenker, 2005). Confounders include coordinated amplification, account size/verification, exogenous events, and algorithm changes; future models should add timing/network controls. Sensitivity checks (winsorization, zero-tweet removal, alternative binnings/classifiers) yield consistent qualitative conclusions.

Generalizability and Scope

The case study is U.S.-centric and time-bounded, but the methodology is platform-agnostic wherever text, labels, and engagement exist. Multilingual settings and platform affordances may shift baselines; cross-language/platform evaluation is future work.

Research Agenda

Next steps include: adding timestamps for diffusion/pump-priming and persistence; integrating coordination/community structure and cross-platform flows; linking to downstream outcomes via natural experiments or policy shocks; testing countermeasures (labels, friction, re-ranking) with A/B designs; and enriching affect models beyond polarity (discrete emotions, moral foundations, toxicity).

Ethical and Policy Considerations

Operationalizing sentiment for defense must balance safety with expression. Risk controls include transparent criteria for flags, narrow scopes tied to measurable harms, audits for disparate impact, and appeal pathways. Research reproducibility (code, metrics, and pre-registered analyses where possible) further reduces overreach and supports policy trust.

Reproducibility and Ethics. We analyze only public data. No attempt was made to deanonymize users. Code and analysis scripts will be made available upon acceptance to support reproduction of results.

Limitations

The FakeNewsNet dataset lacks certain metadata (e.g., timestamps, retweet destinations, user-level details) that could further enhance our understanding of narrative dynamics. The dataset largely reflects disinformation and engagement patterns from the 2016–2018 period, limiting real-time generalizability without more recent data. Our methodology relies on PolitiFact as a baseline for "true" or "false" narrative labels; this data primarily comprises US-based topics and posts, limiting the scope of the research.

Additionally, this paper primarily covers the proliferation (spread) of emotional online content and does not examine the end goal of attackers—specifically, the susceptibility, perception, and behavioral changes of users who encounter the messaging. These projects remain future research goals.

Conclusion

This paper introduced a generalizable methodology for measuring cyber cognitive attack effectiveness and a three-phase *Cyber Cognitive Attack Effects Chain*: *Resonance → Proliferation → Influence*. Applying the framework to the FakeNewsNet PolitiFact corpus, we showed that false narratives are more emotionally polarized than real news, rely disproportionately on negative valence, and achieve substantially higher median engagement—evidence of heavy upper tails in their diffusion. We further identified topic–sentiment pockets (notably false–health/celebrity–negative) associated with outsized virality, and mapped these empirical signals to DISARM execution tactics, providing quantitative support for Maximise Exposure (TA17) and targeted Deliver Content (TA09), with patterns consistent with Drive Online Harms (TA18).

Practically, our results suggest sentiment-first triage and topic-aware watchlists, coupled with tail-sensitive monitoring (medians/quantiles rather than means), as feasible levers to disrupt the resonance→proliferation link early. Methodologically, the framework is platform-agnostic and can be replicated wherever text, basic engagement signals, and labels are available.

Overall, by aligning doctrinal models (DISARM, ABCs) with measurable observables (sentiment and engagement), this work offers a defensible path from theory to operations for proactive, data-driven cognitive defense.

Author's Note

The views expressed are those of the author and do not reflect the official policy or position of the US Air Force, Department of War, or the US Government.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

The author declares that Grammarly, an AI-assisted writing software, was used to proofread and refine the language of the manuscript. The usage was limited to correcting grammatical and spelling errors and rephrasing statements for accuracy and clarity. The author further declares that, apart from Grammarly, no other AI or AI-assisted technologies have been used to generate content in writing the manuscript. The ideas, design, procedures, findings, analyses, and discussion are originally written and derived from appropriate and systematic conduct of the research.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289–300.

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research, 49*(2), 192–205.

Bontridder, N., & Poullet, Y. (2021). The role of artificial intelligence in disinformation. *Data & Policy, 3*(3), e32.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2nd ed.). Cambridge University Press.

Caramancion, K. M. (2023). Harnessing the power of ChatGPT to decimate mis/disinformation: Using ChatGPT for fake news detection. In *2023 IEEE World AI IoT Congress (AIIoT)* (pp. 42–46).

Cervone, C., Thompson, R., Votta, F., & Tucker, J. A. (2024). Memes, emotional engagement, and politics. *Research and Politics, 11*(2), 1–11.

Claverie, B., & du Cluzel, F. (2024). "Cognitive warfare": The advent of the concept of "cognitics" in the field of warfare. In *STO-MP-HFM-361: Cognitive Warfare. NATO Science and Technology Organization*. (Original work presented March 2022)

DISARM Foundation. (2024). *DISARM framework explorer*. https://www.disarm.foundation/framework

Duch, W. (2021). Memetics and neural models of conspiracy theories. *Patterns, 2*(11), 100353.

Ecker, U. K. H., Lewandowsky, S., Chang, E. P., & Pillai, R. (2014). The effects of subtle misinformation in news headlines. *Journal of Experimental Psychology: Applied, 20*(4), 323–335.

François, C. (2019). *The ABC framework for influence operations: A behavioral model for understanding disinformation*. Harvard Kennedy School, Belfer Center for Science and International Affairs.

Gabielkov, M., Ramachandran, A., Chaintreau, A., & Legout, A. (2016). Social clicks: What and who gets read on Twitter? In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems* (pp. 179–192). ACM.

Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). Cambridge University Press.

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM-14)* (pp. 216–225). AAAI Press.

Juneja, P., & Mitra, T. (2021). Auditing e-commerce platforms for algorithmically curated vaccine misinformation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 186:1–186:27). ACM.

Koenker, R. (2005). *Quantile regression*. Cambridge University Press.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46*(1), 33–50.

Krishnan, S. S., & Sitaraman, R. K. (2013). Understanding the effectiveness of video ads: A measurement study. In *Proceedings of the 2013 Internet Measurement Conference (IMC '13)* (pp. 149–162). ACM.

Matias, J. N., Munger, K., Aubin Le Quere, M., & Ebersole, C. (2021). The Upworthy Research Archive, a time series of 32,487 experiments in U.S. media. *Scientific Data, 8*(1), 195.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman and Hall.

Rushing, B., & Xu, S. (2026). Characterising cyber cognitive attacks. *The RUSI Journal, 171*(2), 100–114. https://doi.org/10.1080/03071847.2026.2638130

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2018). *FakeNewsNet: A data repository with news content, social context and dynamic information for studying fake news on social media*. https://arxiv.org/abs/1809.01286

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

Stieglitz, S., & Dang-Xuan, L. (2013). Social media and political communication: A social media analytics framework. *Social Network Analysis and Mining, 3*(4), 1277–1291.

Su, J., Cardie, C., & Nakov, P. (2023a). *Adapting fake news detection to the era of large language models*. https://arxiv.org/abs/2311.04917

Su, J., Zhuo, T. Y., Mansurov, J., Wang, D., & Nakov, P. (2023b). *Fake news detectors are biased against texts generated by large language models*. https://arxiv.org/abs/2309.08674

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

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

Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019). Systematic literature review on the spread of health-related misinformation on social media. *Social Science & Medicine, 240*, 112552.

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023). Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 436:1–436:20). ACM.

Contact email: bonnie.rushing@uccs.edu