Frontiers in Economics and Management Volume 7 Issue 4, 2026
ISSN: 2692-7608 DOI: 10.6981/FEM.202604_7(4).0007

A Predictive Model for Early Intervention Efficacy of Fake News based on Epistemic Vigilance

Yihan Zhao*

School of Economics and Management, Tongji University, Shanghai 20092, China

*Corresponding Author

Abstract

Early automated intervention against fake news is critical for social media platform governance. This paper proposes a predictive method for intervention efficacy grounded in Epistemic Vigilance. First, using Truth-Default Theory (TDT) as a theoretical lens, empirical analysis on the RumourEval 2019 dataset confirms that "Deny" first-comments effectively break audiences' default inertia, significantly suppressing the subsequent support ratio to 0.525. Second, we translate this mechanism into a probability prediction task, utilizing the ELECTRA architecture and Focal Loss to address extreme sample imbalance. Cross-validation results demonstrate that our method substantially improves the recall rate of high-potential refutational texts to 0.898. By reversely applying stance evolution logic to active intervention evaluation, this study provides a scientific reference for platforms to deploy high-recall early-blocking strategies.

Keywords

Fake News; Epistemic Vigilance; Truth-Default Theory; Early Intervention; Natural Language Processing; Design Science.

1. Introduction

Social media decentralization enables fake news to bypass traditional gatekeepers and rapidly capture public attention[1]. Empirical evidence shows that misinformation exhibits a substantial advantage in diffusion speed, depth, and breadth over true news[2]. Current platform governance mechanisms, which rely primarily on expert fact-checking or post-hoc algorithmic blocking, exhibit significant time lags. By the time debunking interventions occur, misinformation has often surpassed its initial explosive growth phase. Therefore, exploring "early and instant intervention" mechanisms to sever the initial propagation chain has become an urgent research direction in Information Systems (IS) and computational communication.

To identify effective early intervention entry points, understanding the cognitive roots of misinformation susceptibility is essential. According to Truth-Default Theory (TDT), humans exhibit a heuristic "truth-default" bias during information exchange, presupposing information truthfulness unless encountering strong "trigger cues" that awaken "Epistemic Vigilance"[3]. In social media interfaces, the "First Comment" naturally functions as a high-potential trigger cue due to its visual prominence and chronological priority. Lee and Jang[4] demonstrated that first comments exert a powerful anchoring effect, framing subsequent group interactions. Consequently, front-loading a "Deny" or "Query" first-comment during the initial publication stage can trigger rational verification mechanisms before group cognitive biases solidify, thereby inhibiting cascading misinformation acceptance.

Although prior research confirms the positive utility of refutational comments[5] such interventions heavily rely on spontaneous user participation and lack scalability. While advancements in Natural Language Processing (NLP) enable proactive, automated

ISSN: 2692-7608 DOI: 10.6981/FEM.202604_7(4).0007

interventions, the blind deployment of ineffective corpora may trigger backfire effects. Following the Design Science paradigm, this study constructs and evaluates a lightweight, automated screening system for early intervention efficacy.

The core contributions are threefold: 1. Empirical Validation: We confirm the significant blocking effect of first-comment stances on group propagation trajectories (support ratio suppression) using the RumourEval dataset. 2. Algorithm Construction: We develop and compare optimized deep pre-trained language models (e.g., ELECTRA, BERT) to accurately extract "high-trigger-efficacy" interventional corpora. 3.Governance Framework: We provide a "data-driven, early-blocking" algorithmic governance architecture for platforms to guide rational public opinion.

2. Related Work

2.1. Limitations of Automated Fake News Detection and Intervention

Current computational fake news research primarily focuses on post-hoc "Detection" [6,7,8,9,10]. However, these technical interventions often suffer from delayed timeliness, low coverage, and the "implied truth effect"[11]. While deep neural networks leveraging text, user profiles, and propagation features [12,13]. achieve high accuracy, detection alone fails to resolve the governance dilemma. Passive labeling mechanisms may inadvertently increase trust in unlabeled fake news [11]. Thus, the research focus must urgently shift from passive detection to proactive intervention.

2.2. Truth-Default Theory and Epistemic Vigilance Mechanisms

Understanding audience susceptibility is a prerequisite for system design. TDT posits that humans possess a "default-to-truth" heuristic to conserve cognitive resources [3], facilitating the initial fission of fake news. Breaking this default requires activating "Epistemic Vigilance"—a psychological defense mechanism for critically evaluating information[14] During early propagation, audiences lack internal motivation to awaken vigilance due to the absence of authoritative debunking. External trigger cues, such as explicit queries or refutations in comments, act as critical stimuli to induce systematic verification. This cognitive foundation justifies focusing our automated intervention system on extracting refutational cues.

2.3. Truth-Default Theory and Epistemic Vigilance Mechanisms

Stance detection is a core NLP task in rumor governance. Traditional binary (Pro/Con) or ternary frameworks assume users possess clear authenticity judgment capabilities, failing to capture the cognitive transition from default belief to nascent suspicion. This study adopts the SDQC (Support, Deny, Query, Comment) framework[15], widely applied in SemEval tasks[8]. Isolating "Query" accurately maps the initial vigilance stage where audiences demand evidence. Furthermore, isolating "Comment" filters out marginal noise, focusing computational resources on strong trigger cues (Deny/Query).

Although previous studies suggest that social media user groups exhibit a certain degree of self-correction during long-term evolution[16]. Research indicates that during breaking events, initial discussion scales are small, and stances are mixed, with explicit corrections lagging official debunking [7]. This natural absence of rational verification comments during the golden intervention phase provides a realistic basis for introducing automated early interventions, see Table1.

ISSN: 2692-7608 DOI: 10.6981/FEM.202604_7(4).0007

Table 1. Stance Classification Categories and Literature Applications

Literature Source	Comment Stance	Definition	Example	Applied Literature
Procter et al., 2013b [15]	Support	Comment expresses trust, agreement, or reinforcement regarding the news.	"This is true, I saw it too!"	RumourEval 2019; Qazvinian et al., 2011[17] Zubiaga et al., 2016[7]; Derczynski et al., 2017[10] Gorrell et al., 2019[9]
	Deny	Comment explicitly refutes or denies the news.	"Fake, it has been officially debunked."	
	Query	Comment raises questions or requests more information.	"Is there a source? Where did you see this?"	
	Comment	Non-stance comments, such as emotions, jokes, or irrelevant content.	"This world is too crazy..."	
Murungi et al., 2018[18]; rhetorical theory	Rational refutation	Logically clear, evidence-based refutational comments.	"This data is obviously fabricated; the real statistics are..."	Zubiaga et al., 2018[8] Jin et al., 2020[19]
Paek & Hove, 2019[20] SCCT	Refuting / Denying / Attacking Source	Three strategies for responding to rumors in a crisis: refuting, denying, or attacking the source.	"This information is untrue, please do not spread it."	Oh et al., 2013[21]; SCCT literature
Oh et al., 2013 [20]; Liu et al., 2015[22]	Empathetic reassurance	Expressing emotional understanding to alleviate panic or anxiety.	"Don't panic everyone, the officials will explain soon."	Rumour theory; SCCT
Chua & Banerjee, 2017 [23]; health misinformation	Verification reminder	Reminding others to verify sources or wait for official news.	"Please verify the source first, don't be gullible."	Oh et al., 2013[21]

2.4. Potential of Social Bots as Early Intervention Mechanisms

Given users' initial lack of self-correction, automated programs offer a promising governance approach. While social bots historically amplify low-credibility content[24], timely early communication can alter propagation trajectories [25]. Recent research validates bots' potential in positive social interactions; posts with bot-generated comments experience higher engagement[26], aligning with the "Computers Are Social Actors" (CASA) paradigm[27]. Unlike traditional research focused on passive stance classification, we reversely apply stance theory to actively filter high-efficacy "Deny/Query" texts, exploring platform-level automated first-comment generation to stimulate epistemic vigilance without direct factual adjudication.

ISSN: 2692-7608 DOI: 10.6981/FEM.202604_7(4).0007

3. Task Formulation & Model Framework

3.1. Mapping Theory to Computational Tasks

TDT posits that individuals maintain default trust unless encountering strong "trigger cues" [3] A first comment with a "Query" or "Deny" stance acts as a third-party warning cue, breaking trust inertia and activating "epistemic vigilance"[14]. We operationalize this macro-level vigilance enhancement as the dynamic decay of the "Support Ratio" (SR) in propagation threads. For a subsequent comment set $C = \{c_1, c_2, \dots, c_N\}$, and a classification mapping function $S(c_i)$ extracting the SDQC label, the group-level SR is defined as:

$$SR = \frac{\sum_{i=1}^{N} I(S(c_i)=Support)}{\sum_{i=1}^{N} I(S(c_i)=Support)+\sum_{i=1}^{N} I(S(c_i)=Deny)} \quad (1)$$

Where $I(\cdot)$ is the indicator function. A significant SR drop mathematically corroborates a shift from "blind belief" to "epistemic vigilance". We reconstruct early intervention screening into a goal-oriented probability prediction task. The model learns a mapping $f(x) \to [0,1]$ to predict the probability (p_lowSR) that a candidate text x_i leads to a low subsequent support ratio, optimizing directly for "propagation blocking efficacy".

3.2. Quantitative Labeling and Hard Thresholding Strategy

To establish a clear optimization objective, we utilize RumourEval 2019 trajectories to quantitatively label efficacy. We set a propagation control safety threshold τ. If the subsequent support ratio $SR \leq \tau$, the first comment is assigned a positive label $y_i = 1$ ("high-potential"); otherwise, $y_i = 0$("weak-efficacy").

During inference, the predictor outputs a confidence score $p\_lowSR \in [0,1]$. Given the strict time sensitivity of early intervention, maximizing recall is paramount. Based on validation set Precision-Recall analysis, we introduce an asymmetric Hard Thresholding Strategy, lowering the decision boundary to maximize the inclusion of high-trigger-efficacy corpora into the intervention repository.

3.3. Algorithm Selection and Baseline Strategy

Traditional machine learning models (e.g., TF-IDF + Logistic Regression) struggle with the complex discourse dependency of social media refutations. We employ Transformer-based deep pre-trained language models (BERT and ELECTRA) as core classifiers. Their Self-Attention mechanisms precisely capture contextual dependencies. Specifically, ELECTRA, relying on the Replaced Token Detection (RTD) mechanism[28], demonstrates high sample efficiency in capturing discriminative semantic boundaries.

4. Empirical Analysis

4.1. Data Acquisition and Preprocessing

We utilize the RumourEval 2019 dataset[9], a benchmark corpus widely validated in stance evolution research [29,30]. It contains 446 real breaking event threads across Twitter and Reddit, comprising 8,010 user interaction comments. We operationalize the variables by extracting the "First Comment" and mapping its SDQC attribute. We use the early window "Support Ratio" (SR) as a proxy for group attitudes and Shannon Entropy to measure SDQC stance heterogeneity, representing activated Epistemic Vigilance.

ISSN: 2692-7608 DOI: 10.6981/FEM.202604_7(4).0007

4.2. Exploratory Data Analysis

Analysis of 294 valid propagation threads reveals significant heterogeneity. "Deny" first-comments suppress the mean subsequent SR to 0.525, significantly lower than the "Support" group's 0.744 ($p = 0.0081 < 0.01$). Furthermore, Deny-type threads exhibit peak stance diversity (Entropy = 0.897) compared to the Support group (Entropy = 0.680). This indicates that refutational cues successfully interrupt truth-default inertia, breaking echo chambers and prompting rational gaming structures, see Table2.

Table 2. Propagation Indicators under Different First-Comment Stances

First-Comment Stance	Support Ratio (SR)	Entropy
Support	0.744	0.680
Query	0.715	0.656
Deny	0.525	0.897

4.3. Predictive Model Construction

4.3.1. Algorithm Selection and Loss Function Optimization

To screen high-trigger-efficacy corpora, we implement event-grouped 5-fold cross-validation to prevent feature leakage. We utilize Logistic Regression (TF-IDF) as a shallow baseline, compared against BERT-base-uncased and ELECTRA. Addressing sample imbalance (valid texts constitute only 33.3%), we introduce Focal Loss [31] during fine-tuning, shifting computational focus to hard-to-recognize marginal contexts to improve minority-class recall.

4.3.2. Model Testing and Performance Evaluation

The LR baseline achieves a meager Recall Rate of 0.295, missing over 70% of potential texts. The ELECTRA model with Focal Loss achieves optimal discriminability (AUC = 0.589) and substantially increases recall to 0.898. In early interventions, prioritizing candidate scope over strict precision is core business logic; ELECTRA's advantage validates its application valuem, see Table3.

Table 3. Comparison Table of Prediction Performance of Models on the Test Set

Models	AUC	Precision	Recall Rate	F1-Score
Logistic Regression (TF-IDF)	0.581	0.371	0.295	0.325
BERT-base-uncased	0.585	**0.420**	0.840	**0.551**
ELECTRA + Focal Loss	**0.589**	0.404	**0.898**	0.550

4.3.3. Case Study on High-Trigger-Efficacy Intervention Corpus Screening

To guarantee system recall, we establish a hard truncation threshold ($\tau = 0.18$). Texts with predicted probability $p\_lowSR \geq 0.18$ are automatically incorporated into the repository. Qualitative analysis (Table 4) confirms the model accurately assigns high confidence to strong cognitive interventions (e.g., explicit falsification commands or official source verification requests), avoiding extreme aggression while aligning with TDT's "third-party warning cues"., see Table4.

Table 4. Examples of High-Trigger-Efficacy Intervention Texts Screened by the System

First-Comment Text	Blocking Probability (p_lowSR)	Decision Result (pred@thr=1)
Stop spreading misinformation. This is false.	0.8199	1
I doubt this claim. Any official source?	0.8143	1
Can someone confirm with a link? Looks suspicious.	0.7226	1

ISSN: 2692-7608 DOI: 10.6981/FEM.202604_7(4).0007

5. Conclusion and Future Work

5.1. Research Summary

Addressing the time lag in post-hoc debunking, this study validates an early automated intervention framework based on epistemic vigilance. Empirical analysis confirms "Deny" first comments suppress subsequent fake news support ratios to 0.525 ($p < 0.01$) and drive discussion entropy to 0.897. Mapping these theories to NLP tasks, our ELECTRA + Focal Loss model demonstrates superior robustness against imbalanced samples, increasing the high-potential corpus recall rate to 0.898 and validating automated early cognitive guidance feasibility.

5.2. Theoretical and Practical Contributions

Theoretically, this study integrates TDT and epistemic vigilance into computational communication, expanding the SDQC framework from passive detection to proactive intervention design. Reversely applying stance evolution logic provides a new paradigm for "Theory-driven Design Science". Practically, we provide platforms with a scalable algorithmic governance solution, utilizing an asymmetric hard truncation threshold ($\tau = 0.18$) to establish a "high recall priority" screening criterion for blocking initial misinformation cascades.

5.3. Limitations and Future Research Directions

While providing macro-level insights, this study relies on secondary historical data, precluding direct observation of individual cognitive mechanisms. Real social contexts involve confounding variables (e.g., source credibility, UI layout) not fully isolated in current evaluations. Future research must extend to micro-level Randomized Controlled Trials (RCTs) using high-fidelity prototypes. Manipulating AI first-comment stimuli and combining psychometric scales will directly measure the mediating effect of "epistemic vigilance," completing the theoretical and empirical closed loop.

References

[1] Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2), 211-236.

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

[3] Levine, T. R. (2014). Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. Journal of Language and Social Psychology, 33(4), 378–392.

[4] Lee, E. J., & Jang, Y. J. (2010). What do others' reactions to news on Internet portal sites tell us? Effects of presentation format and readers' need for cognition on reality perception. Communication Research, 37(6), 825-846.

[5] Bode, L., & Vraga, E. K. (2015). In related news, that was wrong: The correction of misinformation through related stories functionality in social media. Journal of Communication, 65(4), 619-638.

[6] Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. European Review of Social Psychology, 32(2), 348-384.

[7] Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS ONE, 11(3), e0150989.

[8] Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. ACM Computing Surveys (CSUR), 51(2), 1-36.

[9] Gorrell, G., Kochkina, E., Liakata, M., Aker, A., Zubiaga, A., Lukasik, M., & Bontcheva, K. (2019). SemEval-2019 Task 8: RumourEval 2019: Determining rumour veracity and support for rumours. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 845-854).

ISSN: 2692-7608 DOI: 10.6981/FEM.202604_7(4).0007

[10] Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Wong Sak Hoi, G., & Zubiaga, A. (2017). SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 69-76).

[11] Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. Management Science, 66(11), 4944-4957.

[12] Ma, J., Gao, W., Mitra, P., Kwon, S., Perez, D. J., Wong, K. F., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. IJCAI.

[13] Wu, L., Rao, Y., Zhao, Y., Liang, H., & Nazir, A. (2019). Trace fake news in social media: A unified framework with text, comment and propagation. KDD.

[14] Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. Mind & Language, 25(4), 359–393.

[15] Procter, R., Crump, J., Karstedt, S., Voss, A., & Cantijoch, M. (2013). Reading the riots: what were the police doing on Twitter?. Policing and society, 23(4), 413-436.

[16] Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter Under Crisis: Can we trust what we RT?. In Proceedings of the First Workshop on Social Media Analytics (pp. 71-79).

[17] Qazvinian, V., Rosengren, E., Radev, D. R., & Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (pp. 1589-1599).

[18] Murungi, D. M., Purao, S., & Yates, D. (2018). Beyond facts: A new spin on fake news in the age of social media. In Proceedings of the 24th Americas Conference on Information Systems (AMCIS).

[19] Jin, Y., van der Meer, T. G., Lee, Y. I., & Lu, X. (2020). The effects of corrective communication and employee backup on the effectiveness of fighting crisis misinformation. Public Relations Review, 46(3), 101910.

[20] Paek, H. J., & Hove, T. (2019). Effective strategies for responding to rumors about risks: The case of radiation-contaminated food in South Korea. Public Relations Review, 45(3), 101762.

[21] Oh, O., Agrawal, M., & Rao, H. R. (2013). Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. MIS Quarterly, 37(2), 407-426.

[22] Liu, B. F., Fraustino, J. D., & Jin, Y. (2015). Social media use during disasters: How information form and source influence intended behavioral responses. Communication Research, 42(5), 626-646.

[23] Chua, A. Y., & Banerjee, S. (2017). To share or not to share: The role of epistemic belief in online health rumors. International Journal of Medical Informatics, 97, 108-115.

[24] Shao, C., Ciampaglia, G. L., Varol, O., Yang, K. C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. Nature Communications, 9(1), 4787.

[25] Lee, E. J., & Jang, Y. J. (2010). What do others' reactions to news on Internet portal sites tell us? Effects of presentation format and readers' need for cognition on reality perception. Communication Research, 37(6), 825-846.

[26] Gao, Y., Zhang, M. M., & Lysyakov, M. (2025). Does Social Bot Help Socialize? Evidence from a Microblogging Platform. Information Systems Research. https://doi.org/10.1287/isre.2024.1089

[27] Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. Human-Machine Communication, 1, 71-85.

[28] Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. In 8th International Conference on Learning Representations (ICLR 2020).

[29] Fajcik, M., Smrz, P., & Burget, L. (2019). BUT-FIT at SemEval-2019 Task 7: Determining the rumour stance with pre-trained deep bidirectional transformers. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 1097-1104).

ISSN: 2692-7608 DOI: 10.6981/FEM.202604_7(4).0007

[30] Li, Q., Zhang, Q., & Si, L. (2019). Rumor detection by exploiting user credibility information, attention and multi-task learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 119-129).

[31] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (ICCV) (pp. 2980-2988).