

Yaroslav Lashyn, Oleksandr Voitko, Vitalii Fedoriienko

National Defence University of Ukraine, Kyiv, Ukraine

HYBRID NEURO-MATHEMATICAL MODEL FOR DETECTING INFORMATION THREATS: INTEGRATING LARGE LANGUAGE MODELS AND MODIFIED DIFFUSION PROCESSES

Abstract. Background. In the context of the rapid evolution of cyber threats, where destructive narratives spread with unprecedented speed, traditional isolated monitoring systems lose their effectiveness. Subject of research: analytical systems for detecting coordinated information operations in modern cognitive security. **Purpose of research:** to develop a unified neuro-mathematical architecture that maps non-linear semantic features onto differential equations of spatio-temporal diffusion. **Research objectives.** To construct a hybrid model based on a large language model, transform semantic outputs into continuous kinematic parameters, and empirically validate the model's efficacy during real information attacks. **Research methods.** A mixed-methods empirical approach was employed, integrating quantitative mathematical modeling with computational linguistics, the Llama-3.1-8B-Instruct model utilizing the Adaptive Self-Consistency algorithm, and modifying the Bass diffusion model alongside the Neural Hawkes process. **Research results.** The implementation of thermodynamic logit scaling reduced semantic classification entropy to 0.160 bits. Integrating continuous kinematic parameters into differential equations allowed the system to identify the phase transition of an attack 3 to 4 hours prior to its statistical peak. It was revealed that granular emotional subtype classification without fine-tuning leads to mathematical collapse, entropy surging to 1.349 bits. **Conclusions.** The research expands the Bass diffusion theory and Hawkes processes by replacing static constants with dynamic, AI-derived forcing functions. It mathematically proves the necessity of macro-polarization in feature spaces for stable differential modeling without network fine-tuning.

Keywords: information security; information operations; large language models; natural language processing; social network analysis; hybrid neuro-mathematical model; diffusion processes; stochastic models.

1. Introduction

1.1 Research Background

The contemporary digital landscape is characterized by an unprecedented degree of complexity, wherein sociotechnical information dissemination networks are continuously subjected to coordinated manipulation and cognitive warfare [1]. The transition from scripted, deterministic automated agents to autonomous, cognitively capable social bots has fundamentally altered the parameters of online social interactions. Modern artificial intelligence algorithms, specifically generative multimodal architectures and Large Language Models (LLMs), are increasingly deployed to execute highly sophisticated information operations that exhibit emergent properties such as long-term strategic planning, emotional mimicry, and hyper-personalized persuasion [2]. Furthermore, the democratization of these advanced generative models has enabled malicious actors to turbocharge disinformation operations, creating synthetic content farms that disseminate propaganda across multiple languages with near-zero marginal cost [3].

Under these modern conditions, it is necessary to conduct scientific research on hybrid threat detection models because existing analytical approaches are systematically inadequate. Analyzing text semantics using natural language processing [4, 5] without understanding their kinematic propagation vectors [6, 7] fails to distinguish between organic, user-driven virality and algorithmic, state-sponsored amplification [8]. Conversely, classical mathematical diffusion models [9, 10] rely on static coefficients that mathematically cannot account for the rapid, dynamic interventions characteristic of modern botnets.

In practice, the results of this research can provide security practitioners and threat analysts with a unified

framework to filter out organic social media noise and enable the early detection of artificially amplified cognitive threats. By mathematically transforming discrete linguistic classifications into continuous temporal indicators, the results will allow these cognitive variables to function as dynamic forcing functions within stochastic mass-servicing systems, facilitating the neutralization of bot-driven panic injections prior to mass audience infection [11, 12]. Therefore, research dedicated to the development of integrated hybrid neuro-mathematical models for detecting information threats is highly relevant.

1.2 Literature review and problem statement

In works [4, 8], deep learning and Natural Language Processing (NLP) are utilized to analyze the semantic payload of individual messages. These studies successfully identify discrete elements of state-sponsored propaganda and synthetic text. However, the evaluation of the temporal dimension of information dissemination remains an unresolved issue in these works. This question remains unresolved because the applied semantic transformer architectures inherently analyze texts in a vacuum and lack mechanisms to track network dynamics or propagation vectors over time [13–15].

To mitigate the information entropy of LLMs, works [16] and [17] propose Adaptive Self-Consistency (ASC) decoding algorithms. While this approach improves reasoning on complex tasks by dynamically adjusting the sampling budget, the stable integration of these stochastic classifications into rigid mathematical diffusion models remains unresolved. This issue remains unresolved because standard ASC implementations lack deliberate thermodynamic temperature scaling, which is fundamentally necessary to mathematically bound the variance of the neural network's outputs before feeding them into sensitive differential equations [18, 19].

Works [9, 10], and [20] apply the Bass diffusion model to evaluate the macroscopic kinematic spread of information based on external influence and internal imitation. However, adapting this model to the hyperdynamic environment of modern botnets remains an unresolved problem. This limitation persists because the proposed classical models strictly rely on static coefficients, which mathematically cannot account for the rapid, dynamic interventions and artificial volume injections characteristic of modern cyber-warfare [21, 22].

The microscopic dynamics of information cascades are considered in works [23–25], which propose Neural Hawkes point processes to dynamically learn non-linear intensity functions from temporal data. Nevertheless, evaluating the underlying cognitive payload that drives the organic virality is an unresolved issue in these studies. This question remains unresolved because the employed point processes are mathematically blind to semantic features; they rely exclusively on the metadata of temporal event clustering and ignore the real-time emotional volatility of the content [26].

Systematizing the identified local problems reveals a fundamental methodological gap: semantic models lack a temporal dimension, kinematic diffusion models lack cognitive awareness, and advanced generative algorithms lack the rigid mathematical bounding required for continuous differential systems. Therefore, the general unresolved problem is the systematic inability of existing, isolated analytical frameworks to reliably identify anomalous structural shifts within the state vector of a sociotechnical network during coordinated information operations. Resolving this problem requires the formalization of a non-linear mathematical mapping that systematically projects a highly dimensional space of linguistic features into a continuous state vector of spatio-temporal diffusion.

1.3 Research Aim and Objectives

The aim of this research is to construct and empirically validate a unified hybrid neuro-mathematical architecture that seamlessly integrates tensor-based semantic inference with analytical spatio-temporal diffusion models to detect anomalous structural shifts in sociotechnical networks during coordinated information operations. To achieve this aim, the following objectives were formulated:

- to develop a non-linear semantic contour based on a quantized large language model Llama-3.1-8B-Instruct utilizing a modified Adaptive Self-Consistency algorithm with thermodynamic logit scaling to minimize classification entropy;
- to synthesize continuous early warning indicators, specifically the Sentiment Volatility Coefficient and Bot-Injection Index, by mathematically transforming discrete semantic classifications into dynamic time-series scalars;
- to construct a dynamic differential contour by modifying the classical Bass diffusion model and the Neural Hawkes point process, substituting their static adoption constants with the AI-derived kinematic functions;
- to empirically validate the predictive sensitivity of the proposed hybrid model on real-world data from a live

information operation and determine the mathematical boundary limits of zero-shot generative models in granular emotional classification.

2. Materials and Methods

The object of the research is the sociotechnical information dissemination networks that are continuously subjected to coordinated manipulation and cognitive warfare during information operations. The paper considers the non-linear semantic features and continuous kinematic parameters, formed by generative large language models and diffusion equations, for early detection of anomalous structural shifts within the state vector of a sociotechnical network. The mathematical layer transposes the discrete semantic classifications into continuous temporal sequences to govern stochastic mass-servicing models. The main hypothesis of the research: the implementation of a unified hybrid neuro-mathematical architecture that maps non-linear semantic features onto differential equations of spatial-temporal diffusion will allow for reducing the semantic classification entropy and identifying the phase transition of coordinated cognitive attacks hours before traditional volume-based anomaly detectors. The method is based on the foundational theoretical postulate that modern information warfare possesses an inextricable dual nature: it is simultaneously semantic, comprising cognitive narratives, emotional manipulation, and rhetorical framing, and kinematic (comprising propagation velocity, dissemination channels, and network topology) [7]. This will ensure the early detection of artificially amplified cognitive threats prior to mass audience infection. When developing the hybrid neuro-mathematical method, the following conditions were used:

Condition 1. Organic information diffusion generally traces a smooth cumulative trajectory, whereas coordinated bot injections manifest mathematically as unit impulses, closely approaching the Dirac delta function or Heaviside step function [27].

Condition 2. The rate of adoption in classical diffusion linearly depends on the fraction of the population that has already adopted the information [21].

Condition 3. To maintain mathematical convergence without fine-tuning, the neural feature space is strictly confined to macro-level polarization, simplifying the highly granular emotional subsets that typically cause entropy surges.

Condition 4. The semantic outputs are transformed into continuous kinematic parameters, which dynamically govern the phase-space of a modified Bass diffusion model and a Neural Hawkes process. In the process of developing the method for detecting information threats, a number of various methods and algorithms were applied.

When developing the mathematical model of the Semantic Non-linear Contour, the principle of the Adaptive Self-Consistency decoding algorithm was used [16]. The principle consists in applying thermodynamic temperature scaling to the model's output logits to generate a stochastic ensemble of independent conclusions, diverging from standard greedy decoding. The probability distribution for the generation of the i -th

token (class) is computed utilizing a thermodynamically modified Softmax function:

$$p_i = \frac{\exp(z_i/\tau)}{\sum_j \exp(z_j/\tau)}, \quad (1)$$

where $p_i \in [0, 1]$ represents the normalized probability of selecting the i -th class from the total vocabulary of features, ensuring the sum of all p_i equals 1; $z_i, z_j \in \mathbb{R}$ represent the unscaled logits extracted from the final fully connected layer of the neural network, indicating the raw algorithmic “confidence” in class i or j ; and $\tau \in (0, \infty)$ denotes the generative temperature hyperparameter.

The application of this mathematical principle when building the dynamic process model provides the following advantages:

A detailed mathematical analysis of its boundary conditions reveals that as $\lim_{\tau \rightarrow 0} p_i$, the distribution degenerates into a discrete Dirac delta function (equivalent to $\arg \max$), losing its capacity to navigate semantically ambiguous texts. Conversely, as $\lim_{\tau \rightarrow \infty} p_i = \frac{1}{|V|}$, all exponential terms approach unity, transforming the distribution into absolute uniformity, whereby the system loses classification capability. Therefore, extensive multi-factor experimentation dictated the strict structural bifurcation of the parameter τ . For the classification of narrative and thematic topics, an elevated temperature of $\tau_{topic} = 0.85$ deliberately amplifies the sampling variance, compelling the neural network to traverse peripheral semantic association graphs, which is critical for defeating adversarial attempts to bypass monitoring systems via synonymic masking. For the classification of sentiment, the temperature was strictly compressed to $\tau_{sent} = 0.40$, which forces the probability distribution to cluster tightly around the dominant emotional vector, ensuring absolute convergence among the ensemble's independent inferences.

When quantifying algorithmic certainty in the Semantic Non-linear Contour, the principle of the mathematical confidence coefficient (Confidence margin) was used. The principle consists in calculating the differential between the volume of votes for the triumphant majority class and the closest competing alternative class within a stochastic ensemble. The mathematical margin is formalized as follows:

$$\gamma_i = \frac{\sum_{j=1}^N \mathbb{I}(y_j = c^*) - \max_{k \neq c^*} \sum_{j=1}^N \mathbb{I}(y_j = k)}{N}, \quad (2)$$

where N is the total ensemble size; $\mathbb{I}(\cdot)$ is the discrete indicator function returning 1 if the condition is true and 0 if false; c^* is the triumphant majority class; and k is the closest competing alternative class.

The application of this principle when evaluating algorithmic certainty provides the following advantages: it acts as a vital algorithmic entropy penalty. By purging classification outputs where $\gamma_i < \theta_{conf}$ was empirically optimized to 0.75, it creates an impenetrable safety barrier, preventing the infiltration of semantic noise into the highly sensitive differential equations of the dynamic contour. When computing the constituent elements of the system's macro-state vector S_t , the principle of five distinct mathematical Early Warning Indicators (EWIs),

specifically the hybridized impact function, was used. The principle consists in weighting each source's emission to counteract the obfuscation of social graph topologies and mitigate the impact of automated bot inflation, commonly referred to as Sybil attacks. The hybridized impact function is calculated as:

$$EIS_{it} = \alpha \frac{v_{it}}{\max(v_t)} + \beta \frac{s_{it}}{\max(s_t)}, \quad (3)$$

where v_{it} represents the median number of unique content views for the i -th source; s_{it} denotes the subscriber count of the emission source; $\max(v_t)$ and $\max(s_t)$ are denominators enforcing strict Euclidean normalization within the interval; $\alpha = 0.75$ and $\beta = 0.25$ are specific influence constants.

The application of this principle when computing the impact of sources provides the following advantages:

It enforces the mathematical dominance of kinematic interaction (views) over static topological metrics (subscribers), which is a fundamental prerequisite for gradient stability in downstream neural network layers. This effectively nullifies the efficacy of bot networks where the subscriber count approaches the maximum $s_i \rightarrow \max(s)$ but genuine human engagement approaches zero $v_i \rightarrow 0$. When isolating kinematic anomalies, the principle of the Early Amplification Ratio (EAR) was used [27]. The principle consists in computing the area under the intensity curve strictly during the initial hours of an attack's deployment, distinguishing coordinated bot injections, which manifest mathematically as unit impulses, closely approaching the Dirac delta function or Heaviside step function from organic information diffusion, which generally traces a smooth cumulative trajectory.

$$EAR(t_0, \Delta t) = \frac{\int_{t_0}^{t_0 + \Delta t} V_{topic}(t) dt}{\mu_V \cdot \Delta t}, \quad (4)$$

where $V_{topic}(t)$ represents the definite integral of content generation velocity, messages per hour, over a highly restricted analytical window $\Delta t \leq 2$ hours, computed via numerical integration methods; $\mu_V \cdot \Delta t$ represents the historical mathematical expectation of background intensity for the given narrative under non-anomalous conditions.

The application of this principle when isolating kinematic anomalies provides the following advantages:

An $EAR \geq 3.0$ algorithmic threshold signifies a 300% deviation from expected statistical behavior, immediately flagging a critical kinetic anomaly. When formalizing the sociological phenomenon of emotional manipulation – specifically the deliberate injection of panic, fear, or unbridled aggression into an otherwise neutral information stream – the principle of the Sentiment Volatility Coefficient (SVC) was used. The principle consists in measuring the standardized emotional dispersion:

$$SVC_t = \sqrt{\frac{1}{|\mathcal{M}_t|} \sum_{i=1}^{|\mathcal{M}_t|} (\sigma_i - \bar{\sigma}_t)^2}, \quad (5)$$

where $|\mathcal{M}_t|$ is the cardinality of the message set within the temporal window t ; $\sigma_i \in \{-1, 0, 1\}$ is the discrete emotional polarity derived from the semantic contour;

$\bar{\sigma}_t$ is the moving mathematical expectation of the sentiment for the set. The application of this principle when measuring emotional dispersion provides the following advantages:

The utilization of standard deviation over absolute deviation is mathematically vital; squaring the differentials exponentially amplifies the systemic weight of extreme emotional outliers, rendering the metric highly sensitive to artificial emotional injections. When inextricably linking the kinematic anomaly EAR with the cognitive anomaly SVC into a singular, unified threat marker, the principle of the Bot-Injection Index (BII) was used. The principle consists in calculating a synthetic, non-linear, multi-dimensional scalar utilizing the hyperbolic tangent function:

$$BII_t = \tanh\left(w_{ear} \cdot EAR_t + w_{svc} \cdot \frac{SVC_t}{SVC_{norm}}\right), \quad (6)$$

where $w_{ear} = 0.15$ and $w_{svc} = 0.35$ are weighting coefficients empirically optimized via gradient descent methodologies on historical datasets of prior information operations; SVC_{norm} represents a constant of normal background volatility, required to bring disparate multidimensional values to a common denominator. The application of this principle when linking anomalies provides the following advantages:

The integration of the hyperbolic tangent function constitutes a critically important architectural safeguard that explicitly eliminates the pervasive problem of gradient explosion when these metrics are subsequently fed into differential equations. It guarantees the compression of potentially infinite inputs, massive bot spam causing $EAR \rightarrow 100$, into a highly stable, bounded computational space of $[0, 1)$, ensuring the computational resilience of the overall system. When concentrating the model's computational resources exclusively on the destructive vector of the data stream, the principle of the Thematic Risk Factor (TRF) was used. The principle consists in acting as a complex spatial filter by aggregating the aforementioned scalars:

$$TRF_t = \sum_{m_i \in \mathcal{M}_t} EIS_{it} \cdot \mathbb{I}(\theta_i \in \Theta_{destruct}) \cdot \gamma_i, \quad (7)$$

where $\mathbb{I}(\theta_i \in \Theta_{destruct})$ is an indicator function that strictly zeroes out any message where the identified topic θ_i does not belong to the set of predefined destructive narratives $\Theta_{destruct}$; γ_i ensures that classifications with low neural certainty contribute proportionally less to the total accumulated risk. The application of this principle when filtering the data stream provides the following advantages:

It completely eliminates organic, neutral noise from the risk calculation, accelerating system response time. When evaluating the persistence of an operation, the principle of the geometric Campaign Longevity metric (NTP) was used. The principle consists in calculating the definite integral of the area where the risk exceeds the background norm:

$$NTP = \int_{T_{obs}} \max(0, TRF_t - \mu_{TRF}) dt, \quad (8)$$

where: TRF_t is the Thematic Risk Factor; μ_{TRF} is the background norm; and $\max(0, \dots)$ is a non-linear

operation functioning identically to a Rectified Linear Unit (ReLU) activation within neural networks, truncating all sub-threshold risk fluctuations to zero. The application of this principle when evaluating operation persistence provides the following advantages:

It ensures the integral accumulates area exclusively above the baseline anomaly threshold, allowing analysts to mathematically differentiate transient, organic information noise from sustained, highly resourced cognitive campaigns executed by state-sponsored actors. When achieving the mathematical transformation of classical, isotropic diffusion models into dynamic, predictive systems, the principle of non-linear dynamic modification of the Bass diffusion model was used [20]. The principle consists in replacing static constants with dynamic control functions derived directly from the cognitive-kinematic EWIs. The coefficient of external injection p_t^* is calculated as:

$$p_t^* = p_0 \cdot (1 + \lambda_p \cdot BII_t \cdot \overline{EIS}_t), \quad (9)$$

where p_0 is the base innovation rate; λ_p is a hyperparameter dictating sensitivity to bot attacks; $\overline{EIS}_t$ represents high-impact channels; and BII_t represents automated spam.

The application of this principle when modifying external injection provides the following advantages:

It mathematically simulates the real-world operational tactics of bot-farms, which utilize high-impact channels and automated spam to artificially inflate the external visibility of a narrative, driving an exponential surge in p_t^* . Concurrently, when calculating the coefficient of social imitation, the principle of emotional resonance function was used. The principle consists in mathematically modeling the proven sociological phenomenon that high-volatility content, inducing fear or anger, spreads horizontally at exponential rates [28]:

$$q_t^* = q_0 \cdot \left(1 + \lambda_q \cdot \max\left(0, \frac{SVC_t}{SVC_{norm}} - 1\right)\right), \quad (10)$$

where q_0 is the base organic imitation rate; λ_q reflects the audience's sensitivity to emotional fluctuations.

The application of this principle when modifying social imitation provides the following advantages:

If the emotional dispersion SVC_t spikes relative to the norm due to the injection of panic-inducing narratives, the multiplier becomes positive, mathematically accelerating the network's inherent horizontal virality. When mapping the microscopic topology of information cascades, the exact sequence of individual retweets over time, the principle of self-exciting Hawkes processes was used [23]. The principle consists in exponentially amplifying the baseline intensity of the point process by the Thematic Risk Factor derived from the LLM contour:

$$\lambda(t) = \mu_0 \cdot \exp(\alpha \cdot TRF_t) + \sum_{t_i < t} \phi(t - t_i), \quad (11)$$

where $\lambda(t)$ is the conditional intensity of the Poisson point process, the probability of a new publication in the next microsecond; $\phi(\tau) = \kappa \cdot e^{-\omega\tau}$ is the decay kernel representing the viral echo effect of all previous events t_i

[25]. The application of this principle when mapping information cascades provides the following advantages:

By embedding the TRF_t scalar directly into the exponent of the base intensity, the system generates a critical mathematical alert indicating an impending structural phase shift, the transition to mass “Attack” phase 2 to 4 hours prior to full operational deployment, significantly outperforming traditional delayed-response models.

3. Results of the Development and Empirical Validation of the Hybrid Neuro-Mathematical Model

3.1 Development of the Semantic Non-linear Contour and Entropy Minimization

To evaluate the efficacy of the proposed method and achieve complete scientific validation of the developed neuro-mathematical apparatus, an extensive computational experiment was executed utilizing real-world data. The empirical dataset comprised $N = 999$ lemmatized messages extracted from $K = 50$ high-density Telegram channels over an observation window of $\Delta t = 24$ hours, meticulously capturing the live deployment and dissemination of a verified information operation. Comprehensive data preprocessing protocols were applied, including the removal of stop words and strict tokenization constrained to a maximum of 512 tokens to

ensure complete compatibility with the LLM’s context window. Following the execution of the semantic contour utilizing the Adaptive Self-Consistency algorithm on the quantized Llama-3.1 model, the system successfully categorized 711 messages as neutral, 20 as positive, and 268 as destructive narratives $P_{neg} = 26.8\%$. A parallel unsupervised topic modeling process (utilizing BERTopic) independently corroborated these findings, isolating four robust topical clusters with an exceptionally high Adjusted Rand Index $ARI = 0.925$. The implementation of the thermodynamic logit scaling mechanism yielded an unprecedented reduction in classification entropy, which plummeted to $H_{topic} = 0.160$ bits, accompanied by a high mean algorithmic confidence margin of $\gamma = 0.878$. This definitively established the efficacy of the proposed thermodynamic technique in neutralizing the semantic uncertainty that typically plagues zero-shot LLM inference.

3.2 Synthesis of Continuous Early Warning Indicators

The continuous aggregation of these validated semantic metrics into the macro-state vector S_t , with an hourly discretization step $\Delta t = 1$ hr, provided empirical proof of the model’s extreme predictive sensitivity. The results of the modeling for these indicators are presented in Table 1.

Table 1 – Fragment of the calculation matrix of IO macro-indicators

Temporal Slice (t_i)	Intensity (V)	Sentiment ($\bar{\sigma}$)	Volatility SVC_t	EAR_{t_0}	Bot Index BII_t	Risk Factor TRF_t	Z-score TRF
Background t_{-1}	11	-0.15	0.34	1.0	0.05	1.45	0.12
Trigger t_0	48	-0.42	0.68	4.36	0.874	4.8	2.95
Attack t_1	95	-0.65	0.82	8.63	0.96	11.2	4.8

An exhaustive analysis of the mathematical behaviors observed during the critical phase transition at t_0 , the preparatory “Trigger” phase, illustrates the core superiority of the hybrid framework. The absolute volume metric $V = 48$ messages/hour remained insufficiently high to breach the static thresholds of conventional monitoring systems, which would have inevitably resulted in a false negative. However, the advanced kinematic apparatus detected severe localized deviations: the EAR surged to 4.36, signifying an immediate, non-organic injection of volume that was over four times the expected background rate. Concurrently, the emotional dispersion SVC experienced a 100% amplification jumping from a baseline of 0.34 to 0.68, mathematically confirming the artificial induction of panic and aggression into the information flow. Inserting these discrete scalars into the normalized BII formula yielded irrefutable mathematical evidence of algorithmic manipulation:

$$BII_{t_0} = \tanh\left(0.15 \cdot 4.36 + 0.35 \cdot \frac{0.68}{0.34}\right) \approx 0.874. \quad (12)$$

3.3 Construction of the Dynamic Differential Contour

The BII value of 0.874, an asymptotic push toward the theoretical limit of 1.0, alongside a standardized Z-score of 2.95 for the TRF, comfortably exceeding the critical 2.0σ anomaly threshold, triggered the

immediate, non-linear exponential scaling of the innovation parameter p_t^* in the modified Bass equation. This explicit neural-to-differential linkage enabled the system’s detection module to mathematically declare a phase transition to the “Attack” state 3 to 4 hours prior to the actual empirical peak volume occurring at t_1 .

3.4 Empirical Validation and Mathematical Limits of Generative Models

A secondary, yet equally profound result of the experiment was the mathematical documentation of the fundamental boundary limits inherent in contemporary artificial intelligence paradigms. During an extended experimental phase, the semantic feature space Σ was expanded beyond basic macro-polarization: positive, negative and neutral to encompass highly granular emotional subsets: anger, fear, anxiety, sorrow. The ensuing statistical analysis recorded an absolute collapse in semantic inference quality. The mean information entropy H for granular sentiment spiked exponentially to a critical 1.349 bits – dangerously approaching the maximum theoretical entropy limit for a 3 to 4 class distribution – while the ensemble confidence margin collapsed to a marginal $\gamma = 0.292$. The semantic space experienced severe divergence; the LLM ensemble began outputting near-uniform probability distributions, entirely incapable of distinguishing between adjacent

psycho-emotional states, confusing raw geopolitical anger with general hate speech. This empirically proved that zero-shot generative architectures are structurally incapable of executing stable, highly granular emotional classifications on complex texts without extensive, parametric fine-tuning, thereby defining the absolute operational limits of the hybrid model.

4. Discussion of the Results of the Hybrid Neuro-Mathematical Model Empirical Validation

The proposed hybrid neuro-mathematical model of the information threat detection process. The specific feature of this model is its dual-nature architecture, integrating semantic non-linear filtering with dynamic differential diffusion equations. The criteria and quality indicators of the threat detection process were determined. The specific feature of the proposed Early Warning Indicators (EWIs) is the application of continuous kinematic scalar aggregation, formulas for EAR, SVC, and BII. This allowed for evaluating the absolute mathematical limits of generative models and formulating the dynamic control parameters for the modified Bass diffusion model and Neural Hawkes process.

The evaluation of the proposed model's efficacy on Table 1, showed the following results: the implementation of thermodynamic logit scaling yielded an unprecedented reduction in classification entropy, which plummeted to $H_{topic} = 0.160$ bits, accompanied by a high mean algorithmic confidence margin of $\gamma = 0.878$. During the critical phase transition, the kinematic apparatus detected an EAR surge to 4.36 and a 100% amplification in emotional dispersion SVC, which mathematically confirmed the artificial induction of panic. The explicit neural-to-differential linkage enabled the system's detection module to mathematically declare a phase transition to the "Attack" state 3 to 4 hours prior to the actual empirical peak volume. Expanding the semantic feature space beyond basic macro-polarization to encompass highly granular emotional subsets without fine-tuning resulted in an absolute collapse in semantic inference quality, with entropy spiking to a critical 1.349 bits. The obtained results of the proposed model research can be explained by the application of thermodynamic stabilization and the integration of the hyperbolic tangent function within the BII metric, which explicitly eliminates the pervasive problem of gradient explosion when these metrics are fed into differential equations. In contrast to [4, 5], where deep learning sequence models or standard classification algorithms are proposed to parse complex disinformation, the proposed model incorporates the temporal dimension necessary to forecast the velocity and trajectory of information propagation. In addition, the proposed model, unlike [6], transforms rigorous semantic assessments into continuous kinetic variables, resolving the core deficiency of pure mathematical epidemiology when applied to highly volatile social networks.

In contrast to [27], where traditional systems remain perpetually reactive by isolating the cognitive payload but ignoring the dissemination physics, the proposed model guarantees system stability under duress by

mathematically bounding the parameters. Also, unlike, where it is assumed that massive, pre-trained models possess generalized psychological profiling capabilities out-of-the-box, the proposed model mathematically proves that zero-shot generative architectures are structurally incapable of executing stable, highly granular emotional classifications on complex texts without extensive parametric fine-tuning. Therefore, the developed hybrid neuro-mathematical model allowed for the early detection and mitigation of information anomalies long before they reach critical mass within the target population. This made it possible to filter out organic social media noise and mathematically differentiate transient information noise from sustained cognitive campaigns. But the proposed results should be applied under the following limitations:

- the application of the model is strictly confined to macro-level polarization of the neural feature space;
- the presence of a mathematically sound thermodynamic temperature scaling setup tailored for specific classification tasks: topic and sentiment.

In addition, the proposed model is expedient to use if the text corpus is sufficiently large and the temporal window provides a reliable historical background intensity baseline. As a drawback of this study, it should be noted that the paper lacks an analysis of the model's performance when integrated with specialized Low-Rank Adaptation (LoRA) or extensive supervised fine-tuning protocols. To eliminate this drawback, the hybrid model needs to be expanded with an additional procedure to align the model's latent space with the specific nuances of granular emotional classification in the operational environment. As a development of this research, the following can be noted.

First, it is necessary to investigate the integration of advanced LoRA techniques to expand the emotional granularity of the semantic contour while maintaining the convergence of the dynamic differential equations.

Second, it is necessary to conduct research regarding the dynamic optimization of the weighting coefficients w_{ear} and w_{svc} using automated machine learning (AutoML) frameworks across diverse types of sociotechnical platforms beyond Telegram.

5. Conclusions

1. The proposed mathematical model of the information threat detection process. The specific feature of this model is its dual-nature architecture, integrating semantic non-linear filtering with dynamic differential diffusion equations. The application of Adaptive Self-Consistency decoding with thermodynamic logit scaling yielded an unprecedented reduction in classification entropy. This allowed for formulating strict boundary conditions to prevent the collapse of coupled differential systems and minimize semantic uncertainty.

2. The criteria and quality indicators of the threat detection process were determined. The specific feature of the proposed Early Warning Indicators is the application of continuous kinematic scalar aggregation. This allowed for assessing the temporal dimension of information dissemination and formulating continuous dynamic control functions for the Bass diffusion equation and Neural Hawkes point process.

3. The developed unified hybrid neuro-mathematical architecture seamlessly integrates tensor-based semantic inference with analytical spatio-temporal diffusion models. In the first stage, the Llama-3.1-8B-Instruct model resolves semantic ambiguity via the ASC algorithm; in the second stage, linguistic outputs are transformed into dynamic early warning indicators. The results allowed for evaluating the efficacy of the developed method. During empirical testing, the method significantly outperformed traditional detectors by accurately identifying the phase shift of cognitive attacks hours before the empirical peak volume. But, expanding the feature space to highly granular emotional subsets resulted in an absolute collapse in inference quality, establishing the mathematical limitations of zero-shot models without fine-tuning.

Conflicts of interest

The authors declare that they have no conflicts of interest regarding this research, including financial, personal, authorship, or any other nature that could influence the research and the results presented in this article.

Funding

The research was conducted without financial support.

Data Availability

The manuscript has no associated data.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies in the creation of the presented work.

REFERENCES

1. Parmar, M. (2026), "Safety, security, and cognitive risks in world models", *arXiv*, arXiv.2604.01346, doi: <https://doi.org/10.48550/arXiv.2604.01346>
2. Li, Y., Yang, Y., and Huang, H. (2025), "Advances in social bots [Editorial]", *Electronics*, vol. 14, no. 24, 4885, doi: <https://doi.org/10.3390/electronics14244885>
3. Sadeghi, M., Dimitriadis, D., Padovese, V., Pozzi, G., Badilini, S., Vercellone, C., Huet, N., Fishman, Z., Pfaller, L., and Adams, N. (2026), "Tracking AI-enabled misinformation: 3,749 AI content farm sites (and counting), plus the top false claims generated by artificial intelligence tools", *NewsGuard*, available at: <https://www.newsguardtech.com/special-reports/ai-tracking-center>
4. Sabri, M. I. F., Hettiarachchi, H., and Ranasinghe, T. (2025), "Detecting fake news in the era of language models", *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing*, pp. 1036–1043, available at: <https://aclanthology.org/2025.ranlp-1.119/>
5. Burghardt, K., Chen, K., and Lerman, K. (2024), "Large language models reveal information operation goals, tactics, and narrative frames", *Proceedings of the ICWSM Workshops*, doi: <https://doi.org/10.48550/arXiv.2405.03688>
6. Stanovov, V., Grabljevec, S., Akhmedova, S., Semenkin, E., Stojanović, R., Rozman, Č., and Škraba, A. (2022), "Identification of COVID-19 spread mechanisms based on first-wave data, simulation models, and evolutionary algorithms", *PLOS One*, doi: <https://doi.org/10.1371/journal.pone.0279427>
7. Whiteside, R. D. (2025), "Narrative manipulation, malinfluence operations, and cognitive warfare through large language model poisoning with adversarial noise", *Military Intelligence Professional Bulletin*, available at: <https://mipb.ikn.army.mil/>
8. Ezzeddine, F., Ayoub, O., Giordano, S., Nogara, G., Sbeity, I., Ferrara, E., and Luceri, L. (2023), "Exposing influence campaigns in the age of LLMs: A behavioral-based AI approach to detecting state-sponsored trolls", *EPJ Data Science*, vol. 12, no. 1, 46, doi: <https://doi.org/10.1140/epids/s13688-023-00423-4>
9. Cosguner, K., and Seetharaman, P. B. (2022), "Dynamic pricing for new products using a utility-based generalization of the Bass diffusion model", *Management Science*, vol. 68, no. 3, pp. 1904–1922, doi: <https://doi.org/10.1287/mnsc.2021.4257>
10. Liang, L. (2021), "Quantitative analysis of diffusion of public opinion by the modified Bass model for the dyeing steamed bread incident in China", *Saudi Journal of Business and Management Studies*, vol. 6, no. 1, pp. 24–27, doi: <https://doi.org/10.36348/sjbms.2021.v06i01.004>
11. Ask, T. F., Sütterlin, S., Müller, L., Lugo, R., Østby, L., Jensen, M., Nordheim, K., Gjøvik, J., and Lervik, E. (2025), "Cognitive security: The study and practice of protecting the human mind and other cognitive assets from cognitive threats", *OSF*, doi: https://doi.org/10.31234/osf.io/2ftqc_v1
12. Stella, M., Ferrara, E., and De Domenico, M. (2018), "Bots increase exposure to negative and inflammatory content in online social systems", *Proceedings of the National Academy of Sciences*, vol. 115, no. 49, pp. 12435–12440, doi: <https://doi.org/10.1073/pnas.1803470115>
13. Munir, B., Abbasi, M. Z., Wilson, W. B., and Colombo, A., Jr. (2025), "Evaluating AI in legal operations: A comparative analysis of accuracy, completeness, and hallucinations in ChatGPT-4, Copilot, DeepSeek, Lexis+ AI, and Llama 3", *International Journal of Legal Information*, vol. 53, no. 2, pp. 103–114, doi: <https://doi.org/10.1017/iji.2025.10052>
14. Barbera, I. (2025), *AI privacy risks & mitigations - large language models (LLMs). Practical guide for conducting risk assessments to mitigate privacy and data protection risks in systems using LLMs*, European Data Protection Board Support Pool of Experts, available at: <https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf>
15. Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., and Hamin, M. (2025), *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations* (NIST AI 100-2e2025), National Institute of Standards and Technology, doi: <https://doi.org/10.6028/NIST.AI.100-2e2025>
16. Jeong, E., and Choi, Y. S. (2025), "ACR: Adaptive confidence re-scoring for reliable answer selection among multiple candidates", *Applied Sciences*, vol. 15, no. 17, 9587, doi: <https://doi.org/10.3390/app15179587>
17. Alcaraz, A. (2023), *Achieving greater self-consistency in large language models*, Towards Data Science, available at: <https://towardsdatascience.com/achieving-greater-self-consistency-in-large-language-models-6e6cb5f3c5b7>
18. Xie, Z., Guo, J., Yu, T., and Li, S. (2024), *Calibrating reasoning in language models with internal consistency*, NeurIPS 2024, available at: <https://neurips.cc/virtual/2024/poster/93260>
19. Taherinezhad, F., Momeni Nezhad, M. J., Karimi, S., et. al (2026), "Large language model adaptation strategies in speech-based cognitive screening: Systematic evaluation", *JMIR AI*, vol. 5, e82608, doi: <https://doi.org/10.2196/82608>

20. Bass, F. M., Krishnan, T. V., and Jain, D. C. (1994), "Why the Bass model fits without decision variables", *Marketing Science*, vol. 13, no. 3, pp. 203–223, doi: <https://doi.org/10.1287/mksc.13.3.203>
21. Manshadi, V., Johari, R., and Misra, S. (2015), *A modified Bass model for product growth in social networks*, available at: https://www-2.rotman.utoronto.ca/MSOM-2015/Documents/MSOM2015_0325.pdf
22. Baek, J., Farias, V. F., Georgescu, A., Levi, R., Peng, T., Sinha, D., Wilde, J. T., and Zheng, A. T. (2025), "The limits to learning a diffusion model", *Management Science*, doi: <https://doi.org/10.1287/mnsc.2022.02953>
23. Laub, P. J., Lee, Y., Pollett, P. K., and Taimre, T. (2025), "Hawkes models and their applications", *Annual Review of Statistics and Its Application*, vol. 12, pp. 233–258, doi: <https://doi.org/10.1146/annurev-statistics-112723-034304>
24. Li, J., Yang, Y., Zhang, Y., Hu, Q., Zhao, A., and Gao, H. (2025), "Public opinion field effect and Hawkes process join hands for information popularity prediction", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 11, pp. 12076–12083, doi: <https://doi.org/10.1609/aaai.v39i11.33315>
25. Okawa, M., Iwata, T., Tanaka, Y., Toda, H., Kurashima, T., and Kashima, H. (2021), "Dynamic Hawkes processes for discovering time-evolving communities' states behind diffusion processes", *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1276–1286, doi: <https://doi.org/10.1145/3447548.3467248>
26. Zeng, F., and Gao, W. (2022), "Early rumor detection using neural Hawkes process with a new benchmark dataset", *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4105–4117, doi: <https://doi.org/10.18653/v1/2022.naacl-main.302>
27. Kuntur, S., Wróblewska, A., Paprzycki, M., and Ganzha, M. (2025), "Under the influence: A survey of large language models in fake news detection", *IEEE Transactions on AI*, vol. 6, no. 2, pp. 458–476, doi: <https://doi.org/10.1109/TAI.2024.3471735>
28. Eryarsoy, E., Delen, D., Davazdahemami, B., and Topuz, K. (2021), "A novel diffusion-based model for estimating cases, and fatalities in epidemics: The case of COVID-19", *Journal of Business Research*, vol. 124, pp. 163–178, doi: <https://doi.org/10.1016/j.jbusres.2020.11.054>

Received (Надійшла) 21.06.2026

Accepted for publication (Прийнята до друку) 02.09.2026

Publication date (Дата публікації) 25.09.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Лашин Ярослав Олександрович – ад'юнкт кафедри соціальної комунікації та публічної дипломатії, Інститут стратегічних комунікацій, Національний університет оборони України, Київ, Україна;

Yaroslav Lashyn – Ph. D. Student at the Department of Social Communication and Public Diplomacy, Institute of Strategic Communications, National Defence University of Ukraine, Kyiv, Ukraine;

e-mail: y.lashin@edu.nuou.org.ua; ORCID Author ID: <https://orcid.org/0009-0000-5533-9210>;

Войтко Олександр Володимирович – доктор військових наук, доцент, начальник інституту стратегічних комунікацій, Національний університет оборони України, Київ, Україна;

Oleksandr Voitko – Doctor of Military Sciences, Associate Professor/Head of the Institute of Strategic Communications, National Defence University of Ukraine, Kyiv, Ukraine;

e-mail: o.voitko@edu.nuou.org.ua; ORCID Author ID: <https://orcid.org/0000-0002-4610-4476>;

Scopus Author ID: <https://www.scopus.com/authid/detail.uri?authorId=57210362201>.

Федорієнко Віталій Анатолійович – кандидат технічних наук, професор кафедри соціальної комунікації та публічної дипломатії, Інститут стратегічних комунікацій, Національний університет оборони України, Київ, Україна;

Vitalii Fedorienko – PhD, Professor at the Department of Social Communication and Public Diplomacy, Institute of Strategic Communications, National Defence University of Ukraine, Kyiv, Ukraine;

e-mail: v.fedorienko@edu.nuou.org.ua; ORCID Author ID: <https://orcid.org/0000-0002-0921-3390>;

Scopus Author ID: <https://www.scopus.com/authid/detail.uri?authorId=57221718554>.

Гібридна нейроматематична модель виявлення інформаційних загроз: інтеграція великих мовних моделей та модифікованих дифузійних процесів

Я. О. Лашин, О. В. Войтко, В. А. Федорієнко

Анотація. Актуальність. У контексті стрімкої еволюції кіберзагроз, коли деструктивні нарративи поширюються з безпрецедентною швидкістю, традиційні ізольовані системи моніторингу втрачають ефективність. **Предмет дослідження:** аналітичні системи виявлення скоординованих інформаційних операцій у сучасному середовищі когнітивної безпеки. **Мета дослідження:** розробити єдину нейроматематичну архітектуру, яка відображає нелінійні семантичні ознаки за допомогою диференціальних рівнянь просторово-часової дифузії. **Завдання дослідження:** побудувати гібридну модель на основі великої мовної моделі, перетворити семантичні вихідні дані на неперервні кінематичні параметри та емпірично оцінити ефективність моделі під час реальних інформаційних атак. **Методи дослідження:** застосовано змішаний емпіричний підхід, що поєднує кількісне математичне моделювання, комп'ютерну лінгвістику, модель Llama-3.1-8B-Instruct з використанням алгоритму Adaptive Self-Consistency, а також модифіковану модель дифузії Басса та нейронний процес Гокса. **Результати дослідження:** застосування термодинамічного логіт-масштабування знизило ентропію семантичної класифікації до 0,160 біта. Інтеграція неперервних кінематичних параметрів у диференціальні рівняння дала змогу системі виявити фазовий перехід інформаційної атаки за 3–4 години до досягнення її статистичного піку. Встановлено, що детальна класифікація емоційних підтипів без тонкого налаштування моделі призводить до математичної нестабільності, унаслідок чого ентропія зростає до 1,349 біта. **Висновки:** дослідження розширює теорію дифузії Басса та процеси Гокса шляхом заміни статичних констант динамічними функціями впливу, отриманими за допомогою штучного інтелекту. Математично обґрунтовано необхідність макрополяризації у просторі ознак для забезпечення стабільного диференціального моделювання без тонкого налаштування нейронної мережі.

Ключові слова: інформаційна безпека; інформаційні операції; великі мовні моделі; оброблення природної мови; аналіз соціальних мереж; гібридна нейроматематична модель; дифузійні процеси; стохастичні моделі.