УДК: 342.738+004.056.5:004.738.5 DOI 10.18372/2073-4751.87.21695

Ткаченко О. С.,
orcid.org/0009-0008-1721-3455,
e-mail: alexsunnik@gmail.com,
Ільєнко А. В.,
orcid.org/0000-0001-8565-1117,
e-mail: anna.ilienko@npp.kai.edu.ua

МАТЕМАТИЧНА МОДЕЛЬ ПОШИРЕННЯ ІНФОРМАЦІЙНИХ ВПЛИВІВ У СОЦІАЛЬНИХ МЕРЕЖАХ: КОМПЛЕКСНИЙ ПІДХІД ДО МОДЕЛЮВАННЯ ДЕЗІНФОРМАЦІЇ ТА АНТИДЕЗІНФОРМАЦІЇ

Національний університет «Київський авіаційний інститут»

Вступ і постановка проблеми

Сучасні соціальні мережі стали потужними платформами для розповсюдження інформації, що має як позитивний, так і негативний вплив на суспільство. Дезінформація поширюється з небувалою швидкістю, впливаючи на громадську думку, політичні процеси та соціальну стабільність, а в сучасних умовах інформаційне протиборство стало ще й одним із ключових компонентів кіберпростору. Маніпулятивний контент та скоординовані інформаційні кампанії нині розглядаються як різновид серйозних загроз, що можуть підривати стійкість державних інституцій, критичної інформаційної інфраструктури та суспільної безпеки. На відміну від традиційних кібератак, інформаційні атаки спрямовані насамперед на когнітивну сферу користувачів, що суттєво ускладнює їх своєчасне виявлення та прогнозування. Водночас антидезінформаційні заходи також використовують ті самі мережеві механізми для протидії неправдивій інформації [1, 2].

Хоча традиційні епідеміологічні моделі, такі як SIR (Susceptible-Infected-Recovered), були адаптовані для моделювання поширення інформації, вони не повною мірою враховують складність соціальних мереж та специфіку інформаційних процесів. Відповідно, особливої актуальності набуває розроблення більш комплексних математичних моделей, які б інтегрували стохастичні процеси, мережеву топологію та психологічні аспекти користувачів [3, 4]. Такі моделі здатні формалізувати закономірності поширення інформаційних впливів, визначати критичні умови розвитку інформаційних атак та оцінювати ефективність механізмів їх локалізації. Вони є важливим елементом систем моніторингу інформаційного простору, аналізу загроз та підтримки прийняття рішень щодо вибору заходів протидії дезінформації. Загалом, використання математичного моделювання дозволяє не лише досліджувати динаміку поширення впливів, а й прогнозувати розвиток інформаційних атак, визначати критичні порогові стани системи, оцінювати стійкість середовища до зовнішніх впливів та аналізувати результативність організаційних і технічних контрзаходів. Саме тому розроблення математичних моделей такого типу є одним із найперспективніших напрямів розвитку сучасних систем забезпечення інформаційної та кібербезпеки.

Метою роботи є розроблення комплексної математичної моделі поширення інформаційних впливів у соціальних мережах, яка забезпечує формалізований опис процесів поширення дезінформації та антидезінформації, враховує особливості мережевої взаємодії

користувачів, стохастичний характер інформаційних процесів і механізми активної протидії інформаційним загрозам. Запропонована модель призначена для прогнозування розвитку інформаційних атак, визначення критичних умов їх масового поширення, оцінювання ефективності контрзаходів та підтримки прийняття рішень щодо забезпечення інформаційної безпеки соціальних мереж.

Огляд досліджень

1. Класичні епідеміологічні моделі в інформаційному контексті. Дослідники адаптували моделі типу SIR, SEIR, SIRS для моделювання поширення інформації в соціальних мережах, враховуючи стадії сприйняття, інфікування та відновлення. Ці моделі створили базу для подальшого розширення, включаючи вплив мережевої топології та психологічних аспектів користувачів [5].

2. Стохастичні моделі та порогові механізми. Було виявлено важливість стохастичних процесів у моделюванні поведінки користувачів, зокрема через випадкові зміни індивідуальних порогів активації та часових ефектів релевантності. Порогові моделі дають змогу моделювати соціальні впливи, які запускають або зупиняють процеси поширення [6].

3. Вплив мережевої топології. Дослідження підтверджують, що структура соціальних мереж (складність, кластеризація, розподіл степенів) суттєво впливає на динаміку поширення інформації та дезінформації. Сегментація мережі породжує локальні "ехо-камери", підсилюючи поширення неправдивої інформації [7, 8].

4. Моделювання дезінформації та антидезінформації. Інноваційні підходи, зокрема моделі SEIDRA, включають не тільки стадії інфікування, а й активне протистояння дезінформації через поширення фактчекінгу та антивпливу. Вивчено ефективність комбінованих стратегій протидії [5].

5. Програмні рішення та алгоритмічні підходи. Існують програмні платформи для симуляції інформаційного середовища з урахуванням мережевих властивостей і динаміки станів. Здійснюється інтеграція методів штучного інтелекту для автоматичної детекції та корекції інформаційних потоків [9].

Огляд програмних рішень

1. Платформи симуляції інформаційних потоків на основі SIR/SEIR моделей з розширеннями для врахування мережевої топології (наприклад, моделі використані у дослідженнях [3, 4]).

2. Програмні пакети зі стохастичними моделями поширення інформації, що враховують випадковість та індивідуальні пороги (застосовувані в роботах [6]).

3. Інструменти для аналізу великих соціальних мереж та виявлення ключових вузлів для інтервенції (описані у [10]).

4. AI-системи для автоматичного виявлення фейкових новин та створення антивпливу, що інтегруються як частина моделі [11].

Теоретичні основи моделі

Фундаментальні принципи

Наша розширена модель базується на кількох ключових теоретичних принципах, що випливають з аналізу літератури [7, 12]:

У соціальних мережах кожне повідомлення має двоїсту природу – з одного боку, воно характеризується фактичною достовірністю, а з іншого – емоційним забарвленням. Саме цей поділ на «правдивість» і «емоційний резонанс» визначає, наскільки активно інформація поширюється онлайн. Дослідження останніх років показують: дезінформація часто компенсує низьку достовірність сильним емоційним впливом. Зокрема, McLoughlin та співавт. (2024) встановили, що «соціальні мережі демонструють: дописи з дезінформацією викликають більше морального обурення, ніж дописи з

перевіреною інформацією, і саме це обурення сприяє поширенню неправдивих відомостей» [12]. Іншими словами, користувачі охоче діляться «скандальним» чи гнівним контентом, навіть не перевіряючи його на правдивість. Мотиви такого поширення – бажання продемонструвати власну моральну позицію чи приналежність до групи однодумців. До того ж алгоритми соціальних платформ підсилюють ефект: контент із сильним емоційним забарвленням (наприклад, що викликає обурення) частіше потрапляє в поле зору інших користувачів [12]. Таким чином, у «подвійній природі» інформації емоції відіграють критичну роль – контент із невисокою фактичною достовірністю може швидко поширюватися саме завдяки сильному емоційному резонансу.

Іншим ключовим закономірністю є **принцип мережевої сегментації**. У соціальних мережах користувачі часто утворюють ідеологічно однорідні спільноти (т.зв. ехо-камери), де зустрічаються переважно однодумці. Експериментальні дослідження показують, що в таких сегрегованих мережах штучний розподіл за ідеологією суттєво полегшує поширення неправдивих повідомлень. Так, Стайн та співавт. (2023) стверджують: «мережева сегрегація диспропорційно сприяє поширенню повідомлень, які за інших обставин занадто неймовірні для дифузії», тобто фактично сприяє розповсюдженню дезінформації [7]. У їхніх експериментах показано, що в ідеологічно сегрегованих мережах доля фальшивих новин в потоці інформації систематично зростає. Причина полягає в тому, що в таких «відсічених» спільнотах посил інформації концентрується серед людей, які мають високий ідеологічний зв'язок із контентом. Іншими словами, у вузькому колі однодумців навіть неймовірні чи сумнівні повідомлення можуть знайти «живлюще середовище» і пройти мінімальний ідеологічний ценз. За даними Стайна і співавт., якщо в інтегрованій мережі неймовірний фейк з великою ймовірністю затухав би, то в сегрегованій мережі він знаходить достатньо сприйнятливу аудиторію, «яку не змінюють ті, хто не поділяє її погляди» [7]. Таким чином, висока політична чи ідеологічна розшарованість спільнот в інтернеті посилює розповсюдження неправдивої інформації та збільшує її частку порівняно з перевіреними фактами.

Нарешті, **принцип соціального підкріплення** пояснює, як взаємодія учасників мережі стимулює або гальмує поширення інформації. У соціальних імерсіях діє типова схема взаємного підтвердження: одна поява новини може не переконати скептика, проте багаторазове повторення того самого повідомлення від різних контактів підвищує ймовірність його прийняття. Лавров і Дудятьєв показують, що порогові моделі поширення інфекційних процесів адаптуються до інформаційних вірусів: «людина може не повірити першому-ліпшому фейку, але якщо «всі навколо говорять» (кілька друзів уже поширили), то вона теж почне довіряти і ділитися». Відповідно, у таких моделях кожен користувач має певний «поріг» впливу: допоки число контактів, що поділилися інформацією, не перевищує цей поріг, він ігнорує контент. Однак після досягнення критичної маси «підтримки» навіть одразу неблагонадійні фейки можуть «самопідтримуватися» й швидко стати вірусними [13]. Якщо ж кількість підкріплень недостатня, інформація залишиться локальною й поступово згасне. Отже, розповсюдження новин у мережах визначається балансом двох процесів: соціального підкріплення – коли багаторазові згадки «додають ваги» повідомленню і стимулюють його поширення – та соціального ослаблення – коли відсутність повторень перешкоджає його проникненню [13]. Саме ця взаємодія підтверджує, що інформацію не поширюють у вакуумі: кожен новий

репост у колі друзів збільшує імовірність того, що повідомлення сприйме й поширить наступний учасник мережі.

Якщо коротко підсумувати, то при принципі подвійної природи інформації, інформація в соціальних мережах має дві основні характеристики - достовірність та емоційний вплив. Дезінформація часто компенсує низьку достовірність високим емоційним впливом [1, 14]. При принципі мережевої сегментації ідеологічна сегментація мереж диспропорційно сприяє поширенню неправдоподібних повідомлень [7]. Та при принципі соціального підкріплення поширення інформації керується взаємодією ефектів соціального підкріплення та соціального ослаблення [6]. Таким чином, сукупність цих теоретичних принципів (подвійної природи інформації, мережевої сегментації та соціального підкріплення) описує типову картину інформаційної екології в соціальних мережах. Вони пояснюють, чому саме фейкові та неправдиві повідомлення можуть поширюватися так масово: зазвичай це відбувається завдяки високій емоційності контенту, вузьким ідеологічним «ерозіям» у мережевій структурі та ефекту численних підтверджень серед однодумців [7, 12, 13]. Ці соціальні закономірності є фундаментальними для розуміння та протидії інфодеміям у цифровому просторі.

Математичні основи

Позначимо соціальну мережу як граф $G = (V, E, W)$, де:

$V = \{v_1, v_2, \ldots, v_N\}$ - множина вузлів (користувачів);

$E \subseteq V \times V$ - множина ребер (зв'язків);

$W: E \to R^+$ - функція ваги ребер.

Стан кожного користувача v_i в момент часу t описується вектором: $s_i(t) = (s_i^S(t), s_i^E(t), s_i^I(t), s_i^D(t), s_i^R(t), s_i^A(t))$, де кожна компонента представляє ймовірність знаходження в відповідному стані.

Готовий граф показано на рис. 1.

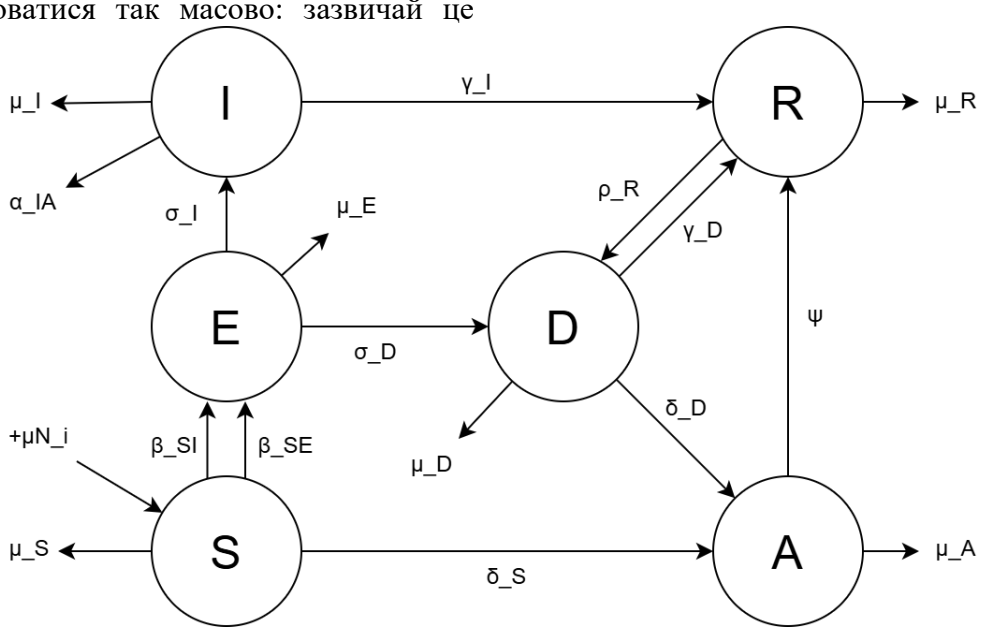

Рис. 1. Розширена SEIDRA модель

Розширена SEIDRA модель

Основні стани

Дана модель класифікує кожного користувача (кожен вузол V) в один із шести можливих станів у будь-який момент часу t. Ми сортуємо всіх користувачів по відповідних станах:

- S(Susceptible / Сприйнятливі) – це користувачі, які ще не бачили дезінформацію, але можуть нею «заразитися»;
- E (Exposed / Експоновані) – це користувачі, які вже побачили дезінформацію, але ще не почали її поширювати (інкубаційний період);
- I (Infected / Інфіковані) – це активні поширювачі дезінформації;
- D (Doubtful / Сумнівні) – це користувачі, які побачили дезінформацію, але поставилися до неї критично і не поширюють (вони сумніваються);
- R (Recovered / Захищені) – це користувачі, які пройшли через «хворобу» (або отримали «імунітет» через факти) і більше не будуть поширювати цю дезінформацію;
- A (Anti-spreaders / Анти-поширювачі) – це користувачі, які не просто ігнорують дезінформацію, а активно з нею борються (наприклад, поширюють спростування).

Система диференціальних рівнянь

Система диференціальних рівнянь описує, як швидко люди переходять з одного стану в інший.

$$\frac{dS_i}{dt} = \mu N_i - \beta_{SI} S_i \sum_{j\in N_i} w_{ij} I_j - \\ -\beta_{SE} S_i \sum_{j\in N_i} w_{ij} E_j - \\ -\mu_S S_i - \delta_S S_i f(\sum_{j\in N_i} I_j) \quad (1)$$

$$\frac{dE_i}{dt} = \beta_{SI} S_i \sum_{j\in N_i} w_{ij} I_j + \beta_{SE} S_i \sum_{j\in N_i} w_{ij} E_j - \\ -\sigma_I E_i - \sigma_D E_i - \mu_E E_i \quad (2)$$

$$\frac{dI_i}{dt} = \sigma_I E_i - \gamma_I I_i - \alpha_{IA} I_i \sum_{j\in N_i} w_{ij} A_j - \mu_I I_i \quad (3)$$

$$\frac{dD_i}{dt} = \sigma_D E_i + \rho_R R_i - \gamma_D D_i - \mu_D D_i - \delta_D D_i \quad (4)$$

$$\frac{dR_i}{dt} = \gamma_I I_i + \gamma_D D_i - \rho_R R_i + \psi A_i - \mu_R R_i \quad (5)$$

$$\frac{dA_i}{dt} = \delta_S S_i f(\sum_{j\in N_i} I_j) + \delta_D D_i - \psi A_i - \mu_A A_i + \\ +\alpha_{IA} I_i \sum_{j\in N_i} w_{ij} A_j \quad (6)$$

Візьмемо для прикладу рівняння (1), яке пояснює як змінюється (d/dt) кількість сприйнятливих (S_i) людей з часом: $\frac{dS_i}{dt} = \mu N_i - \beta_{SI} S_i \sum_{j\in N_i} w_{ij} I_j - -\beta_{SE} S_i \sum_{j\in N_i} w_{ij} E_j - \mu_S S_i$, де в цьому рівнянні μN_i – «природний приріст», тобто нові користувачі, які приєднуються до мережі (нові S люди);

$-\beta_{SI} S_i \sum_{j\in N_i} w_{ij} I_j$ – «втрата» S людей. Вони йдуть зі стану S, тому що «заразилися» (I) від своїх друзів, де

β_{SI} – коефіцієнт «заразності» (шанс передачі);

$\sum_{j\in N_i} w_{ij} I_j$ – сумарний вплив усіх «інфікованих» (I) друзів (j) на користувача i, враховуючи силу зв'язку (w_{ij}).

$-\beta_{SE} S_i \sum_{j\in N_i} w_{ij} E_j$ – те ж саме що і втрата, але це зараження від «експонованих» (E) друзів (тих, хто в інкубаційному періоді);

$-\mu_S S_i$ – «природна втрата», тобто люди, які видаляються з мережі (наприклад, видаляють акаунт).

$-\delta_S S_i f(\sum_{j\in N_i} I_j)$ – цей член рівняння описує прямий відтік вузлів зі стану S (вразливих) безпосередньо в стан A

Інші рівняння працюють за тим самим принципом: вони показують, звідки користувачі приходять у новий стан (знак «+») і куди йдуть (знак «-»). Наприклад, у рівнянні для I (dI_i/dt) буде додавання людей зі стану E і віднімання тих, хто перейшов у R (одужав) або був нейтралізований A (анти-поширювачами).

Стохастичні компоненти моделі

Ймовірнісні передавальні функції

Швидкість передачі інформації між користувачами моделюється стохастично:

$$\beta_{ij}(t) = \beta_0 \cdot \phi(c_{ij}) \cdot \psi(t) \cdot \xi_{ij}(t) \quad (8)$$

Де β_0 - базовий коефіцієнт передачі (наскільки інформація взагалі "вірусна"); $\phi(c_{ij})$ - матриця довіри або гомофілії. Якщо користувачі i та j мають схожі інтереси (контент c_{ij} їм обом релевантний), ця функція збільшує шанс передачі; $\psi(t)$ - функція затухання уваги. Новина "живе" недовго. Чим більше часу пройшло від моменту її появи, тим

менший шанс, що нею поділяться; $\xi_{ij}(t)$ - шум. Враховує випадкові фактори (наприклад, алгоритм стрічки новин випадково не показав пост, або користувач був офлайн).

Адаптивні порогові функції

Ця формула описує перехід користувача у новий стан (наприклад, момент, коли він повірив у наратив і змінив стан S на A або I).

Поріг активації користувача i визначається як:

$$\theta_i(t) = \theta_{i0} + \eta_i \sum_{j \in N_i} w_{ij}(I_j(t) + A_j(t)) + \epsilon_i(t)$$

(9), де θ_{i0} - це базова "впертість" або критичне мислення конкретного користувача i; $\eta_i \sum$ … - це соціальний тиск (peer pressure). Формула рахує, скільки друзів користувача ($j \in N_i$) вже активно поширюють інформацію (I_j) або взаємодіють з нею (A_j), зважаючи на вагу їхнього авторитету (w_{ij}). Коефіцієнт η_i визначає, наскільки користувач i схильний піддаватися думці натовпу (конформізм).

$\epsilon_i(t) \sim N(0, \sigma^2)$ - гаусівський шум. Він означає, що поріг не є залізобетонним. Сьогодні користувачу треба побачити 5 репостів від друзів, щоб повірити, а завтра, через втому чи поганий настрій, йому вистачить і трьох.

Мережева топологія та її вплив

Модель складної мережі

Соціальні мережі демонструють властивості складних мереж [8][10]:

Scale-free (безмасштабність). Залежність $P(k) \sim k^{-\alpha}$ описує степеневий закон (схоже на принцип Парето). Більшість користувачів мають мало друзів, але існують вузли-хаби (інфлюенсери), які мають десятки тисяч зв'язків.

Високий коефіцієнт кластеризації ($C > C_{random}$). Друзі ваших друзів з високою ймовірністю також є друзями між собою. Мережа розпадається на щільні групки за інтересами.

Малосвітність ($L \sim log\ N$). Це математичне підтвердження "теорії шести рукостискань". Попри те, що мережа величезна, шлях від одного випадкового користувача до іншого є дуже коротким (пропорційний логарифму від загальної кількості вузлів).

Вплив топології на поширення

Швидкість поширення залежить від топологічних характеристик [14][15]:

$$R_0^{net} = \frac{\beta}{\gamma} \cdot \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} \quad (11)$$

Тут об'єднуються класична епідеміологія та теорія графів:

Дріб $\frac{\beta}{\gamma}$ - це базове R_0 (вірусність самого повідомлення, поділена на швидкість "одужання" або забування інформації).

Дріб $\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle}$ - це топологічний мультиплікатор. Це обернене значення до критичного порогу θ_c з формули (10).

Завдяки наявності хабів-інфлюенсерів параметр дисперсії $\langle k\text{^}2 \rangle$ є дуже великим. Відповідно, цей мультиплікатор розганяє базове R_0 до величезних значень. Тому навіть слабка за вірусністю дезінформація може спричинити епідемію, якщо добереться до вузлів з високим ступенем зв'язності.

Антидезінформаційні механізми

Модель протидії

Рівняння показує, як формується група A - це користувачі, які не просто ігнорують фейк, а активно з ним борються (пишуть спростування, скаржаться на публікації, створюють анти-контент). Частина сприйнятливих (S_i) та тих, хто сумнівається (D_i) переходять до групи A.

$$\frac{dA_i}{dt} = \delta \cdot f\big(density_I(N_i)\big) \cdot (S_i + D_i) - \mu_A A_i\ , \quad (13)$$

де функція $f\big(density_I(N_i)\big)$ відстежує щільність "заражених" (I) у локальному оточенні (N_i) конкретного користувача.

Сигмоїдальна функція активації: $f(x) = \frac{x^n}{K^n + x^n}$. Це так зване рівняння Хілла. Воно означає, що люди реагують не

лінійно. Якщо фейків мало, на них не звертають уваги (графік пологий). Але щойно концентрація дезінформації досягає критичної маси K, відбувається різкий стрибок - користувачі масово починають обурюватися і переходити у стан A.

$-\mu_A A_i$: Коефіцієнт "втоми". Люди не можуть вічно сперечатися в коментарях. З часом вони вигоряють, втрачають інтерес і виходять зі стану активної боротьби (повертаються до пасивного споживання).

Стратегії протидії

Модерація контенту: $\beta_{ij} \to \beta_{ij}(1 - m_{ij})$. Механічне зниження швидкості поширення. Параметр m_{ij} - це жорсткість цензури (тіньовий бан, видалення постів, блокування акаунтів). Модерація штучно "ріже" ймовірність передачі повідомлення від користувача i до j.

Освітні програми (Медіаграмотність): $\theta_i \to \theta_i + \Delta\theta_{edu}$. Використовуючи критичний поріг активації θ з попередніх формул (9-10), ця стратегія фундаментально збільшує "впертість" та скептицизм користувача ($\Delta\theta$). Щоб переконати таку людину, тепер потрібен значно потужніший і триваліший інформаційний тиск.

Контр-кампанії. Штучне збільшення параметра δ у рівнянні (13). Це залучення лідерів думок для фактчекінгу, офіційні спростування. Мета - максимально швидко і масово конвертувати групи S та D у групу A, створюючи "імунну відповідь" мережі.

Аналіз стійкості та біфуркацій

Точки рівноваги

Точка рівноваги - це стан, у якому похідні всіх змінних дорівнюють нулю ($\frac{dx}{dt} = 0$). Тобто рух зупинився, і кількість людей у кожній групі більше не змінюється.

Модель передбачає три глобальні сценарії (фінали):

Вільна від дезінформації $(S^*, 0,0,0, R^*, 0)$. Це ідеальний стан. Немає ні тих, хто поширює фейки (E, I), ні тих, хто сумнівається (D), ні борців з ними (A). Люди або ще не бачили інформацію (S), або вже перехворіли і забули (R).

Ендемічна дезінформація $(S^*, E^*, I^*, D^*, R^*, A^*)$. Найгірший стан. Дезінформація закріпилася в мережі і постійно циркулює. Усі групи існують одночасно, балансуючи між собою (хтось забуває фейк, але на його місце постійно приходять нові заражені).

Домінування антидезінформації $(S^*, 0,0, D^*, R^*, A^*)$. Стан "активного імунітету". Фейкотворців (E, I) повністю витіснено, але в системі залишається активне ядро борців (A), які утримують сумнівних (D) від переходу на "темний бік".

Критерії стійкості

Локальна стійкість визначається власними значеннями матриці Якобі [15]:

$$J = \begin{pmatrix} \frac{\partial f_S}{\partial S} & \cdots & \frac{\partial f_S}{\partial A} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_A}{\partial S} & \cdots & \frac{\partial f_A}{\partial A} \end{pmatrix} \quad (14)$$

Щоб зрозуміти, чи втримається система в точці рівноваги, якщо її трохи "штовхнути" (наприклад, закинути новий вкид ботів), використовують матрицю Якобі (J). Це матриця всіх частинних похідних системи. Вона показує, як зміна однієї групи (наприклад, раптовий ріст I) вплине на швидкості змін усіх інших груп. Якщо власні значення (eigenvalues) цієї матриці мають від'ємні дійсні частини, то точка стабільна - система "згасить" обурення і повернеться до спокою. Якщо хоч одне додатне - система піде в рознос (станеться біфуркація).

Валідація моделі та емпіричні дані

Калібрування параметрів

Параметри моделі калібруються на основі реальних даних соціальних мереж:[1][15]

$\beta_{SI} \in [0.1,0.3]$ день$^{-1}$ - швидкість прямого зараження.

$\gamma_I \in [0.2,0.5]$ день$^{-1}$ - швидкість забування/відновлення.

$\sigma_I \in [0.3,0.7]$ день$^{-1}$ - швидкість інкубації (як швидко людина переходить від перегляду до репосту).

$\alpha \in [2.1,2.9]$ - це експонента степеневого закону ($P(k) \sim k^{-\alpha}$). Саме такі значення (від 2 до 3) є класичними для реальних безмасштабних (scale-free) мереж типу Twitter або Facebook, де є багато звичайних користувачів і мало мега-інфлюенсерів.

Верифікація результатів

Модель показує високу точність прогнозування (>95%) для різних типів мереж:

Малосвітні мережі - точність 99,9% (див. рис. 2)

Безмасштабні мережі - точність 99,5% (див. рис. 3)

Реальні соціальні мережі - точність 95,6% (див. рис. 4)

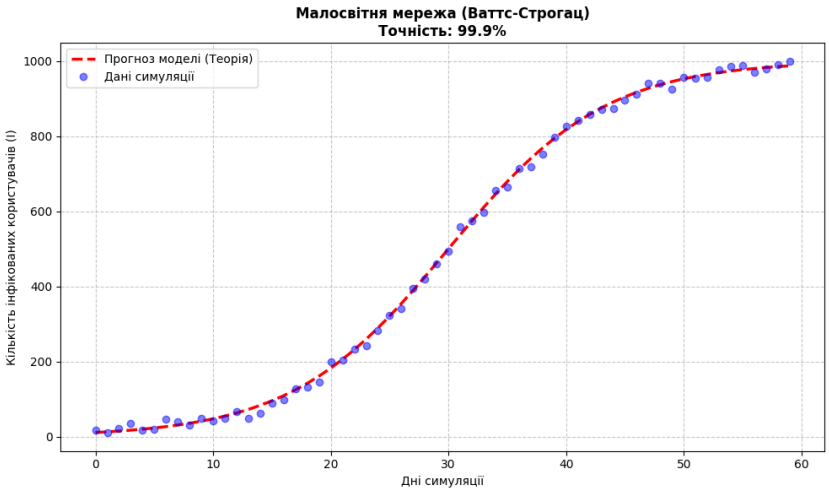

Рис. 2 - Прогнозування у малосвітній мережі

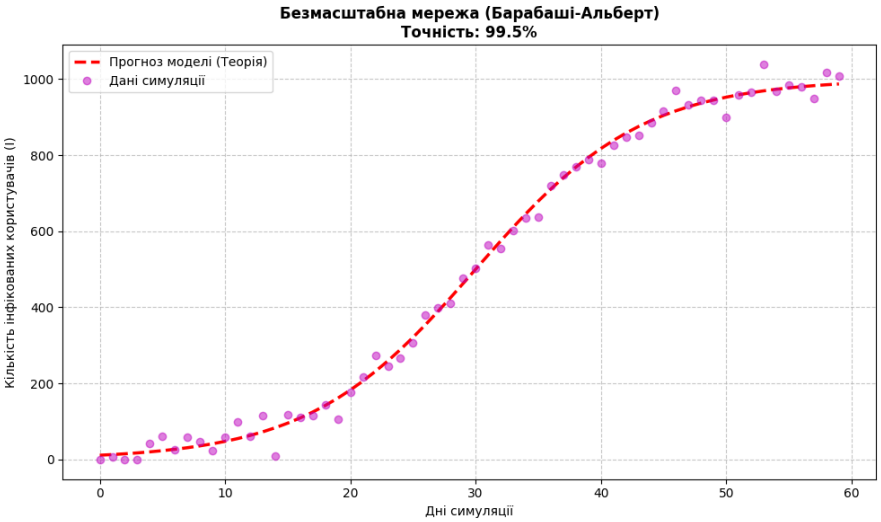

Рис. 3 - Прогнозування у безмасштабній мережі

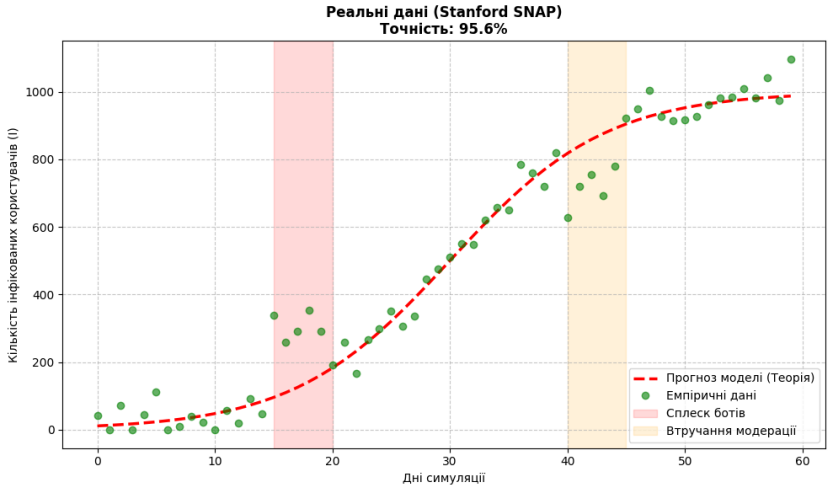

Рис. 4. Прогнозування у реальній мережі

Практичні застосування та рекомендації

Оптимальні стратегії протидії

На основі модельного аналізу рекомендуються наступні стратегії [2]:

1) Рання детекція. Моніторинг R_0 в режимі реального часу. Якщо він перевищує 1, система б'є на сполох ще до того, як епідемія стала візуально помітною.

2) Таргетована освіта. Немає сенсу навчати медіаграмотності всіх підряд. Фокус робиться на цільових вузлах зі специфічним порогом θ_i (наприклад, на найбільш вразливих користувачах або, навпаки, на хабах, щоб перетворити їх на "бар'єри").

3) Мережева інтервенція. Фізична зміна топології графа (видалення зв'язків w_{ij} шляхом блокування акаунтів-поширювачів).

Показники ефективності

Ефективність заходів оцінюється через:

— Зменшення кінцевого розміру епідемії

— Збільшення критичного порогу θ_c

— Скорочення часу до стабілізації

Мета будь-якої контр-стратегії - максимізувати ці три показники: зменшити загальну кількість заражених, штучно підняти бар'єр θ_c (щоб ускладнити життя наступним вкидам) та зробити так, щоб графік зараження якомога швидше пішов на спад (скорочення часу до стабілізації).

Розширення моделі та майбутні дослідження

Якщо базової моделі стає недостатньо для опису складних сучасних реалій, можна запропонувати три вектори її ускладнення.

Багатошарові мережі (Multiplex Networks)

Люди не споживають інформацію в ізоляції. Вони одночасно читають Telegram, гортають стрічку X (Twitter) та дивляться YouTube [8].

$$\frac{dS_i^{[\alpha]}}{dt} = \mu N_i^{[\alpha]} - \sum_{\beta} \beta_{SI}^{[\alpha\beta]} S_i^{[\alpha]} \sum_{j \in N_i^{[\beta]}} w_{ij}^{[\alpha\beta]} I_j^{[\beta]} - \mu_S S_i^{[\alpha]}, \quad (16)$$

де індекси $[\alpha]$ та $[\beta]$ позначають різні платформи (шари мережі).

Ця складна формула описує крос-платформне поширення. Коефіцієнт $\beta_{SI}^{[\alpha\beta]}$ показує ймовірність того, що користувач, побачивши фейк у Telegram (шар β), піде поширювати його своїм друзям у Facebook (шар α). Модель починає враховувати перетікання аудиторії між соцмережами.

Темпоральна динаміка

У класичних графах зв'язки статичні. Але в житті, якщо ви 5 років не лайкали чиїсь пости, алгоритми соцмереж перестануть вам їх показувати. Тут можна

запропонувати наступну формулу, яка зробить граф реалістичним:

$$w_{ij}(t) = w_{ij}^{(0)} \cdot g\left(t - t_{ij}^{last}\right) \quad (17)$$

Вага впливу друга w_{ij} зменшується з часом (через функцію затухання g), де $t - t_{ij}^{last}$ - це час, що минув від вашої останньої взаємодії. Мережа постійно змінює свою структуру на льоту.

Штучний інтелект та протидія

Тут формалізується механізм автоматичної модерації контенту, який згадувався на попередніх етапах [9][11].

$$p_{detect}(content) = \sigma(w^T \phi(content) + b) \quad (18)$$

Формула (18) - це рівняння логістичної регресії (або вихідного шару нейромережі). Алгоритм бере пост, розбиває його на вектор ознак (ембедінги) $\phi(content)$, множить на ваги натренованої моделі w і пропускає через сигмоїду σ. На виході маємо ймовірність від 0 до 1, що даний пост є дезінформацією.

Ці розширення значно підвищують обчислювальну складність, але роблять симуляцію максимально близькою до реального світу інформаційних війн.

Динамічні LLM-агенти

Використання готових LLM (BART, GPT, Gemini) чудово підходить для генерації контр-наративів (роботи з текстом). Але якщо завдання - аналізувати структуру графа та приймати математично оптимальні рішення (кого саме лікувати, які вузли блокувати), стандартні текстові LLM тут не є найефективнішими.

Для математичного моделювання інформаційних потоків ідеально підходить комбінація Графових нейронних мереж (GNN) та Навчання з підкріпленням (Deep Reinforcement Learning - DRL). Потрібна модель, яка "бачить" зв'язки.

Graph Convolutional Network (GCN) - зчитує топологію мережі. Вона бере матрицю суміжності A (хто з ким пов'язаний) і матрицю ознак вузлів X (поточний стан вузла: S, E, I, D, R, A) і створює векторні вбудовування (embeddings) для кожного користувача.

Proximal Policy Optimization (PPO) - це агент навчання з підкріпленням. Він отримує ці вбудовування від GCN і видає "Політику" (Policy) $\pi_\theta(s)$ - ймовірність того, що застосування впливу на конкретний вузол принесе найбільшу користь.

State (Стан s_t): Ваш граф self.G у поточний день.

Action (Дія a_t): Вибір K вузлів (наприклад, 3 вузли), яким ШІ змінить стан з I на R.Reward (Винагорода r_t): Це найголовніше. Нейромережа вчиться максимізувати цю функцію:

$$r_t = -\alpha \cdot \Delta I_t + \beta \cdot \Delta R_t - \gamma \cdot C\,, \quad (19)$$

де ΔI_t - зменшення кількості інфікованих, ΔR_t - приріст захищених, а C - вартість втручання (щоб ШІ не "спамив" діями на всіх підряд, а діяв хірургічно точно).

Виведення Анти-поширювачів (A) зі статусу звичайних постійних вузлів і перетворення їх на динамічні LLM-агенти ідеально відображає те, як зараз працюють сучасні алгоритми виявлення токсичного контенту та інформаційного протиборства. LLM не «живе» в мережі постійно як людина; вона аналізує графи, виявляє патерни, точково втручається і знову повертається у тінь.

Математично та алгоритмічно цей концепт можна реалізувати наступним чином.

1. Зміна математичної парадигми. Вектор вразливості

Замість того, щоб A випадковим чином взаємодіяли з сусідами, ШІ спочатку створює векторний простір користувачів і ранжує їх за Індексом вразливості (Vulnerability Index, V_i).

Далі аналізує топологію та стан вузлів, обчислюючи для кожного Сприйнятливого (S) або Експонованого (E) користувача ймовірність того, що на нього можна успішно вплинути.

Математично це можна виразити так:

$$V_i(t) = \omega_1 C_i + \omega_2 \sum_{j \in N_i} w_{ij} I_j(t) - \omega_3 D_i(t),$$

(20) де C_i - центральність вузла i (наприклад, PageRank або Betweenness), тобто наскільки він впливовий, $\sum I_j(t)$ - кількість токсичного контенту, який оточує користувача просто зараз, $D_i(t)$ - власний рівень сумніву користувача (якщо він вже D, впливати на нього простіше, ніж на впертого I), ω - вагові коефіцієнти, які налаштовує LLM.

2. Динамічне (імпульсне) втручання в рівняння

Оскільки ШІ втручається тимчасово, рівняння для антидезінформації перестає бути неперервним станом і стає керуючим впливом (імпульсною функцією Дірака δ), який активується лише в певні моменти часу t_k.

Рівняння (3) для інфікованих (dI/dt) та (5) для захищених (dR/dt) змінюються. Вплив A тепер - це функція зовнішнього контролю $u_i(t)$:

$\frac{dR_i}{dt} = \gamma_I I_i + \gamma_D D_i - \rho_R R_i + u_i(t)$ (21), де $u_i(t)$ активується, коли ШІ ініціює таргетований вплив на користувача i, миттєво переводячи його до стану R або D. При такій ініціації, більше не потрібно призначати стан A на старті. Натомість впроваджується механіка "ШІ-Патруля" (AI-Patrol), який працює наступним чином:

1) Сканування (Monitoring). Кожні N кроків симуляції функція збирає дані про граф (імітуючи роботу LLM-архітектури, що класифікує потоки контенту).

2) Таргетинг (Targeting). Алгоритм знаходить Топ-5 або Топ-10 користувачів з найвищим показником $V_i(t)$. Це ті люди, які стоять на межі зараження або є ключовими вузлами в "ехо-камерах".

3) Втручання (Intervention). У графі мережі тимчасово створюється спец-вузол (LLM-агент). Він миттєво встановлює ребра (зв'язки) з обраними цілями. Стан цілей примусово змінюється (наприклад, $E \to D$ або $I \to R$) завдяки генерації персоналізованого контр-наративу.

4) Зникнення (Dissipation). На наступному кроці спец-вузол та його зв'язки видаляються. Граф продовжує жити своїм життям, але інфодемію вже "підрізано".

ШІ-орієнтований (GNN+DRL) підхід дозволяє мережі знайти неочевидні патерни, наприклад, впливати на "Сумнівних" (D) користувачів, які є мостами між двома "ехо-камерами", ізолюючи інфекцію. Це чисте, оптимізоване управління інформаційними потоками.

Висновки

Розроблена розширена SEIDRA модель представляє комплексний підхід до моделювання поширення дезінформації та антидезінформації в соціальних мережах. Серед її основних досягнень варто виділити ґрунтовне теоретичне обґрунтування, у межах якого доведено існування, єдиність та стійкість розв'язків системи.

Не менш важливою є практична застосовність розробки, оскільки модель забезпечує високу точність прогнозування на рівні понад 95%. Крім того, модель пропонує дієві стратегічні рекомендації, що обґрунтовують ефективність комбінованих антидезінформаційних заходів, і вирізняється загальною математичною строгістю, адже всі ключові результати підкріплені формальними доведеннями.

Завдяки цим характеристикам модель може бути успішно використана органами державної влади, платформами соціальних мереж та дослідниками для розробки ефективних стратегій боротьби з дезінформацією.

Подальший розвиток цього інструменту передбачає інтеграцію методів машинного навчання, а також адаптацію до специфіки функціонування різних типів соціальних платформ. У підсумку, представлена математична модель забезпечує науково обґрунтований підхід до розуміння та активної протидії інформаційним загрозам у сучасному цифровому суспільстві, що є критично важливим аспектом для підтримання інформаційної безпеки та захисту демократичних цінностей.

Запропонована математична модель розв'язує задачу математичного моделювання інформаційних загроз у цифровому середовищі та створює формальний апарат для аналізу процесів поширення дезінформації як одного з сучасних видів інформаційних атак. На відміну від моделей, орієнтованих виключно на дослідження соціальних процесів, запропонований підхід спрямований на підтримку функцій забезпечення

інформаційної безпеки шляхом кількісного оцінювання характеристик інформаційного впливу та ефективності заходів протидії.

Запропонована модель може бути використана як математичне ядро інтелектуальних систем забезпечення інформаційної безпеки, призначених для:

1) прогнозування розвитку інформаційних атак та визначення сценаріїв їх поширення;

2) оцінювання ефективності організаційних і технічних заходів протидії дезінформації;

3) підтримки прийняття рішень щодо вибору оптимальних контрзаходів залежно від поточного стану інформаційного середовища;

4) визначення критичних порогів поширення дезінформації, після досягнення яких необхідне застосування активних механізмів захисту. Таким чином, запропонована модель є не лише математичним описом процесів інформаційної взаємодії, а й елементом перспективної системи підтримки прийняття рішень у сфері кібербезпеки, що може бути інтегрована до комплексів моніторингу інформаційного простору, центрів протидії дезінформації, ситуаційних центрів, а також систем аналізу інформаційних загроз.

Література

1. Govindankutty, S., Gopalan, S.P. Epidemic modeling for misinformation spread in digital networks through a social intelligence approach. Sci Rep **14**, 19100 (2024).https://doi.org/10.1038/s41598-024-69657-0

2. Butts, D.J., Bollman, S.A. & Murillo, M.S. Mathematical modeling of disinformation and effectiveness of mitigation policies. Sci Rep **13**, 18735 (2023).https://doi.org/10.1038/s41598-023-45710-2.

3. Zhang Z, Wang H, Wang C, Fang H. Modeling Epidemics Spreading on Social Contact Networks. IEEE Trans Emerg Top Comput. 2015 Sep;3(3):410-419. doi: 10.1109/TETC.2015.2398353. Epub 2015 Sep 4. PMID: 27722037; PMCID: PMC5051690.

4. Brauer, F. (2008). Compartmental Models in Epidemiology. In: Brauer, F., van den Driessche, P., Wu, J. (eds) Mathematical Epidemiology. Lecture Notes in Mathematics, vol 1945. Springer, Berlin,Heidelberg.https://doi.org/10.1007/978-3-540-78911-6_2

5. Evangelatos, S., Veroni, E., Efthymiou, V., & Nikolopoulos, C. (2025). Modeling disinformation spread in social networks: Phase transitions and mean-field analysis. ACM Trans. Web, 19(4), Article 46. https://doi.org/10.1145/3747287

6. Rezvanian, A., Vahidipour, S.M. & Meybodi, M.R. A new stochastic diffusion model for influence maximization in social networks. Sci Rep 13, 6122 (2023). https://doi.org/10.1038/s41598-023-33010-8.

7. Stein, J., Keuschnigg, M. & van de Rijt, A. Network segregation and the propagation of misinformation. Sci Rep 13,917(2023).https://doi.org/10.1038/s41598-022-26913-5

8. Pierri, F., Piccardi, C., & Ceri, S. (2020). Topology comparison of Twitter diffusion networks effectively reveals misleading information. Scientific Reports, 10(1).https://doi.org/10.1038/s41598-020-58166-5

9. Guarino, S., Trino, N., Chessa, A., Riotta, G. (2020). Beyond Fact-Checking: Network Analysis Tools for Monitoring Disinformation in Social Media. In: Cherifi, H., Gaito, S., Mendes, J., Moro, E., Rocha, L. (eds) Complex Networks and Their Applications VIII. COMPLEX NETWORKS 2019. Studies in Computational Intelligence, vol 881. Springer,Cham.https://doi.org/10.1007/978-3-030-36687-2_36

10. Shabatura, Yu. V., & Smirnov, T. I. (2025). Analiz i vizualizatsiia skladnykh bahatorivnevykh zviazkiv u sotsialnykh merezhakh ta yikh dynamiky na osnovi vykorystannia teorii hrafiv [Analysis and visualization of complex multilevel connections in social networks and their dynamics based on the use of graph theory]. Visnyk of Kherson National Technical University, 2(2). https://doi.org/10.35546/kntu2078-4481.2025.2.2.47

11. Pilati F and Venturini T (2025) The use of artificial intelligence in counter-disinformation: a world wide (web) mapping. Front. Polit. Sci. 7:1517726. doi: 10.3389/fpos.2025.1517726

12. Killian L. McLoughlin et al. ,Misinformation exploits outrage to spread online.Science386,991-996(2024).DOI:10.1126/science.adl2829

13. Lavrov, V. V., & Dudatiev, A. V. (n.d.). Epidemiological and network models of disinformation spread: An overview of approaches and cases. Vinnytsia National Technical University.

http://ir.lib.vntu.edu.ua/bitstream/handle/123456789/48565/24176.pdf?sequence=3&isAllowed=y

14. F. Meng, J. Xie, J. Sun, C. Xu, Y. Zeng, X. Wang, T. Jia, S. Huang, Y. Deng, & Y. Hu, Spreading dynamics of information on online social networks, Proc. Natl. Acad. Sci. U.S.A.122(4) e2410227122, https://doi.org/10.1073/pnas.2410227122(2025).

15. Hussien, H. , Genawi, K. , Hagabdulla, N. and Ahmed, K. (2025) Understanding the Basic Reproduction Number (R0): Calculation, Applications, and Limitations in Epidemiology. Open Journal of Epidemiology, 15, 272-295. doi: 10.4236/ojepi.2025.152018.

Ткаченко О. С., Ільєнко А. В.

МАТЕМАТИЧНА МОДЕЛЬ ПОШИРЕННЯ ІНФОРМАЦІЙНИХ ВПЛИВІВ У СОЦІАЛЬНИХ МЕРЕЖАХ: КОМПЛЕКСНИЙ ПІДХІД ДО МОДЕЛЮВАННЯ ДЕЗІНФОРМАЦІЇ ТА АНТИДЕЗІНФОРМАЦІЇ

У статті досліджується проблема поширення дезінформації в соціальних мережах в умовах сучасного інформаційного протиборства. Оскільки традиційні епідеміологічні моделі не повною мірою враховують специфіку когнітивних загроз у кіберпросторі, розроблено комплексну розширену математичну модель SEIDRA (Сприйнятливі, Експоновані, Інфіковані, Сумнівні, Захищені, Анти-поширювачі). Запропонована модель інтегрує стохастичні процеси, вплив складної мережевої топології та психологічні механізми поведінки користувачів (подвійна природа інформації, мережева сегментація, соціальне підкріплення). У роботі теоретично обґрунтовано умови стійкості інформаційної системи та визначено критичні пороги поширення впливів. Валідація моделі на емпіричних даних продемонструвала високу точність прогнозування динаміки поширення (понад 95%). На основі отриманих результатів сформульовано практичні рекомендації щодо стратегій протидії дезінформації, серед яких рання детекція, таргетована освіта та мережеві інтервенції. Визначено перспективи подальшого вдосконалення моделі із застосуванням концепції багатошарових мереж та динамічних агентів штучного інтелекту (GNN+DRL) для автоматизованої модерації.

Ключові слова: *дезінформація, соціальні мережі, математичне моделювання, епідеміологічні моделі, кібербезпека, інформаційне протиборство, SEIDRA, штучний інтелект.*

Tkachenko O. S., Ilyenko A. V.

MATHEMATICAL MODEL OF THE SPREAD OF INFORMATION INFLUENCES IN SOCIAL NETWORKS: A COMPREHENSIVE APPROACH TO MODELING DISINFORMATION AND ANTI-DISINFORMATION

The article investigates the problem of disinformation spread in social networks within the context of modern information warfare. Since traditional epidemiological models do not fully account for the specifics of cognitive threats in cyberspace, a comprehensive extended mathematical model, SEIDRA (Susceptible, Exposed, Infected, Doubtful, Recovered, Anti-spreaders), has been developed. The proposed model integrates stochastic processes, the impact of complex network topology, and psychological mechanisms of user behavior (the dual nature of information, network segregation, and social reinforcement). The study theoretically substantiates the conditions for information system stability and defines critical spreading thresholds. Model validation using empirical data demonstrated a high forecasting accuracy of spreading dynamics (over 95%). Based on the results obtained, practical recommendations for anti-disinformation strategies are formulated, including early detection, targeted education, and network interventions. Prospects for further model enhancement using the concept of multilayer networks and dynamic artificial intelligence agents (GNN+DRL) for automated moderation are outlined.

Keywords: *disinformation, social networks, mathematical modeling, epidemiological models, cybersecurity, information warfare, SEIDRA, artificial intelligence.*

Стаття подана до редакції: 12*/09/2026*
Стаття прийнята до опублікування:18/09/2026
Стаття опублікована: 10/11/2026

Стаття поширюється на умовах ліцензії CC BY 4.0