Мультирепрезентаційна GNN-модель з узгодженням і адаптивним злиттям для детекції спаму

Ірина Замрій[1], Іван Шахматов[2]

[1] *Державний університет інформаційно-комунікаційних технологій, Київ, Україна*
[2] *Державний університет інформаційно-комунікаційних технологій, Київ, Україна*

Анотація
Веб-форми залишаються однією з головних цілей спам-кампаній, де координовані боти поєднують масові відправлення, повторне використання полів, часові сплески та технічне маскування, а обсяг розмітки обмежений і діють вимоги приватності. Пропонується модель контрастивної графової нейромережі на основі кількох представлень для фільтрації веб-спаму. Модель об'єднує три представлення подій. Представлення відправок описує зв'язки між заявками за дублюванням полів, часовою близькістю та контактами. Поведінкове представлення є гетерографом взаємодій користувачів, форм і сторінок разом із технічними атрибутами. Семантичне представлення будується як kNN-граф за векторними поданнями тексту та URL. Для кожного представлення застосовуються окремі енкодери GNN з використанням R-GCN для типізованих відносин та GAT або GCN для семантичних і зв'язків відправок. Далі виконується адаптивне злиття представлень, що автоматично зміщує вагу між поведінковими та контентними сигналами на рівні окремої відправки. Навчання поєднує бінарну перехресну ентропію на позначених прикладах і контрастивне узгодження між представленнями з аугментаціями типу drop-edge, feature-mask і стохастичний kNN, а також легку регуляризацію для стабільності. Підхід орієнтовано на практичне розгортання з часовими сплітами без витоків, інтерпретованістю через ваги злиття та механізми уваги, анонімізацією чутливих атрибутів, масштабованим оновленням графів і контролем балансу помилкових спрацьовувань і пропусків відповідно до бізнес-метрик.

Ключові слова
Штучний інтелект, кібербезпека, веб-спам, вебзастосунки, виявлення спаму, графові нейромережі (GNN), сканери вразливостей, інформаційна безпека.

1. Вступ

Веб-спам уже давно не зводиться до «поганих слів» у тексті. Сьогодні це скоординовані кампанії: масові надсилання форм, повторне використання шаблонів, одночасні сплески у часі, однакові технічні ознаки пристроїв. Класичні контентні фільтри швидко застарівають, великі мовні моделі потребують багато позначених прикладів, а прості графові підходи зазвичай дивляться лише на один тип зв'язків. Потрібна модель, що поєднує поведінкові й текстові сигнали, добре працює, коли мало міток, і залишається стабільною, коли тактики спаму змінюються.

Існуючі дослідження охоплюють контентно-семантичні підходи, де текст і посилання перетворюють на граф або використовують сучасні мовні моделі (BERT, ELMo). Це дає помітний приріст, але майже не враховує поведінку користувачів і не узгоджує різні джерела сигналів [1,3,14]. У гетерогенних/багатовідносних графах для ботів, шахрайства та змовних груп важливими є типи зв'язків, уважний відбір корисних сусідів і методи, що враховують відмінність класів у пов'язаних елементах. Також все ще лишаються проблеми зі стійкістю між платформами та поєднанням із контентом [2,4,7,8,11,13]. Напрям контрастивного та попереднього навчання зменшує потребу в розмітці, зіставляючи різні варіанти графа або навчаючись на парах графів, але зазвичай працює в межах одного типу графа без об'єднання кількох представлень на дані [6,9,10]. Отже, бракує рішення, яке одночасно поєднує кілька подань (користувачі/відправки/зміст) і узгоджує їх між собою, дане дослідження направлено

Corresponding authors
i.zamrii@duikt.edu.ua (І.В. Замрій); i.shahmatov@duikt.edu.ua (І.О. Шахматов)
0000-0001-5681-1871 (І.В. Замрій); 0009-0004-9628-0365 (І.О. Шахматов)

на створення нової моделі Multi-View Contrastive GNN яка закриє ці прогалини.
Графове подання природне для цього завдання, так як користувачі взаємодіють із формами та сторінками, відправки утворюють групові сплески, тексти й URL мають смислову схожість. Важливо не обмежуватися одним графом. Різні подання (views) відкривають різні закономірності: у поданні користувача видно шаблони взаємодій і спільні технічні ознаки, у поданні відправок прослідковуються дублікати полів і часові координації, у семантичному є схожість контенту та посилань. Завдання дослідження — узгодити ці сигнали й отримати з них стійкі ознаки.
У дослідженні пропонується модель із кількома поданнями та контрастивним попереднім навчанням. Кожне подання кодується окремою графовою нейромережею (R-GCN для типізованих взаємодій, GAT/GCN для семантичних і зв'язків відправок). Усередині подань виділяються сусідство, між поданнями та гнучко змішуються сигнали, щоб автоматично надавати більше ваги найкориснішому для конкретної відправки. Попереднє контрастивне навчання узгоджує ознаки різних подань і зменшує залежність від великої розмітки. Підсумкове рішення це двокласова класифікація відправок із налаштовуваним порогом під потрібні бізнес-метрики.
В роботі спершу формується практична схема (Рис. 1) з кількома поданнями веб-взаємодій (user, submission, semantic) і набором ознак, що зберігають приватність. Далі будується архітектуру, яка поєднує обробку всередині кожного подання та узгодження між поданнями з гнучким змішуванням сигналів. Поєднується контрастивне попереднє навчання між поданнями з навчанням на мітках, щоб зменшити потребу у великих наборах даних і підвищити переносимість моделі. На додачу надаємо відтворюваний процес обробки на Python і план оцінювання з розділенням даних за часом, сильними базовими порівняннями та експериментами з вимкненням окремих частин моделі. Дослідження орієнтоване на практику враховується потік даних, вимоги до швидкості, потребу мінімізувати зайві спрацювання та питання етики (анонімізація персональних даних, додаткові перевірки для рішень із високою впевненістю).

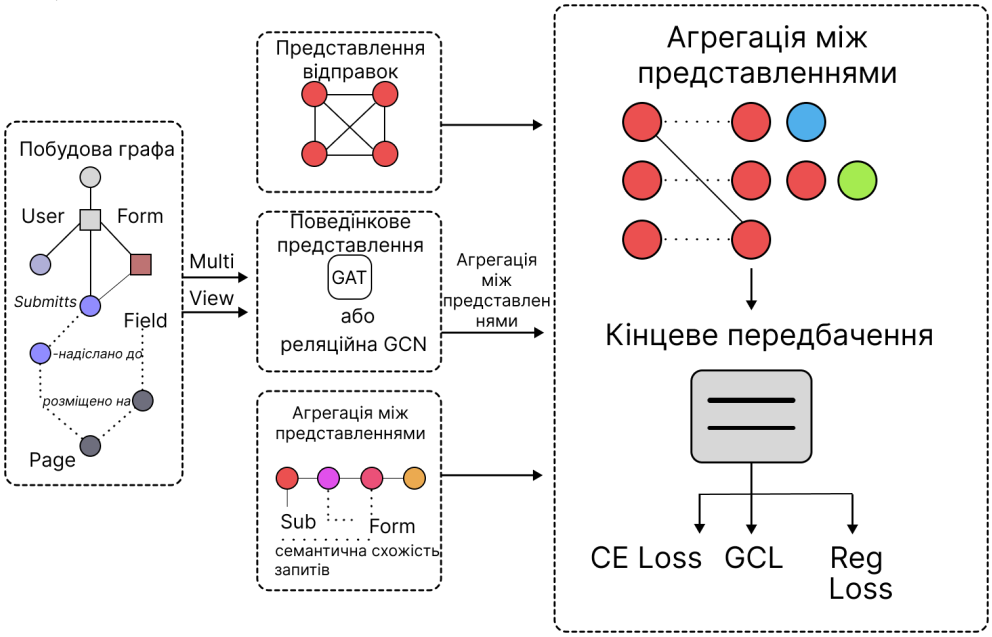

Рис. 1 Багатопредставленнєва графова модель фільтрації веб-спаму (з контрастивним навчанням)

2. Модель

Маємо множину підграфів (view) $G = \{G^{(v)}\}_{v\in\{s,u,t\}}$, де кожне $G^{(v)} = \left(V^{(v)}, E^{(v)}\right)$, тут: $v = s$ — submission-view (вузли: відправки форм), $v = u$ — user-view (гетерогенний граф користувачів/форм/сторінок із типами ребер), $v = t$ — semantic/text-view (граф kNN за

семантичною близькістю тексту/URL) Запропонована мультипредставленнєва контрастивна GNN, що оперує трьома представленнями (s, u, t).

Щоб кодувати дані, для кожного v маємо матрицю ознак вузлів $X^{(v)} \in R^{|V^{(v)}| \times d_v}$, (де d_v — розмірність ознак у представленніv) та матрицю суміжності: $A^{(v)} \in \{0,1\}^{|V^{(v)}| \times |V^{(v)}|}$ У поведінковому представленні v є типи відносин $R^{(u)}$ $(submits, shares\_device, shares\_IP, located\_on)$, тому суміжність розкладаємо на $\{A_r^{(u)}\}_{r \in R^{(u)}}$ для подальшої R-GCN-агрегації. Супервізія здійснюється над підмножиною відправок $L \subseteq V^{(s)}$ з мітками $y_i \in \{0,1\}$. Для локального контексту позначимо $N_i^{(v)}$ — сусідство вузла i (включно із self-loop) у представленні v. Архітектуру описуємо пошарово: $l = 0, ..., L-1$, перехід $d_l! \to d_{l+1}$ параметризується матрицями $W^{(v,l)} \in R^{d_{l+1} \times d_l}$; у R-GCN додатково використовуємо $W_r^{(l)}$ (для кожного типу ребра) та $W_0^{(l)}$ (self-loop).

Цільові мітки задані для підмножини відправок: $L \subseteq V^{(s)}$, $y_i \in \{0,1\}$, шукаємо $f_\theta: V^{(s)} \to [0,1])$з$(\hat{y}_i = f_\theta(i)$ — імовірність спаму для submission-вузла i використовуючи структуру та ознаки всіх представлень. Для вирівнювання між представленнями існують якірні відповідності для одних і тих самих сутностей (відправка i у s має відображення на відповідний вузол/ембеддинг у t та суміжні сутності у u); ці відповідності використовуються надалі в контрастивному навчанні.

Часова організація даних: кожен вузол/ребро має часову мітку τ; тренування/валідація/тест відокремлюються за зрізами часу $\left(T_{train} < T_{val} < T_{test}\right)$ для уникнення витоку. В онлайні граф будується у ковзному вікні ширини W днів (де W — фіксований гіперпараметр).

Виходячи з того, що мітки можуть бути неповними й шумними: спаму мало, частина позначень помилкова або відсутня, ознаки теж можуть спотворюватися: тексти маскують, IP та User-Agent часто спільні або підмінені (через NAT, VPN, емулятори). Відповідності між різними поданнями не завжди повні: для частини відправок немає коректної «прив'язки» у всіх поданнях, тому допускаємо пропуски. Розподіли ознак і зв'язків змінюються з часом, тож оцінювання проводимо у хронологічному розрізі (розділення за часом). Для приватності ідентифікатори (e-mail, телефон, IP) анонімізуємо: хешуємо, узагальнюємо IP до рівня /24 тощо, зберігаємо лише агреговані або хешовані атрибути. Дані великі й розріджені, тому навчаємося пакетно, з вибіркою сусідів. У графі користувачів і у графі відправок зв'язки спрямовані, а в семантичному графі їх зазвичай робимо двосторонніми (kNN) для стабільності. Вартість помилок у продакшені несиметрична: хибні спрацьовування й пропуски коштують по-різному, тож поріг класифікації налаштовуємо під потрібні бізнес-метрики.

Є три подання даних (Рис. 1): відправки (s), користувачі/технічні ознаки (u) та зміст — тексти/URL (t). У t-поданні вузлами також є відправки, які з'єднано kNN-ребрами за схожістю їхніх векторів тексту/URL. Для кожного подання використовується простий графовий шар: для s і t — стандартні GAT/GCN, для u — шар із типами зв'язків (аналог R-GCN). Усередині кожного подання виконуємо агрегування сусідів (GAT/GCN для s і t; R-GCN для u). Далі три вектори зливаються із маскою на відсутні подання та адаптивними вагами, щоб модель надавала більшу вагу найінформативнішому сигналу. Об'єднаний вектор подається до простого класифікатора, який видає ймовірність спаму. Навчання моделі поєднує звичайну бінарну ціль на мічених прикладах і контрастивне узгодження між поданнями; це зменшує потребу у великій розмітці та підвищує стійкість.

Після інтра-агрегації для кожної відправки i маємо верхньошарові вектори $\overline{h_i^{(v)}}$ з представлень $v \in \{s, u, t\}$. Далі працює модуль злиття, який за допомогою навчуваних параметрів обчислює вузло-залежні ваги $\gamma_i^{(v)}$. Ці ваги інтерпретуються як внесок відповідного представлення у рішення та нормуються по доступних каналах; відсутність каналу фіксуємо бінарною маскою

$m_i^{(v)} \in \{0,1\}$ (0 — канал відсутній, 1 — присутній), тож злиття враховує лише ті $\overline{h_i^{(v)}}$, для яких $m_i^{(v)} = 1$. Отримане підсумкове подання позначаємо z_i: це зважена (через $\gamma_i^{(v)}$ та $m_i^{(v)}$) комбінація верхньошарових векторів із внутрішньою нелінійністю модуля злиття. Далі класифікатор із параметрами (w, b) перетворює z_i на $\hat{y}_i \in [0,1]$ — оцінку ймовірності спаму. Завдяки $\gamma_i^{(v)}$ модуль злиття є інтерпретованим: для кожної відправки видно, з якого каналу (поведінкового u, семантичного t чи відправок s) модель бере основний сигнал (найбільше $\gamma_i^{(v)}$) Навчання поєднує бінарну ціль на мічених вузлах (L_C) та контрастивне узгодження між представленнями (L_{GGL}), яке зближує подання тієї самої відправки і відштовхує випадкові пари, що зменшує залежність від обсягу розмітки та підвищує робастність до дрифтів. Для ефективності застосовуємо семплінг сусідства, а семантичні зв'язки у t - представленні попередньо формуємо через kNN-індексування текстів/URL.

Для submission-представлення використовуються числові ознаки форми: довжина тексту, символьна ентропія, частка цифр, глибина/довжина URL, година доби та показник сплеску активності (кількість відправок у вікні Δt) Для semantic/text-представлення вузлами також є відправки, а ознаками — попередньо обчислені векторні подання тексту/URL (напр., 256-вимірні sentence embeddings); суміжність утворюємо методом kNN за косинусною подібністю. Для behavior/user-представлення формуємо гетерограф: для вузлів типу user/device/ip/page беремо агреговані показники активності (частоти дій, інтервали, різноманітність сторінок, кількість зв'язків), категоріальні ознаки подаємо у вигляді векторних подань фіксованої розмірності; суміжність розкладаємо за типами відносин (submits, shares_device, shares_IP, located_on). Навчання та оцінювання відбуваються у часових сплітах (Train/Val/Test), усі перетворення (скейлінг, PCA, kNN-індекси) налаштовуються лише на Train; через дисбаланс класів основною метрикою є PR-AUC, додатково звітуємо F1@τ (τ обирається на Val) та Recall@FP-rate при 10^{-3} і 10^{-4} . Порівнюється модель з поширеними базовими підходами (TF-IDF+LR, Char+XGB, BERT-CLS, одно-представленнєві GNN); окремо оцінюється внесок компонентів шляхом послідовного вимкнення окремих представлень, модуля GCL та застосування альтернативних схем злиття й аугментацій. Додатково метод зіставляється з двома простими правилами — фільтром за списком ключових слів та фіксованим лімітом на кількість відправок з одного IP/пристрою за інтервал часу — і звітуються значення PR-AUC, F1 за порогом, визначеним на валідації, а також Recall за фіксованих рівнів FP 10^{-3} та 10^{-4} з 95% довірчими інтервалами на однакових часових сплітах.

Висновки

Запропонована модель, як спосіб фільтрувати веб-спам на основі графів. Задача в тому, що дані розглядаються з кількох боків — відправки форм, поведінка користувачів і зміст текстів/посилань — а модель попередньо навчається так, щоб ці джерела добре узгоджувалися між собою. Для кожного боку використовується окрема модель; далі їхні сигнали поєднуються так, щоб для кожної відправки більшу вагу отримували саме корисні ознаки. Такий підхід менше залежить від великої кількості розмічених прикладів, краще переноситься на нові кампанії і його простіше пояснити — видно, які зв'язки вплинули на рішення. Додатковий блок узгодження між «поведінкою» і «змістом» підвищує стійкість до маскування текстів, змін тактик і шуму в технічних полях. Поєднання навчання на мітках із попереднім узгодженням ознак дає змогу ефективно вчитись і за браку міток, і за умов регулярного оновлення.

Якість і швидкість побудови графів залежать від даних (особливо від пошуку схожих елементів у дуже великих наборах). Підхід чутливий до неповних відповідників між частинами графа; крім того, потрібне регулярне оновлення списків пошуку і повторне навчання моделі, коли дані з часом змінюються. Загалом, підхід із кількома поданнями та попереднім узгодженням ознак забезпечує збалансоване поєднання точності, стійкості та

зручності для реальних систем фільтрації веб-спаму, а також дає повторювані результати і контрольовані ризики під час експлуатації.

Список літератури

[1] Weisen Pan, Jian Li, Lisa Gao, Liexiang Yue, Yan Yang, Lingli Deng, Chao Deng, Semantic Graph Neural Network: A Conversion from Spam Email Classification to Graph Classification, Scientific Programming, vol. 2022, Article ID 6737080, 8 pages, 2022. doi:10.1155/2022/6737080.

[2] Chensu Zhao, Yang Xin, Xuefeng Li, Hongliang Zhu, Yixian Yang, Yuling Chen, An Attention-Based Graph Neural Network for Spam Bot Detection in Social Networks, Applied Sciences, 10(22) (2020) 8160. doi:10.3390/app10228160.

[3] Linjie Shen, Yanbin Wang, Zhao Li, Wenrui Ma, SMS spam detection using BERT and multi-graph convolutional networks, International Journal of Intelligent Networks, 6 (2025) 79–88. doi:10.1016/j.ijin.2025.06.002.

[4] Iryna Zamrii, Ivan Shakhmatov, Vladyslav Yaskevych, BlockchainSQLSecure: Integration of Blockchain to Strengthen Protection Against SQL Injections, Bulletin of Taras Shevchenko National University of Kyiv. Series: Physics and Mathematics, 1 (2024). doi:10.17721/1812-5409.2024/1.29.

[5] Jiangnan Tang, Youquan Wang, Jie Cao, Haicheng Tao, Guixiang Zhu, Inter- and Intra-Graph Attention Aggregation Learning for Multi-relational GNN Spam Detection, Procedia Computer Science, 214 (2022) 1522–1530. doi:10.1016/j.procs.2022.11.339.

[6] Cheng Wu, Chaokun Wang, Jingcao Xu, Ziyang Liu, Kai Zheng, Xiaowei Wang, Yang Song, Kun Gai, Graph Contrastive Learning with Generative Adversarial Network, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), ACM, 2023, pp. 2721–2730. doi:10.1145/3580305.3599370.

[7] Xuxu Zheng, Chen Feng, Zhiyi Yin, Jinli Zhang, Huawei Shen, Research on Fraud Detection Method Based on Heterogeneous Graph Representation Learning, Electronics, 12(14) (2023) 3070. doi:10.3390/electronics12143070.

[8] Jinbo Chao, Chunhui Zhao, Fuzhi Zhang, Network Embedding-Based Approach for Detecting Collusive Spamming Groups on E-Commerce Platforms, Security and Communication Networks, vol. 2022, Article ID 4354086, 11 pages. doi:10.1155/2022/4354086.

[9] Luzhi Wang, Yizhen Zheng, Di Jin, Fuyi Li, Yongliang Qiao, Shirui Pan, Contrastive Graph Similarity Networks, ACM Transactions on the Web, 18(2) (2024) Article No. 17, pp. 1–20. doi:10.1145/3580511.

[10] Yupeng Hou, Binbin Hu, Wayne Xin Zhao, Zhiqiang Zhang, Jun Zhou, Ji-Rong Wen, Neural Graph Matching for Pre-training Graph Neural Networks, in: Proceedings of the 2022 SIAM International Conference on Data Mining (SDM), SIAM, 2022, pp. 738–747. doi:10.1137/1.9781611977172.20.

[11] Ijeoma A. Chikwendu, Xiaoling Zhang, Chiagoziem C. Ukwuoma, Okechukwu C. Chikwendu, Yeong Hyeon Gu, Mugahed A. Al-antari, Spectrum-Constrained and Skip-Enhanced Graph Fraud Detection: Addressing Heterophily in Fraud Detection with Spectral and Spatial Modeling, Symmetry, 17(4) (2025) 476. doi:10.3390/sym17040476.

[12] Saif Safaa Shakir, Leyli Mohammad Khanli, Hojjat Emami, Convolutional Graph Network-Based Feature Extraction to Detect Phishing Attacks, Future Internet, 17(8) (2025) 331. doi:10.3390/fi17080331.

[13] Zhiwei Liu, Yingtong Dou, Philip S. Yu, Yutong Deng, Hao Peng, Alleviating the Inconsistency Problem of Applying Graph Neural Network to Fraud Detection, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), ACM, 2020, pp. 1569–1572. doi:10.1145/3397271.3401253.

[14] María Novo-Lourés, David Ruano-Ordás, Reyes Pavón, Rosalía Laza, Silvana Gómez-Meire, José R. Méndez, Enhancing representation in the context of multiple-channel spam filtering, Information Processing & Management, 59(2) (2022) 102812. doi:10.1016/j.ipm.2021.102812.