УДК 004.056.55

Назаров В.О.

Національний університет «Одеська політехніка»

АВТОМАТИЗОВАНИЙ МЕТОД РИЗИК-ОРІЄНТОВАНОГО ВИЯВЛЕННЯ ПРОБЛЕМНИХ ПРОФІЛІВ У СОЦМЕРЕЖАХ

Вступ. У соціальних мережах системно діють проблемні профілі, що використовуються для розповсюдження спаму, фішингових повідомлень [1] і скоординованих інформаційних операцій з боку країни-агресорки. Їх відстеження ускладнюють короткий життєвий цикл і ротація обліковок, мімікрування під легітимні спільноти, варіативність мовних патернів та обфускація посилань, а також масштаби потоку контенту, які не покриваються ручною модерацією.

Правила й «чорні списки» швидко деградують, знижуючи відтворюваність результатів у часі та між мовами. Доцільним є компактний підхід, що оперує небагатьма, але інформативними ознаками на рівні повідомлення (можлива винагорода, часовий тиск, емоційне забарвлення, наявність зовнішнього посилання) з імовірнісним рішенням на базі сигмоїдної активації [2].

Водночас для каналу електронної пошти існують продуктивні гібридні нейромережеві методи [3]; однак у контексті соціальних мереж із публічними API перевага запропонованого підходу полягає в мілісекундній латентності, інтерпретованості чинників і можливості прямого калібрування та налаштування порогів під канал/мову, що спрощує прозоре агрегування ризику до рівня профілю за часовими вікнами активності в реальному часі.

Мета: Сконструювати та обґрунтувати технічне рішення для автоматизованого пошуку й відсікання проблемних профілів у соціальних мережах на базі чотирифакторної моделі ризику з імовірнісним класифікаційним правилом та агрегуванням ознак на рівні профілю.

Основна частина

Методика ґрунтується на припущенні, що ознаки зловживань у соціальних мережах проявляються спершу на рівні окремих повідомлень (пости, коментарі, приватні звернення), а вже потім - на рівні профілю як сукупності таких дій. Для кожного повідомлення обчислюються чотири інформативні показники: R_m- ступінь обіцянки винагороди, T_m- виразність часових обмежень/терміновості, E_z- інтенсивність емоційного тиску (імперативи, залякування, нав'язливі обіцянки), R_z - наявність зовнішнього посилання. Перші три оцінюються у шкалі від 0 до 10, останній є бінарним (0 - відсутність, 1 - наявність). Таке компактне подання дозволяє застосувати швидку, інтерпретовану модель ризику та підтримувати низьку латентність у великих потоках даних.

Імовірність фішингової/маніпулятивної природи повідомлення визначається логістичною моделлю :

$$p = \sigma(\beta_0 + \beta_1 R_m + \beta_2 T_m + \beta_3 E_z + \beta_4 R_z) \quad (1)$$

де σ(·) - логістична функція, а $\beta_0 \dots \beta_4$ - параметри, отримані під час навчання на розміченому корпусі. Рішення на рівні повідомлення формулюється словами: якщо оцінка ризику p не нижча за обраний поріг, повідомлення вважається

підозрілим; інакше - звичайним. Поріг добирають з урахуванням допустимої частоти хибних спрацьовувань і специфіки каналу/мови; для підвищення стабільності модельні ймовірності калібруються окремо для різних каналів і локалей. Для профілю користувача ризик агрегується в межах рухомого часового вікна W за правилом:

$$S_u(W) = 1 - \prod_{t \in W}(1 - p_t) \quad (2)$$

де p_t - імовірності для окремих повідомлень профілю; величина $S_u(W)$ інтерпретується як інтегральний ризик профілю за період, що зростає як з частотою підозрілих повідомлень, так і з їх індивідуальними оцінками.

Технічна реалізація методу передбачає наскрізний контур «дані → фактори → інференс → агрегування → політика». Наскрізний контур у цій роботі розглядається як окремий програмний продукт, спроєктований для експлуатації в реальному часі у середовищі соціальних мереж. Конкретні технічні засоби для кожного етапу контуру наведено у структурованому вигляді (таблиця 1).

На етапі отримання даних застосовується офіційний інтерфейс платформи (наприклад, публічні ендпоінти Facebook для сторінок і груп) з дотриманням політик доступу, обмежень швидкості та вимог конфіденційності; приватні повідомлення не обробляються.

Текст нормалізується за мовою (токенізація, лематизація, усунення стоп-слів), після чого визначаються R_m, T_m, E_z за контрольованими лексиконами та правилами, а R_z - шляхом аналізу URL (пунікод-нормалізація, виявлення піддоменів, редиректів, параметрів відстеження).

Обчислення p виконується у легкому інференс-сервісі, здатному працювати з мілісекундними затримками; далі формується $S_u(W)$ і приймається рішення на рівні профілю з урахуванням обраного порога для агрегованого ризику. Для підвищення надійності передбачено калібрування ймовірностей, регулярний підбір порогів під цільові метрики (наприклад, утримання FPR у заданих межах), реєстр версій моделі та журналювання рішень для періодичного донавчання.

Адаптивність забезпечується керованим оновленням лексиконів: прикордонні випадки передаються на експертну перевірку (людина-в-контурі), підтверджені нові тригери надходять до кандидатного словника, який після частотних і контекстних фільтрів промотується в робочий. Моніторинг дрифту ознак і якості моделі (розподіли R_m, T_m, E_z, R_z, стабільність калібрування, динаміка FPR/TPR) дозволяє виявляти зміни атакувальних патернів і своєчасно коригувати як словники, так і параметри моделі без втрати відтворюваності.

Таблиця 1 – Технічні засоби та методи реалізації

Збір і підготовка даних		Модель і експлуатація	
Етап	Засіб / метод	Етап	Засіб / метод
1	2	3	4
Доступ до платформи	Публічні API (напр., Graph API), OAuth, квоти	Класифікація повідомлень	Логістична модель (ONNX Runtime), ймовірність у [0;1]
Потік подій	Kafka / RabbitMQ; retries з exponential backoff + jitter	Калібрування ймовірностей	Platt або Isotonic (per-channel / per-locale)

1	2	3	4
Визначення мови	fastText LID, langdetect	Підбір порогів повідомлення/профілю	ROC/PR-аналіз; line/grid search під цільовий FPR
Нормалізація тексту	spaCy / Stanza; токени, леми, стоп-слова	Агрегування ризику профілю	Інтегральний ризик у рухомому вікні; згладжування
Оцінка R_m, T_m, E_z	Керовані лексикони та правила (шкала 0–10)	Метрики якості та затримок	AUC, FPR/FNR, Recall@FPR≤x, latency p95
Виявлення R_z (посилання)	idna/punycode, tldextract, редиректи, URL-патерни	Моніторинг дрифту	PSI/JS для факторів ризику; алерти
Безпека даних	TLS/mTLS, hashing PII, шифрування на диску	Деплой та масштабування	Docker, Kubernetes; канарейкові релізи, rollback
Human-in-the-Loop	Адмін-UI, Label Studio; модерація «прикордонних» кейсів	Інтеграції / реагування	Webhooks; реєстр профілів із ризиком ≥ порога; маршрути модерації

Висновок. Подано автоматизований ризик-орієнтований підхід до виявлення проблемних профілів у соціальних мережах, що поєднує чотирифакторну оцінку на рівні повідомлень з агрегуванням ризику у часовому вікні профілю та реалізується як завершений програмний продукт за контуром «дані → фактори → інференс → агрегування → політика». Підхід вирізняється компактністю, інтерпретованістю й низькою латентністю, забезпечує калібрування під канали та мови і керованість хибних спрацьовувань через порогові налаштування. Практичну придатність і засоби розгортання систематизовано у структурі технічних засобів та процедур (таблиця 1), що гарантує інтеграцію з офіційними API, журналювання та контроль якості в експлуатації.

Перелік використаних джерел.

1. Штонда Р., Черниш Ю., Терещенко Т., Терещенко К., Цикало Ю., Поліщук С. Класифікація та методи виявлення фішингових атак. Кібербезпека: освіта, наука, техніка. 2024. Т. 4, № 24. С. 69–80. DOI: 10.28925/2663-4023.2024.24.6980.

2. Назаров В. О., Садченко А. В., Кушніренко О. А. Алгоритм виявлення фішингу в листах месенджерів та електронної пошти із використанням нейронних мереж з фіксованою кількістю ранжованих чинників ризику. Інформатика та математичні методи в моделюванні. 2025. Т. 15, № 2. С. 247–259.

3. Фещенко Є. О., Заболотня Т. М. Метод автоматизованого виявлення фішингу в електронних листах на основі гібридної нейромережевої архітектури. Наукові праці Вінницького національного технічного університету. 2025. № 2. С. 145–154. DOI: 10.31649/2307-5376-2025-2-145-154.