КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 4 (32), 2026
ISSN 2663 – 4023
DOI 10.28925/2663-4023.2026.32.1168
УДК 004.056.5:004.853:004.774.1

Бучик Сергій Степанович
Доктор технічних наук, професор, професор кафедри кібербезпеки та захисту інформації
Київський національний університет імені Тараса Шевченка, Київ, Україна
ORCID: 0000-0003-0892-3494
buchyk@knu.ua

П'ятигор Віталій Петрович
Аспірант кафедри кібербезпеки та захисту інформації
Київський національний університет імені Тараса Шевченка, Київ, Україна
ORCID: 0000-0002-7621-1299
vp5gor@knu.ua

МОДЕЛЬ ВИЗНАЧЕННЯ АВТОМАТИЗОВАНИХ АКАУНТІВ СОЦМЕРЕЖ НА ОСНОВІ БАГАТОВИДОВОГО МЕТОДУ З МЕХАНІЗМОМ УВАГИ

Анотація. У статті розглянуто задачу виявлення бот-акаунтів у соціальних мережах як проблему аналізу різнорідних даних, що включають поведінкові, атрибутивні, текстові та візуальні характеристики користувача. Зростання кількості автоматизованих акаунтів, здатних імітувати реальну поведінку, обумовлює необхідність розробки нових підходів, які забезпечують не лише високу точність класифікації, але й пояснювальність отриманих результатів. У роботі запропоновано багатовидову модель виявлення бот-акаунтів, яка базується на використанні багатовидового методу Gradient Boosting та механізму уваги для інтеграції результатів аналізу окремих представлень. Запропонований підхід передбачає розгляд кожного типу даних як окремого представлення, для якого формується власний вектор ознак і навчається локальна модель. Поведінкове представлення враховує часові характеристики активності користувача, атрибутивне – властивості профілю, контентне – текстові особливості дописів із використанням семантичних моделей, а візуальне - ознаки зображень, включаючи аналіз метаданих та OCR. Для кожного представлення обчислюється локальна оцінка ймовірності бот-активності та показник якості даних. Інтеграція результатів виконується за допомогою механізму уваги, який визначає вагу кожного представлення залежно від його інформативності та достовірності. Це дозволяє адаптивно враховувати неповноту або неоднорідність даних та підвищує стійкість моделі до відсутності окремих типів інформації. Фінальна оцінка формується на основі зваженого об'єднання представлень із використанням метамоделі Gradient Boosting. Особливістю запропонованого підходу є можливість інтерпретації результатів за рахунок оцінки внеску кожного представлення та визначення рівня впевненості моделі. Це забезпечує прозорість прийнятих рішень і дозволяє використовувати модель у практичних системах аналізу соціальних мереж. Запропонований підхід розширює існуючі методи виявлення ботів за рахунок поєднання багатовидового навчання, адаптивної агрегації та врахування якості даних, що підвищує ефективність розробленої системи.

Ключові слова: автоматизовані акаунти; соціальні мережі; Gradient Boosting; багатовидова модель; multi-view learning.

ВСТУП

Зі зростанням популярності соціальних мереж значно збільшується кількість автоматизованих акаунтів (ботів), які використовуються для поширення спаму, маніпуляцій інформацією та впливу на суспільну думку. Розвиток цифрових платформ сприяє масштабуванню таких загроз, що негативно впливає на достовірність інформації та довіру користувачів. Дослідження показують, що поширення ботів є системною проблемою сучасних онлайн-систем і потребує ефективних методів виявлення [1]. У

зв'язку з цим актуальним є створення інтелектуальних систем аналізу, здатних автоматично ідентифікувати підозрілу активність.

Постановка проблеми. Існуючі підходи до виявлення ботів часто базуються на аналізі окремих типів даних, таких як поведінкові характеристики, профіль користувача або текстовий контент. Проте такі методи не враховують комплексну природу сучасних акаунтів та їх здатність маскуватися під реальних користувачів. Крім того, сучасні соціальні мережі мають складну структуру взаємозв'язків, що ускладнює аналіз через високу розмірність і неоднорідність даних . Відсутність єдиного підходу до інтеграції різнорідних джерел інформації обмежує ефективність існуючих моделей.

Стаття присвячена розробці підходу, який дозволяє ефективно інтегрувати різнорідні джерела даних, враховувати їхню якість та забезпечувати інтерпретованість результатів класифікації.

Аналіз останніх досліджень і публікацій. Останні дослідження у сфері виявлення ботів активно використовують методи машинного навчання, а саме глибокі нейронні мережі та графові моделі. Зокрема, Graph Neural Networks дозволяють враховувати структуру соціальних зав'язків та покращують точність виявлення [2]. Інші підходи поєднують мовні моделі та графові представлення для інтеграції семантичних і структурних ознак [3]. Водночас існуючі рішення часто мають обмежену інтерпретованість або недостатньо ефективно інтегрують різні типи даних.

Мета статті. Метою статті є розробка багатовидової моделі виявлення бот-акаунтів з механізмом уваги, яка інтегрує різнорідні дані користувача та забезпечує інтерпретованість результатів.

ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

Системи виявлення бот-акаунтів у соціальних мережах є складними інформаційно-аналітичними системами, що поєднують методи аналізу даних, машинного навчання та обробки природної мови. Основною метою таких систем є автоматизоване визначення ймовірності того, що конкретний акаунт або його активність має штучний, автоматизований характер. У загальному вигляді ці системи формулюються як задачі бінарної класифікації, де кожному об'єкту ставиться у відповідність мітка «бот» або «реальний користувач» [4].

Системи виявлення ботів мають модульну архітектуру, яка включає підсистеми збору даних, формування ознак, аналізу та агрегації результатів [5]. Підсистема збору даних забезпечує отримання інформації про користувача, включаючи поведінкові характеристики, метадані профілю, текстовий та мультимедійний контент. На етапі формування ознак виконується перетворення сирих даних у структурований вектор ознак, що відображає різні аспекти активності користувача.

Підсистема аналізу, як правило, реалізується у вигляді набору спеціалізованих моделей, орієнтованих на різні типи даних. Поведінковий аналіз дозволяє виявляти аномальні часові патерни активності, аналіз атрибутів профілю – нетипові характеристики акаунта, тоді як контентний аналіз використовує методи обробки природної мови для виявлення шаблонності або маніпулятивного змісту [6].

Одним із перспективних методів аналізу є багатовидовий підхід, у межах якого кожен тип даних розглядається як окреме представлення, а їх інтеграція виконується на рівні метамоделі [7]. Такий підхід дозволяє ефективно поєднувати різнорідні джерела інформації та враховувати складні залежності між ними.

Новизна полягає у розробці багатовидової моделі виявлення бот-акаунтів, яка поєднує правила-орієнтовані ознаки та методи машинного навчання в єдиній архітектурі, використовує підхід Multi-View Gradient Boosting для інтеграції результатів окремих підсистем, а також включає механізм уваги, який включає якість даних і забезпечує інтерпретованість результатів через оцінку внеску кожного представлення у фінальне рішення.

Таким чином, запропонований підхід розширює існуючі методи за рахунок поєднання гетерогенних джерел даних, адаптивної агрегації та пояснюваності, що підвищує ефективність і практичну придатність систем виявлення ботів.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

В результаті даного дослідження авторами пропонується використання багатовидової моделі на основі Gradient Boost з механізмом уваги для систем виявлення ботів у соціальних мережах, яка може бути застосована на етапі аналізу таких систем.

У запропонованій моделі аналізу бот-акаунтів інформація про користувача формується з кількох незалежних джерел: ознаки поведінки користувача, атрибутів користувача, текстового контент допису та візуальних ознак, які обчислюються за допомогою окремих моделей у відповідних сервісах. Кожен із цих типів даних описує один і той самий акаунт, але з різних точок зору, будучи специфічним представленням (або view) у межах цього методу. Таким чином вхідний вектор ознак для представлень можна визначити як множину ознак.

Кожне з представлень даного методу визначає локально ймовірність того, чи є акаунт ботом чи ні використовуючи лише притаманні даному представленню ознаки. Тому кожне представлення має власну локальну модель на основі Gradient Boosting, яку можна описати формулою 1.

$$F^{(v)}(x^{(v)}) = \sum_{m=1}^{M_v} \gamma_m^{(v)} h_m^{(v)}(x^{(v)}) \quad (1)$$

де v – тип представлення, $x^{(v)}$– вектор ознак відповідного представлення, h_m – дерево для відповідного представлення, γ_t – вага дерева.

Також для кожного етапу формується ймовірність того, що акаунт є автоматизованим за формулою 2, та показник впевненості системи у отриманому результаті за формулою 3. Дана інформація є корисною для аналізу впливу представлення на фінальну оцінку та пояснення рішення про акаунт.

$$s^{(v)} = \sigma(F^{(v)}(x^{(v)})) \quad (2)$$

де σ(·) – функція нормалізації на проміжку [0,1].

$$c^{(v)} = \alpha \cdot | s^{(v)} - 0.5 | + \beta \cdot q^{(v)} \quad (3)$$

де $q^{(v)}$ – якість даних для окремого представлення, α, β– коефіцієнти ваг.

Як зазначено у формулі 1, для кожного представлення визначено певний набір ознак, які характерні для даного представлення, та використовуються для прийняття рішення щодо акаунту. Тому розглянемо, які ознаки акаунту користувача використовуються у запропонованій системі.

1. Представлення поведінки користувача.

Це представлення відповідає за аналіз інтенсивності, регулярності та ритму постингу, виявлення автоматизованих шаблонів поведінки у часі та формування ймовірнісних поведінкових сигналів. Представлення оперує ознаками користувача типу «частота», які були обчислені на етапі вибірки ознак. Використовується декілька паралельних правил, що визначають середню людську поведінку і встановлюють прапори бот активності, коли значення є аномальними. Можна визначити 5 основних груп правил:

1. Правила інтенсивності створення дописів
   a. Надмірна кількість дописів за день.
   b. Відносно велика кількість дописів за день при новоствореному акаунті.
2. Правила затримки між постами
   a. Дуже мала середня затримка.
   b. Неможлива для людини швидкість створення дописів.
   c. Стабільна затримка між дописами.
3. Правила регулярності активності
   a. Майже ідеальний ритм.
   b. Стабільний ритм протягом тривалого часу.
   c. Активність кожен день
4. Правила сплескової поведінки
   a. Рівномірно розподілені у часі дописи.
   b. Класифікація поведінки акаунта за кластерами нормальна людська, сплескова людська, постійна автоматизована та сплескова автоматизована.
5. Правила циркадного ритму
   a. Відсутність природнього добового циклу.
   b. Активність 24/7
   c. Статистичне відхилення від людської норми

Для розрахунку метрик ритмічності використовуються наступні методи. Оцінка регулярності постингу визначається за формулою 4 нормалізується в [0,1].

$$\text{posting\_regularity\_score} = 1 - \frac{std\_delay\_minutes}{mean\_delay\_minutes+\epsilon}, \quad (4)$$

де ε – мале число.

Показник сплесковості вказує чи дописи користувача створюються регулярно або мають сплескові шаблони. Звичайні користувачі зазвичай мають більш сплесковий характер дописів у відповідь на певну подію або участь у обговорені з іншими користувачами. Розраховується за формулою Goh & Barabási [8], що наведена у формулі 5.

$$burstiness\_score = \frac{std\_delay\_minutes - mean\_delay\_minutes}{std\_delay\_minutes + mean\_delay\_minutes} \quad (5)$$

Додатково визначається циркадна ентропія для користувача, що вказує на можливу автоматизацю акаунту, наприклад якщо дописи на акаунті створюються кожну годину, що неможливо для звичайного користувача при проміжку спостереження більше тижня. Для обрахунку групуємо дописи по годинах доби та нормалізуємо за формулою 6.

$$circadian\_entropy = \frac{-\sum_{h=0}^{23} p_h \log_2 p_h}{\log_2 24} \quad (6)$$

де p_h – ймовірність допису в певну годину доби

Використовуючи згруповані дані дописів по годинам визначаємо домінантну годину, в якій користувач створює дописи.

У результаті аналізу, представлення має сформувати додатково до ймовірності та впевненості у результаті масив оцінок для основних типів правил а також показник якості даних що були проаналізовані. Цей показник якості є специфічним для кожного представлення та дає змогу зрозуміти чи є дані повними та достатніми для формування висновку. Для поведінкового представлення таким показником будемо вважати кількість дописів за одиницю часу у порівнянні з попередньо визначеним значенням достатньої кількості.

2. Представлення атрибутів користувача. Представлення атрибутів користувача аналізує статичні та мережеві характеристики профілю користувача соціальної мережі з метою визначення ступеня відповідності профілю типовим характеристикам бот-акаунтів. Використовуються структурні характеристики профілю, властивості імені користувача, параметри опису профілю, мережеві показники підписників, службові прапорці акаунта. Визначено наступні правила для цього сервісу:

1. Вік акаунта
2. Ознаки профілю
   a. Відсутність аватару.
   b. Відсутність банеру.
   c. Дуже коротка біографія.
   d. Низька ентропія біографії.
   e. Наявність зовнішнього посилання.
   f. Наявність локації та її відповідність до країни створення акаунту.
3. Ознаки імені користувача
   a. Висока ентропія імені.
   b. Велика кількість цифр.
   c. Дуже довге ім’я.
   d. Наявність емодзі та спеціальних символів у імені користувача.
   e. Кількість змін імені користувача.
4. Мережеві характеристики
   a. Низьке співвідношення followers/following.
   b. Дуже велика кількість підписок.
   c. Відсутність списків.
5. Прапорці акаунта
   a. Перевірений акаунт.
   b. Захищений акаунт.
   c. Наявність бейджу організації.

Особливу увагу варто звертати на прапорці користувача. Після зміни процесу верифікації акаунтів у соціальній мережі X у кінці 2022 року дописи користувачів з прапором верифікації мають більший вплив на алгоритми пошуку та рекомендації контенту у загальній стрічці дописів [9]. Це означає що для того щоб бот акаунт міг більш ефективно поширювати контент така підписка необхідна, на відміну від попереднього процесу верифікації, де для отримання прапорця необхідно було надавати документ, що підтверджує особу.

Для показника якості у даному представленні використовується кількість полів атрибутів користувача, що були задіяні для аналізу у порівнянні з мінімально допустимим. Так як деякі поля є опціональними вимога наявності всіх полів не має сенсу, тому маємо визначити перелік атрибутів, що мають найбільший вплив.

3. Представлення контенту допису користувача. Представлення контенту допису аналізує лише текстовий вміст обраного допису та формує оцінку того, наскільки цей текст відповідає типовим шаблонам ботоподібного, спамоподібного або автоматично згенерованого контенту. Представлення використовує як обчислені на етапі вибору ознак статистичні данні так і повний текст допису.

Даним представленням аналізується поверхнева структура тексту, лексичні шаблони, ознаки шаблонності, статистику взаємодії з дописом та здійснює лексико-семантичний зміст тексту. Аналіз медіаконтенту не виконується.

Хоча остаточна оцінка формується за формулами 1-3, як і у попередніх представленнях, масив ознак для цього представлення поєднує структурні, шаблонні і контекстні ознаки на основі правил та метод XLM-RoBERTa для глибокого семантичного аналізу тексту.

Частина оцінки на основі правил формує інтерпретовані сигнали, які описують поверхневу ботоподібність тексту та контексту допису. Такими правилами є:

1. Структура тексту
   a. Дуже короткий текст.
   b. Неприродно короткий текст при наявності посилання.
   c. Надмірне використання верхнього регістру або емодзі.
2. Лексичні правила
   a. Надмірна кількість або щільність хештегів.
   b. Надмірна кількість згадок.
   c. Велика кілька URL у дописі.
3. Шаблонність
   a. Низька різноманітність символів.
   b. Високий рівень повтору символів.
   c. Високий рівень n-gram повторів.
   d. Надмірна кількість спеціальних символів.
4. Формату допису
   a. Короткий текст з медіа.
   b. Короткий текст з URL та медіа.
5. Взаємодія з дописом
   a. Низька взаємодія при великій кількості переглядів.
   b. Дисбаланс поширень і відповідей.
   c. Аномально низький рівень відповідей.

Семантична підмодель оцінює зміст тексту, а не лише його форму. Для цього використовується XLM-RoBERTa [10], яка добре підходить для багатомовного аналізу коротких текстів соціальних мереж. Модель обрана тому, що вона підтримує багатомовність, добре працює з короткими текстами, дозволяє враховувати контекст, краще за словникові методи виявляє токсичність, приховану промо-семантику, маніпулятивні повідомлення, політичні заклики та смислову розірваність тексту. Семантична модель повертає наступні оцінки:

- sentiment_score – оцінка тональності тексту;
- subjectivity_score – рівень суб’єктивності й емоційної навантаженості;
- toxicity_score – ознака токсичного / агресивного повідомлення;

– spam_semantic_score – семантична схожість з промо- або clickbait-повідомленнями;
– promotional_intent_score – ймовірність рекламного наміру;
– political_intent_score – ймовірність політичної або агітаційної спрямованості;
– semantic_coherence_score – оцінка внутрішньої зв'язності тексту;
– information_density_score – оцінка змістовної насиченості тексту.

Після виконання аналізу обома підмоделями формується загальний висновок сервісу. Таким чином загальна модель визначає ваги кожної з оцінок та правила коли семантичний аналіз важливіший за правила чи навпаки. Зберігаються лише основні оцінки аналізу без передачі всіх значень семантичного аналізу. Існує можливість коли у дописі відстутній текст і сам він складається лише з медіа контенту, тому зберігаємо у полі якості даних значення чи дані аналізу цього сервісу релевантні.

4. Представлення візуального контенту. Представлення візуального контенту призначене для оцінки того, наскільки візуальний контент допису має ознаки, характерні для штучно згенерованих зображень, зображень із текстовими вставками, типовими для спаму. Для даного представлення застосовується структурний аналіз зображення, виділення візуальних тегів, аналіз метаданих, оцінка ймовірності штучно згенерованного зображення та OCR-аналіз тексту на зображенні.

Представлення використовує схожий до представлення контенту підхід гібридної системи, де вектором ознак для фінальної оцінки представлення є ознаки на основі правил, результат моделі визначення штучно згенерованих зображень та моделі розпізнавання тексту на зображені та семантичного аналізу. Підмодель на основі правил забезпечує інтерпретовані сигнали на основі простих властивостей зображення:

1. Метадані
   a. Відсутність EXIF.
   b. Підозріле поле software.
   c. Наявність метадати пов'язаної з ШІ генерацією.
2. Розпізнавання тексту
   a. Чи є текст на зображенні.
   b. Надмірна кількість тексту на зображені.

Підмодель визначення штучно згенерованих зображень оцінює ймовірність, що зображення було створене генеративною моделлю. Сервіс повертає ймовірність штучної генерації, а не жорстку класифікацію. Для цього використовуємо модель класифікації візуальної трансформації ViT та тренуємо модель на датасеті, які має чітко визначені реальні та штучно згенеровані зображення.

Якщо на зображенні є текст, його треба оцінити семантично. Для OCR розпізнавання використовуємо застосунок PaddleOCR [11], який підтримує багатомовність та стійкіший до шуму у зображені. Після виокремлення тексту з зображення, використовуємо ідентичний до представлення контенту підхід з використанням методу XLM-RoBERTa, для визначення семантичних ознак.

Сервіс повертає загальну оцінку, впевненість та основні результати аналізу трьох моделей аналізу зображень, що використовується сервісом. Як і у випадку сервісу контенту, не у всіх дописах є медіа контент, який можемо аналізувати, тому у якості даних зазначаємо чи можуть дані результати бути використану для формування остаточної оцінки.

5. Формування загальної оцінки. Етап формування загальної оцінки призначений для об'єднання результатів кількох незалежних сервісів аналізу в одну інтегральну оцінку ботоподібності акаунта. Сервіс використовує запропонований метод

багатовидової моделі на основі Gradient Boost з механізмом уваги, у межах якого кожне представлення акаунта розглядається як окреме представлення, а фінальна модель навчається враховувати індивідуальну силу кожного представлення, узгодженість або суперечність між ними, нелінійні залежності між частковими оцінками та якість і повноту даних кожного представлення.

Для даної системи можна визначити такі переваги цього методу:

1. Краще використання різних типів даних – різні джерела інформації мають різну структуру. Запропонована модель дозволяє адаптуватися до кожного типу ознак окремо.

2. Менша залежність від одного джерела сигналів – якщо один тип даних неповний (наприклад, немає зображень), модель все одно може працювати з іншими представленнями.

3. Краща інтерпретація результатів – можна аналізувати, яке представення дає найбільший вклад у передбачення.

4. Підвищення точності – у задачах, де є різнорідні дані, multi-view моделі часто працюють точніше, ніж моделі, які використовують усі ознаки разом.

Для кожного представлення акаунта використовується функція оцінки релевантності $g_k(X_k)$, яка перетворює вектор локальних характеристик представлення у скалярну величину, що відображає інформативність та надійність цього представлення для конкретного об'єкта. До складу вектора X_kвходять локальна оцінка сервісу, показник достовірності, характеристики якості даних та агреговані ознаки представлення. Дані оцінки релевантності вираховуються для кожного представлення за формулою 7.

$$g_k(X_k) = w_k^{\top} X_k + b_k \tag{7}$$

де $w_k^{\top}$ – вектор ваг для k-го view, b_k – зсув.

Для використання методу необхідно обрахувати представлення з урахуванням визначених показників якості даних. Для цього обраховуємо коефіцієнт уваги для кожного представлення в залежності від якості даних. Ці коефіцієнти визначають відносну важливість кожного представлення для конкретного акаунту. Коефіцієнт уваги можна розрахувати як відношення оцінки релевантності певного представлення до загальної суми оцінок релевантності всіх представлень за формулою 8.

$$\alpha_k = \frac{\exp(g_k(X_k))}{\sum_j \exp(g_k(X_k))} \tag{8}$$

Після обчислення ваг кожне представлення масштабується використовуючи власний коефіцієнт за формулою 9 та формується об'єднаний зважений вектор за формулою 10:

$$\tilde{X}_k(u) = \alpha_k(u)\, X_k(u) \tag{10}$$

$$\tilde{X}(u) = (\tilde{X}_b(u), \tilde{X}_a(u), \tilde{X}_c(u), \tilde{X}_i(u)) \tag{11}$$

Після чого даний вектор застосовується у стандартній формулі Gradient Boosting наведеної у формулі 1.

Отже, на рівні формування загальної оцінки Gradient Boosting використовується як метамодель, що приймає на вхід сукупність результатів представлення і формує

єдину підсумкову оцінку. Додатково до оцінки чи є акаунт автоматизованим чи ні та впевненості за допомогою моделі можемо визначити вклад кожного з представлення на фінальну оцінку. Це дає змогу надати більш розширенний класифікатор акаунту ніж звичайне бот або ні.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У роботі розроблено багатовидову модель виявлення бот-акаунтів у соціальних мережах, яка поєднує різнорідні джерела даних, зокрема поведінкові характеристики, атрибути профілю, текстовий та візуальний контент. Запропонований підхід базується на використанні локальних моделей Gradient Boosting для кожного представлення та їх інтеграції за допомогою механізму уваги, який враховує якість і повноту даних.

Розроблена модель дозволяє не лише визначити ймовірність бот-активності, але й оцінити впевненість у результаті та внесок кожного представлення у фінальне рішення. Це забезпечує підвищену пояснювальність моделі та можливість пояснення прийнятих рішень, що є важливим для практичного використання.

Запропонований підхід демонструє переваги у порівнянні з традиційними методами за рахунок інтеграції різнорідних даних, адаптивного зважування представлень та врахування якості вхідної інформації.

Подальші дослідження будуть спрямовані на експериментальну оцінку запропонованої моделі на реальних наборах даних, а також вдосконалення механізму оцінки якості даних та його впливу на фінальну класифікацію.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Liu, Y., et al. (2025). Evolution of malicious social bot detection: From individual profiling to group analysis and beyond. *Journal of Social Computing*, *6*(3), 258–284. https://doi.org/10.23919/jsc.2025.0017
2. Huang, H., et al. (2024). CGNN: A compatibility-aware graph neural network for social media bot detection. *IEEE Transactions on Computational Social Systems*, 1–16. https://doi.org/10.1109/tcss.2024.3396413
3. Li, D., et al. (2025). BotLGT: Social bot detection based on LLM and graph transformer. *Neurocomputing*, 131453. https://doi.org/10.1016/j.neucom.2025.131453
4. Varol, O., et al. (2017). Online human-bot interactions: Detection, estimation, and characterization. *Proceedings of the International AAAI Conference on Web and Social Media, 11*(1), 280–289. https://doi.org/10.1609/icwsm.v11i1.14871
5. (2025). Architecture of automated account (bot) detection systems in social networks. *Information Systems and Technologies Security*, *1*(9), 11–17. https://doi.org/10.17721/ISTS.2025.9.11-17
6. Devlin, J., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). https://doi.org/10.18653/v1/n19-1423
7. Sun, S. (2013). A survey of multi-view machine learning. *Neural Computing and Applications*, *23*(7–8), 2031–2038. https://doi.org/10.1007/s00521-013-1362-6
8. Goh, K. I., & Barabási, A. L. (2008). Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, *81*(4), 48002. https://doi.org/10.1209/0295-5075/81/48002
9. OpenTweet. (n.d.). What is Twitter Blue / X Premium? https://opentweet.io/glossary/twitter-blue
10. Hugging Face. (n.d.). XLM-RoBERTa. https://huggingface.co/docs/transformers/en/model_doc/xlm-roberta
11. PaddlePaddle. (n.d.). *PaddleOCR: Turn any PDF or image document into structured data for AI*. GitHub. https://github.com/PaddlePaddle/PaddleOCR

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 4 (32), 2026
ISSN 2663 – 4023

Serhii Buchyk
DSc (Engin.), Prof., Professor of the Department of Cybersecurity and Information Protection
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
ORCID: 0000-0003-0892-3494
buchyk@knu.ua

Vitalii Piatyhor
PhD student of the Department of Cybersecurity and Information Protection
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
ORCID: 0000-0002-7621-1299
vp5gor@knu.ua

SOCIAL MEDIA BOT ACCOUNT DETECTION MODEL BASED ON MULTI-VIEW METHOD WITH ATTENTION MECHANISM

Abstract. The article proposes a multi-view model for detecting bot accounts, which is based on the use of the ensemble Gradient Boosting method and the attention mechanism for integrating the results of the analysis of individual representations. The proposed approach involves considering each type of data as a separate representation, for which its own feature vector is formed, and a local model is trained. The behavioral representation considers the temporal characteristics of user activity, the attributive representation - the properties of the profile, the content representation - the textual features of posts using semantic models, and the visual representation - the features of images, including metadata analysis and OCR. For each representation, a local estimate of the probability of automated activity and a data quality indicator are calculated. The integration of results is performed using an attention mechanism that determines the weight of each representation depending on its informativeness and reliability. This allows us to adaptively take into account incompleteness or heterogeneity of data and increases the model's resilience to the absence of individual types of information. The final score is formed based on a weighted combination of representations using the Gradient Boosting metamodel. A feature of the proposed approach is the ability to interpret the results by assessing the contribution of each representation and determining the level of confidence of the model. This ensures transparency of the decisions made and allows the model to be used in practical social network analysis systems. The proposed approach extends existing bot detection methods by combining multi-species learning, adaptive aggregation, and data quality consideration, which increases the efficiency of the developed system.

Keywords: automated accounts; social networks; Gradient Boosting; multi-view model; multi-view learning.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Liu, Y., et al. (2025). Evolution of malicious social bot detection: From individual profiling to group analysis and beyond. *Journal of Social Computing*, *6*(3), 258–284. https://doi.org/10.23919/jsc.2025.0017
2. Huang, H., et al. (2024). CGNN: A compatibility-aware graph neural network for social media bot detection. *IEEE Transactions on Computational Social Systems*, 1–16. https://doi.org/10.1109/tcss.2024.3396413
3. Li, D., et al. (2025). BotLGT: Social bot detection based on LLM and graph transformer. *Neurocomputing*, 131453. https://doi.org/10.1016/j.neucom.2025.131453
4. Varol, O., et al. (2017). Online human-bot interactions: Detection, estimation, and characterization. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 280–289. https://doi.org/10.1609/icwsm.v11i1.14871
5. (2025). Architecture of automated account (bot) detection systems in social networks. *Information Systems and Technologies Security*, *1*(9), 11–17. https://doi.org/10.17721/ISTS.2025.9.11-17
6. Devlin, J., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). https://doi.org/10.18653/v1/n19-1423
7. Sun, S. (2013). A survey of multi-view machine learning. *Neural Computing and Applications*, *23*(7–8), 2031–2038. https://doi.org/10.1007/s00521-013-1362-6

8. Goh, K. I., & Barabási, A. L. (2008). Burstiness and memory in complex systems. *EPL (Europhysics Letters), 81*(4), 48002. https://doi.org/10.1209/0295-5075/81/48002
9. OpenTweet. (n.d.). What is Twitter Blue / X Premium? https://opentweet.io/glossary/twitter-blue
10. Hugging Face. (n.d.). XLM-RoBERTa. https://huggingface.co/docs/transformers/en/model_doc/xlm-roberta
11. PaddlePaddle. (n.d.). *PaddleOCR: Turn any PDF or image document into structured data for AI*. GitHub. https://github.com/PaddlePaddle/PaddleOCR

Отримано редакцією журналу / Received: 21.01.26
Прорецензовано / Revised: 15.02.26
Схвалено до друку / Accepted: 26.03.26