КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (33), 2026
ISSN 2663 – 4023
DOI 10.28925/2663-4023.2026.33.1266
УДК 004.056.5:004.85:004.774.1(048.8)

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

МЕТОДОЛОГІЧНІ ЗАСАДИ ВИЯВЛЕННЯ АВТОМАТИЗОВАНИХ АКАУНТІВ У СОЦІАЛЬНИХ МЕРЕЖАХ

Анотація. У статті розглянуто методологічні засади виявлення автоматизованих акаунтів у соціальних мережах. Актуальність дослідження зумовлена використанням соціальних ботів для поширення дезінформації, маніпулювання громадською думкою та зростанням можливостей генеративного штучного інтелекту щодо імітації поведінки реальних користувачів. Систематизовано методи виявлення ботів за основними джерелами ознак: атрибутами користувача, поведінковими характеристиками, текстовим контентом, графовими структурами та їх поєднанням. Порівняння виконано за використаними моделями, датасетами, метриками якості, складністю отримання даних, обмеженнями та рівнем пояснюваності рішень. Встановлено, що профільні методи є доступними й відносно інтерпретованими, але недостатньо стійкими до складних ботів; поведінкові підходи ефективні для виявлення автоматизованих патернів і координації, однак потребують історії активності; текстові моделі забезпечують високу якість розпізнавання за змістом дописів, проте залежать від мови, тематики й розвитку генеративного штучного інтелекту; графові методи є перспективними для виявлення бот-мереж, але характеризуються високими вимогами до доступності даних. Визначено, що найбільш перспективним напрямом є застосування гібридних моделей, які поєднують доступні профільні, поведінкові й текстові ознаки з урахуванням пояснюваності результатів та без критичної залежності від повного соціального графа. Узагальнено основні проблеми сучасних методів: адаптивність ботів, неоднорідність і застарівання датасетів, дисбаланс класів, обмеження доступу до даних і недостатню інтерпретованість складних моделей.

Ключові слова: соціальні мережі; автоматизовані акаунти; соціальні боти; виявлення ботів; машинне навчання; гібридні методи; пояснюваність моделей.

ВСТУП

Соціальні мережі стали одними з основних засобів поширення інформації, формування громадської думки та комунікації. Платформи соціальних мереж активно використовуються як для обміну інформацією між користувачами, так і для проведення маркетингових, політичних та інформаційних кампаній. Водночас стрімке зростання популярності соціальних мереж супроводжується появою великої кількості автоматизованих акаунтів – ботів, які здатні імітувати поведінку реальних користувачів.

Сучасні боти використовуються для розповсюдження дезінформації, маніпуляції суспільною думкою, штучного підвищення популярності контенту, проведення інформаційних атак, політичної пропаганди та спам-активності. Розвиток технологій штучного інтелекту та великих мовних моделей значно ускладнив задачу виявлення ботів. Сучасні боти здатні генерувати природний текст за допомогою засобів штучного інтелекту, адаптувати поведінку до контексту та ефективно маскуватися під реальних користувачів. Це призводить до поступового зниження ефективності традиційних методів детекції ботів, які базуються лише на простих статистичних або поведінкових ознаках.

За результатами дослідження Imperva Bad Bot Report 2025, автоматизований бот-трафік вже перевищує половину глобального інтернет-трафіку, а частка шкідливих ботів становить близько 37% від усього інтернет-трафіку [1]. Крім того, дослідження глобальної активності в соціальних мережах показують, що близько 20% інформаційної активності у соціальних мережах генерується ботами [2].

Проблема ботів також безпосередньо пов'язана із поширенням дезінформації. Дослідження демонструють, що соціальні боти активно беруть участь у розповсюдженні низькодостовірного контенту, підсилюють інформаційні кампанії та сприяють формуванню інформаційних «ехо-камер» [3]. У зв'язку з цим задача розробки ефективних методів виявлення ботів у соціальних мережах є актуальною науковою та практичною проблемою.

Постановка проблеми. Сучасні методи виявлення ботів у соціальних мережах базуються на різних підходах та використовують широкий спектр ознак, зокрема атрибути користувача, поведінкові характеристики, текстовий контент, графові структури та їх комбінування [4]. Різноманіття існуючих методів, відмінності у використовуваних наборах даних, критеріях оцінювання та алгоритмах класифікації ускладнюють проведення об'єктивного порівняння їх ефективності, стійкості та пояснюваності. Крім того, розвиток генеративних мовних моделей поступово знижує ефективність частини традиційних підходів до виявлення ботів.

Аналіз останніх досліджень і публікацій. Значна кількість робіт присвячена аналізу поширення дезінформації ботами, виявленню бот-мереж та застосуванню методів машинного й глибокого навчання для задач класифікації ботів. До аналізу включалися роботи, у яких описано метод виявлення соціальних ботів, наведено використані ознаки або архітектуру моделі та подано результати експериментального оцінювання. Відібрані підходи згруповано за основним джерелом даних: атрибути профілю, поведінка, текст, графові зв'язки та гібридне поєднання ознак.

Мета статті. Метою статті є систематизація та порівняльний аналіз сучасних методів виявлення ботів у соціальних мережах за типами використовуваних ознак, а також визначення їхніх переваг, обмежень і рівня пояснюваності для формування вимог до перспективної моделі виявлення автоматизованих акаунтів.

ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

Боти у соціальних мережах являють собою автоматизовані або частково автоматизовані акаунти, які імітують поведінку реальних користувачів з метою виконання певних дій: поширення інформації, взаємодії з контентом, впливу на громадську думку або маніпуляції активністю в інформаційному просторі. Залежно від функціонального призначення та особливостей поведінки виділяють кілька основних типів ботів.

Імітаційні боти – це автоматизовані акаунти, які імітують соціальну активність людини шляхом створення повідомлень, коментування, підписок та взаємодії з іншими користувачами. Такі боти часто використовуються для формування штучної активності та поширення інформації.

Спам-боти призначені для масового розповсюдження рекламного, шахрайського або небажаного контенту. Їхня діяльність зазвичай характеризується високою частотою публікацій та повторюваністю повідомлень.

Політичні боти використовуються у політичних та інформаційних кампаніях з метою маніпуляції суспільною думкою, поширення політичної пропаганди або штучного формування інформаційних трендів.

Sybil боти – множина фальшивих акаунтів, створених для отримання непропорційного впливу в соціальній мережі. Такі акаунти часто використовуються у бот-мережах.

Кіборг-акаунти – акаунти, частина дій яких виконується автоматично, тоді як інша частина контролюється реальною людиною, що значно ускладнює їх виявлення.

Для оцінювання ефективності методів виявлення ботів використовуються стандартні метрики задач класифікації. Однією з базових метрик є Accuracy, яка визначає частку правильно класифікованих об'єктів серед усіх досліджуваних акаунтів. Проте у задачах виявлення ботів використання лише Accuracy може бути недостатнім через незбалансованість датасетів.

Більш інформативними є метрики Precision та Recall. Precision характеризує частку правильно визначених ботів серед усіх акаунтів, класифікованих як боти, тоді як Recall показує здатність моделі знаходити реальних ботів серед усіх наявних бот-акаунтів.

Для комплексної оцінки якості класифікації часто використовується F1-score, що є гармонійним середнім між Precision та Recall і дозволяє оцінити баланс між точністю та повнотою класифікації.

Також важливою метрикою є ROC-AUC, яка характеризує здатність моделі відокремлювати ботів від реальних користувачів при різних порогах класифікації. Високе значення ROC-AUC свідчить про якісну дискримінаційну здатність моделі.

Окрім класичних метрик класифікації, для оцінювання методів виявлення ботів використовуються додаткові критерії, що характеризують практичну придатність моделей. Одним із таких критеріїв є пояснюваність, яка визначає можливість інтерпретації результатів роботи моделі та розуміння причин прийняття рішень щодо класифікації акаунтів. Рівні пояснюваності можна поділити на три:

- високий – ознаки явно інтерпретовані; модель дозволяє оцінити внесок ознак без додаткових методів пояснюваного ШІ (XAI);
- середній – ознаки зрозумілі, але модель складна; потрібен додатковий аналіз;
- низький – рішення базується на прихованих представленнях без наведеної інтерпретації.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Методи виявлення ботів на основі атрибутів користувача. Методи даної групи використовують ознаки, що характеризують профіль користувача та доступні без повного аналізу його публікацій або соціального графа. До них належать кількість підписників і підписок, вік акаунта, кількість публікацій, статус верифікації, наявність опису профілю, характеристики імені користувача та похідні співвідношення між зазначеними показниками. Перевагою такого підходу є відносна простота отримання даних і менша обчислювальна складність порівняно з контентними та графовими методами. Водночас його ефективність залежить від того, наскільки профільні характеристики ботів відрізняються від характеристик реальних користувачів.

У межах аналізу розглянуто модель DeeProBot [5] та моделі авторів Yang [6] та Lopez-Joya [7], які не мали окремої назви, але використовують атрибути профілю для класифікації ботів (табл. 1).

Таблиця 1

Характеристика методів, що використовують атрибути користувача

Назва методу	Алгоритм	Використані ознаки
Yang	Random Forest	20 ознак профілю користувача
DeeProBot	Гібридна DNN із LSTM та повнозв'язними шарами	11 відібраних ознак профілю, включаючи метадані та текст опису профілю
Lopez-Joya	Класичні ML-моделі з feature engineering та feature selection; в експериментах акцентовано Random Forest	Атрибути профілю, похідні соціальні характеристики та параметри конфігурації акаунта

У методі автора Yang використано двадцять ознак профілю користувача та класифікатор Random Forest. Метод DeeProBot, запропонований Hayawi et al., також використовує виключно профільну інформацію користувача, однак відрізняється способом її обробки. Модель поєднує числові й бінарні метадані профілю з текстом поля description, який подається у вигляді GloVe-векторів до LSTM-шару. До відібраних ознак належать кількість публікацій, підписників, підписок, уподобань і списків, частота публікацій, характеристики імені акаунта та опису профілю. Автори наголошують, що використання лише даних профілю дозволяє уникнути додаткових витрат на збір і обробку історії дописів користувача.

У роботі автора Lopez-Joya повний запропонований підхід використовує і профільні, і контентні ознаки. Однак автори окремо провели часткове дослідження для конфігурації, що використовує лише ознаки користувача, тому саме її результати враховано в цьому підрозділі. Аналіз показав, що найбільш значущими для класифікації є ознаки, пов'язані з кількістю підписників, співвідношенням підписників і підписок, репутацією користувача, віком акаунта та налаштуваннями профілю. За результатами експерименту атрибути користувача виявилися інформативнішими за використання лише контентних ознак на всіх трьох наборах даних, хоча поєднання обох типів ознак забезпечило найкращі результати.

Порівняння ознак демонструє, що найбільш поширеними атрибутами є кількість підписників, кількість підписок і вік акаунта. Вони використовуються майже в усіх розглянутих підходах, оскільки можуть відображати неприродне формування аудиторії або короткий період існування автоматизованих акаунтів. DeeProBot розширює базовий набір профільних характеристик текстом опису профілю й ознаками імені користувача, тоді як Lopez-Joya застосовує ширший процес конструювання та відбору профільних ознак, зокрема характеристики конфігурації акаунта (табл. 2).

Таблиця 2

Порівняльна оцінка методів на основі атрибутів користувача

Критерій	Yang	DeeProBot	Lopez-Joya: режим акаунту
Складність збору даних	Низька	Низька	Низька
Найкращий наведений результат та датасет	0,98 – Hold-out валідація	0,97 – Hold-out та Botwiki-verified	0,991 – Cresci-17
Пояснюваність	Висока/середня	Середня/низька	Висока/середня
Основна перевага	Масштабованість і простота ознак	Краща крос-платформеність за рахунок опису профілю	Розширений аналіз значущості атрибутів і краща інтерпретованість
Основний недолік	Недостатність інформації для виявлення складних ботів	Недостатність інформації для виявлення складних ботів	Недостатність інформації для виявлення складних ботів

На наборах Botwiki-verified та Midterm-18 обрані методи демонструють високі значення AUC: DeeProBot досягає 0,97 та 0,96 відповідно, а Random Forest у Yang – 0,94 та 0,96. Водночас на Cresci-rtbust перевага DeeProBot є суттєвою: AUC становить 0,72 порівняно з 0,48 для Yang. Однак навіть цей результат є значно нижчим, ніж на інших наборах даних, що пояснюється специфікою Cresci-rtbust: його розмітка базується на груповій активності та ретвіт-поведінці, які складно виявити за атрибутами окремого профілю. Для Gilani-17 обидва підходи також показують нижчу результативність, що автори DeeProBot пов'язують зі складнішими типами ботів і меншою роздільністю класів за профільними ознаками. Результати Lopez-Joya підтверджують ефективність профільних ознак на Cresci-15 і Cresci-17, де Accuracy перевищує 0,98. Однак на TwiBot-20 Accuracy лише з профільними ознаками знижується до 0,7679. Автори також показують, що для TwiBot-20 поєднання профільних і контентних ознак підвищує Accuracy до 0,8544. Це свідчить про те, що атрибути користувача є достатньо інформативними для частини наборів даних, але в умовах більшої різноманітності боти можуть потребувати доповнення іншими типами інформації.

Таким чином, аналіз розглянутих робіт показує, що методи на основі атрибутів користувача є ефективними, масштабованими та відносно доступними для практичного застосування. Найбільш інформативними ознаками є характеристики соціальних зв'язків акаунта, зокрема кількість підписників, кількість підписок, їх співвідношення, вік акаунта, а також додаткові характеристики опису і конфігурації профілю. Водночас результати на Cresci-rtbust, Gilani-17 і TwiBot-20 демонструють, що профільні ознаки не завжди достатні для виявлення ботів, діяльність яких проявляється через координацію, часові патерни або складні способи маскування.

З погляду пояснюваності найбільш придатними є підходи на основі явних профільних ознак та класичних моделей машинного навчання, як у Yang і Lopez-Joya. DeeProBot демонструє кращу результативність на окремих крос-платформених наборах, однак використання нейронної мережі, LSTM та векторних представлень тексту ускладнює інтерпретацію рішення. Отже, методи на основі атрибутів користувача доцільно розглядати як базовий і достатньо пояснюваний напрям детекції ботів, який у складніших сценаріях може бути доповнений поведінковими, текстовими або графовими ознаками.

Поведінкові методи виявлення ботів. Поведінкові методи виявлення ботів базуються на аналізі дій акаунта у соціальній мережі та часових закономірностей їх виконання. На відміну від методів на основі атрибутів користувача, вони не спираються на характеристики профілю, а намагаються виявити автоматизованість через послідовність публікацій, взаємодій і відповідей, регулярність активності, синхронність дій різних акаунтів або повторюваність поведінкових шаблонів. У межах цього підрозділу розглядаються лише методи, що використовують відкрито спостережувану активність користувачів та не потребують IP-адрес, даних пристроїв, внутрішньої мережевої телеметрії чи інших закритих платформних даних.

Аналіз наукових робіт показує, що поведінкові ознаки рідше використовуються як самостійна основа для класифікації окремого акаунта. Значна частина таких досліджень спрямована на виявлення груп скоординованих ботів за синхронністю дій або застосовує часову поведінку як додаткову групу ознак у змішаних моделях разом із профільними, текстовими чи графовими характеристиками. Тому в межах цього підрозділу основна увага приділена роботі, де часова поведінка використовується для визначення типу окремого акаунта, а також дослідженню, яке демонструє інформативність поведінкових показників, але не пропонує самостійного чистого account-level методу. Для порівняння обрано методи Bi-LSTM та Time-series pattern mining (табл. 3).

Таблиця 3

Характеристика поведінкових методів виявлення ботів

Назва методу	Алгоритм	Використані ознаки
Bi-LSTM	Bi-LSTM	Часова активність окремого акаунта; частота та інтенсивність публікацій
Time-series pattern mining	Time-series pattern mining; додаткова перевірка через SVM з RBF-ядром	Кількість активностей і видалень за дві години

У роботі Deep Temporal Analysis of Twitter Bots [8] досліджується можливість класифікації окремих акаунтів за їх часовою поведінкою. Запропонований підхід використовує двонаправлену мережу довгої короткочасної пам'яті (Bi-LSTM) для аналізу послідовних даних про активність користувача. На відміну від методів, орієнтованих на пошук синхронних груп ботів, у цій роботі об’єктом класифікації є окремий акаунт, а визначальними ознаками є частота публікацій та темп активності.

У роботі Temporal Patterns in Bot Activities [9] аналізуються часові ряди активності ботів, що містять інформацію про виконання користувачем таких дій, як публікація, репост і видалення повідомлення. Автори досліджують повторювані фрагменти поведінки, аномальні послідовності, періодичність, сплески активності та динамічні кластери. Проте вихідною точкою дослідження є боти, уже знайдені системою DeBot [10], яка виявляє групи акаунтів із незвично синхронною активністю. Отже, ця робота демонструє інформативність часових показників для характеристики окремого акаунта, але не пропонує незалежного методу його первинної класифікації як бота. Порівняльна оцінка поведінкових підходів представлена в табл. 4.

Таблиця 4

Порівняльна оцінка поведінкових підходів

Критерій	Bi-LSTM	Time-series pattern mining
Об’єкт оцінювання	Окремий акаунт	Акаунти, попередньо позначені DeBot як боти
Найкращий наведений авторами результат	0,988	0,8171
Використання неповедінкових ознак	Не зазначено	Так
Пояснюваність	Середня/низька через Bi-LSTM	Висока для опису часових патернів; обмежена як метод виявлення
Основна перевага	Класифікація окремого акаунта лише за часовою поведінкою	Наочне пояснення характерних патернів бот-активності
Основний недолік	Низька інтерпретованість результатів та обмежені відомості про узагальнюваність	Не є самостійним чистим методом

Результат Bi-LSTM свідчить, що часова поведінка окремого акаунта може бути достатньо інформативною для задачі класифікації ботів. Високе значення Точності показує потенціал рекурентних нейронних мереж для аналізу частоти та темпу публікацій. У дослідженні Time-series pattern mining SVM-класифікатор із RBF-ядром досяг середньої Accuracy 0,8171 при десятикратній крос-валідації на збалансованому наборі ботів і добросовісних користувачів. Проте автори прямо наголошують, що цей експеримент не є новим методом виявлення ботів, оскільки боти були попередньо визначені системою DeBot. Крім того, дві з чотирьох використаних ознак (частка повідомлень із URL та частка дублікатів повідомлень) виходять за межі чисто поведінкового аналізу. Тому результат роботи доцільно використовувати лише для підтвердження того, що часові показники, такі як загальна активність і частота видалень, можуть бути корисними додатковими індикаторами автоматизації.

Перевагою підходу Bi-LSTM є відповідність задачі індивідуального виявлення ботів: метод не потребує побудови графа взаємодій, аналізу змісту повідомлень або доступу до закритих технічних даних. Він може бути особливо корисним для акаунтів, поведінка яких характеризується неприродно високою частотою публікацій або стабільними автоматизованими часовими патернами. Недоліком підходу є обмежена пояснюваність Bi-LSTM: навіть якщо модель правильно визначає бот-акаунт, встановити, який саме фрагмент часової послідовності став визначальним для рішення, складніше, ніж у випадку статистичних правил або дерев рішень.

Проведений аналіз демонструє, що поведінка окремого акаунта може бути використана для виявлення ботів, зокрема через аналіз частоти та часової структури публікацій. Найбільш релевантним прикладом є метод Bi-LSTM, який застосовується для індивідуальної класифікації акаунтів за часовою активністю.

Водночас поведінкові ознаки значно частіше використовуються у двох інших сценаріях. По-перше, вони застосовуються для виявлення груп ботів, діяльність яких проявляється через синхронні публікації, координовані взаємодії або подібні часові шаблони. До цього напряму належать DeBot [10], Social Fingerprinting [11] і RTbust [12]. По-друге, поведінкові характеристики часто включаються як додаткові ознаки до профільних, текстових або графових даних у змішаних моделях, оскільки вони доповнюють статичний опис акаунта інформацією про характер його діяльності.

Отже, суто поведінкові методи індивідуального виявлення ботів є перспективними, але недостатньо широко представленими напрямом. Їх основною перевагою є можливість виявляти автоматизацію незалежно від змісту повідомлень та оформлення профілю. Основними обмеженнями залишаються потреба у достатній історії активності, складність виявлення ботів, які навмисно імітують природний темп поведінки людини, а також нижча пояснюваність моделей глибокого навчання.

Методи виявлення ботів на основі текстового контенту. Методи на основі текстового контенту використовують повідомлення, створені акаунтом, для визначення ознак автоматизованої діяльності. Такі підходи можуть спиратися на повторюваність лексики, особливості використання хештегів і посилань, тональність повідомлень, семантичні представлення слів або контекстні векторні подання, сформовані нейронними мовними моделями. На відміну від методів на основі атрибутів користувача, вони аналізують не оформлення профілю, а зміст опублікованого тексту. Це дозволяє виявляти ботів, які поширюють однотипні рекламні, політичні або маніпулятивні повідомлення, однак робить методи залежними від мови, тематики та здатності сучасних ботів генерувати природний текст (табл. 5).

Таблиця 5

Характеристика методів виявлення ботів на основі текстового контенту

Запропонований метод	Використані текстові дані та ознаки	Завдання класифікації
BiLSTM з вставками	Текст дописів; контекстні послідовності слів; GloVe-векторизація; спеціальні Twitter-токени для hashtag, mention та URL	Бінарна: human / spambot
Удосконалений DistilBERT	Текстові дані акаунтів; токенізація та контекстне представлення тексту трансформерами	Багатокласова: human / spam / fake / political bot / Sybil

У методі BiLSTM with Word Embeddings [13] запропоновано визначати, чи є окремий Twitter-акаунт ботом, використовуючи лише зміст його повідомлень. Метод представляє кожне слово через попередньо навчені вектори GloVe, після чого послідовності слів обробляються тришаровою мережею BiLSTM. Автори окремо наголошують, що модель не використовує профільні ознаки, граф дружніх зв'язків або часову історію поведінки акаунта та не потребує ручного конструювання ознак. Специфічні для соціальної мережі Twitter елементи (хештеги, згадування користувачів і скорочені URL) перетворюються на окремі текстові токени, тобто аналізуються як частина контенту повідомлень. Такий підхід відмовляється від вручну сформованих статистичних ознак, таких як кількість URL, частота окремих слів або чисельна оцінка тональності. Замість цього текст повідомлень перетворюється у послідовність GloVe-векторів, а BiLSTM навчається виявляти семантичні та послідовнісні закономірності, характерні для спам-ботів. Така архітектура є простішою за сучасні трансформерні моделі та не потребує великої кількості підготовлених ознак, однак залежить від здатності статичних embedding-представлень відобразити значущі відмінності між ботами й людьми.

У роботі Ellaky [14] запропоновано текстовий підхід до багатокласового розпізнавання соціальних ботів. Автори сформували об'єднаний текстовий набір даних, що містить повідомлення людей, спам-ботів, політичних ботів, фальшивих акаунтів і Sybil-акаунтів, та порівняли 6 трансформерних моделей: BERT, DistilBERT, RoBERTa, DeBERTa, XLNet і ALBERT. На відміну від BiLSTM-підходу, де семантичні представлення слів є статичними, трансформерні моделі формують контекстне представлення токенів залежно від їх оточення у тексті. Найкращий результат у роботі отримано для DistilBERT. Перевагою такого представлення є можливість ураховувати контекст слова у конкретному повідомленні, що є важливим для аналізу політичних, шахрайських або маніпулятивних повідомлень. Крім того, робота переходить від бінарної постановки задачі до визначення типу автоматизованого акаунта. Водночас підготовлений авторами набір даних є об'єднанням кількох джерел, тому модель

потенційно може частково навчатися не лише мовних відмінностей між типами ботів, а й відмінностей у способах формування вихідних датасетів. Нижче в табл. 6 представлені результати текстових методів за наборами даних.

Таблиця 6

Результати текстових методів за наборами даних

Метод	Набір даних	Класи	Precision	Recall	Accuracy	F1-score
BiLSTM з вставками	Cresci-2017, Test set № 1: genuine accounts + social-bot-1	Human/bot	0,940	0,976	0,961	0,963
	Cresci-2017, Test set № 2: genuine accounts + social-bot-3	Human/bot	0,933	0,919	0,929	0,926
Удосконалений DistilBERT	Об'єднаний текстовий набір: Cresci-2015, Cresci-2017, TwiBot-20, Fake Followers, Political Bots, Human Accounts	Human/spam/fake/political bot/Sybil	0,9685	0,9683	0,9683	0,9684

Метод BiLSTM with Word Embeddings показав найкращий результат для Test set № 1, що містить реальні акаунти та ботів, які здійснюють ретвіти політичного контенту: Accuracy становить 0,961, а F1-score – 0,963. Для Test set № 2, що містить спам-ботів, які рекламують товари, показники дещо нижчі: Accuracy дорівнює 0,929, а F1-score – 0,926. Автори також наводять приклади лексичних відмінностей: повідомлення спам-ботів частіше містять рекламно забарвлені слова та посилання на зовнішні вебсторінки, тоді як у повідомленнях людей частіше зустрічається повсякденна лексика. Це підтверджує інформативність тексту для виявлення окремих типів ботів.

У роботі Advanced Text-Based Transformer Architecture for Malicious Social Bots Detection найкращу результативність продемонструвала модель DistilBERT, яка досягла Accuracy 96,83%, Precision 96,85%, Recall 96,83% та F1-score 96,84%. Автори також подають результати за окремими класами: F1-score становить 99% для спам-ботів, 92% для фальшивих акаунтів, 89% для політичних ботів, 98% для людей і 100% для Sybil-акаунтів. Нижчий результат для політичних ботів свідчить про складність їх розмежування за текстом, оскільки політичні повідомлення ботів можуть бути стилістично та тематично подібними до дописів реальних користувачів. Порівняльна оцінка текстових методів представлена в табл. 7.

Таблиця 7

Порівняльна оцінка текстових методів

Критерій	BiLSTM з вставками	Удосконалений DistilBERT
Основний об'єкт аналізу	Послідовності слів у дописах	Контекстне представлення тексту дописів
Підтримка багатокласового визначення типу бота	Не досліджено	Так
Здатність аналізувати контекст слова	Середня/висока	Висока
Мовна залежність	Висока; залежить від embedding-корпусу	Висока; залежить від навчальних даних
Пояснюваність	Низька/середня	Низька
Основна перевага	Висока якість без профільних і графових ознак	Висока якість і розпізнавання кількох типів ботів
Основний недолік	Обмежена перевірка на різних типах ботів і датасетах	Висока залежність від датасету й складність інтерпретації рішення

Перевагою текстових методів є можливість визначати ботів за змістом створюваних повідомлень навіть у випадках, коли профіль акаунта виглядає правдоподібно, а його часові патерни не мають очевидної автоматизованості. Застосування вставок та трансформерів дозволяє відмовитися від значної кількості вручну розроблених характеристик і безпосередньо навчати модель на текстовому наборі.

Основним обмеженням текстових методів є залежність від тематики та мови повідомлень. Модель, навчена на рекламних спам-ботах або політичних ботах певної кампанії, може гірше працювати для ботів іншого типу або для повідомлень іншою мовою. Крім того, сучасні генеративні мовні моделі дозволяють ботам створювати більш різноманітні та природні тексти, що зменшує вираженість лексичних і стилістичних відмінностей між автоматизованими та людськими акаунтами.

З погляду пояснюваності, обидва розглянуті підходи мають суттєві обмеження. У випадку BiLSTM можна встановити загальні лексичні закономірності, наприклад переважання рекламних слів у спам-ботів, однак конкретне рішення нейронної мережі важко пов'язати з визначеним набором прозорих правил. Для DistilBERT проблема є ще виразнішою: модель формує складні контекстні подання тексту, а без додаткових XAI-інструментів, наприклад візуалізації уваги, SHAP або LIME, неможливо однозначно пояснити, які фрагменти тексту визначили результат класифікації.

Графові методи виявлення ботів та проблеми формування датасетів. Графові методи розглядають соціальну мережу як структуру, у якій користувачі є вузлами, а зв'язки між ними – підписками, згадуваннями, відповідями, ретвітами або іншими взаємодіями – ребрами графа. На основі такої структури можуть аналізуватися центральність вузлів, належність до спільнот, щільність взаємодій, характер зв'язків між ботами й реальними користувачами, а також можуть використовуватися графові нейронні мережі (GNN). Такі методи є перспективними для виявлення бот-мереж і координованих кампаній, оскільки враховують не лише властивості окремого акаунта, а й його місце у системі взаємодій.

Доцільність використання графових характеристик підтверджується результатами дослідження «A global comparison of social media bot and human characteristics» [2]. Робота не пропонує нового графового детектора: користувачі у вихідному датасеті були класифіковані як боти або люди алгоритмом BotHunter, після чого автори виконали порівняльний аналіз їх текстових, профільних і мережевих характеристик. Дослідження охоплює приблизно 200 млн користувачів і близько 5 млрд повідомлень із семи подій, зокрема виборів, пандемічних дискусій і суспільно-політичних кампаній.

Для аналізу мережевої структури автори побудували его-мережі користувачів, у яких ребрами вважалися комунікаційні взаємодії: репости, згадування, відповіді та цитування. Порівнювалися такі графові показники, як вхідний, вихідний і загальний ступінь вузлів, а також щільність мережі. Встановлено, що его-мережі ботів були щільнішими за мережі людей на 8,33%. Крім того, боти взаємодіяли з більшою часткою інших ботів: у середньому 9,66% їхніх зв'язків припадало на бот-акаунти проти 7,31% для людей. Водночас переважна частина взаємодій ботів усе одно була спрямована на людей, що може свідчити про використання автоматизованих акаунтів для впливу на людську аудиторію.

Важливим результатом роботи є виявлена відмінність у формі комунікаційних структур. На прикладі найбільш активних комунікаторів під час азійських виборів автори показали, що для ботів характернішою є зіркоподібна структура взаємодій, у якій центральні акаунти створюють або підтримують повідомлення, а периферійні боти підсилюють їх через репости. Для людей типовішою була деревоподібна або ієрархічна структура, що відображає поступове поширення інформації між безпосередніми та опосередкованими контактами. Такий результат підтверджує, що графові ознаки можуть виявляти координацію та механізми підсилення повідомлень, які складно встановити лише за профілем або текстом окремого користувача.

Однак результати цієї роботи одночасно демонструють головне практичне обмеження графових методів – потребу у великих і достатньо повних наборах даних про взаємодії. Для побудови надійного графа необхідно отримати не лише дані окремого акаунта, а й репости, згадування, відповіді, цитування та зв'язки значної кількості пов'язаних користувачів. Якщо такі дані збираються частково, мережа може втратити важливі ребра, унаслідок чого спотворюються показники щільності, центральності та структури спільнот.

Повторне формування їхнього масивного корпусу за сучасних обмежень X API було б практично недоступним: за оцінкою авторів, у безкоштовному режимі збір такого обсягу даних потребував би приблизно 136 986 років, а в режимі Pro – близько 416 років і 25 млн доларів США. Це підтверджує, що графові методи можуть мати високу аналітичну цінність, однак їх відтворюваність і практичне використання без уже наявного якісного датасету істотно обмежені доступністю платформних даних.

Проблеми виникають і на етапі формування датасетів. По-перше, соціальний граф є динамічним: користувачі підписуються та відписуються, видаляють повідомлення, змінюють активність або припиняють існування. Унаслідок цього граф, зібраний у різні моменти часу, може відображати різні стани мережі. По-друге, складною є процедура розмітки акаунтів як ботів, так і людей, особливо для складних чи частково автоматизованих акаунтів. По-третє, набори даних можуть бути незбалансованими або орієнтованими лише на окремі типи ботів, через що модель погано узагальнюється на інші сценарії.

Таким чином, графові методи мають значний потенціал для виявлення скоординованих бот-мереж, однак їх практичне застосування істотно залежить від наявності великих, актуальних і якісно розмічених графових датасетів. Через API-ліміти, вартість збору даних, динамічність соціальних зв'язків і складність розмітки отримання таких наборів є значно складнішим, ніж для методів, що використовують лише профільні або текстові ознаки. Саме тому в межах даного огляду графові підходи розглядаються як перспективний, але ресурсомісткий напрям, детальне експериментальне порівняння якого потребує окремо сформованої якісної графової вибірки.

Гібридні методи виявлення ботів. Змішані або гібридні методи поєднують кілька джерел інформації про акаунт: атрибути профілю, поведінкові характеристики, текстовий контент, сентиментні оцінки та, у найбільш комплексних системах, графові зв'язки. На відміну від однотипних підходів, такі методи дозволяють виявляти ботів, які маскують окрему групу ознак. Наприклад, бот може мати правдоподібно оформлений профіль, але демонструвати неприродний темп активності; або генерувати природний текст, але бути включеним до підозрілої структури взаємодій. Тому комбінування ознак розглядається як спосіб підвищення стійкості детекції до складніших і адаптивних ботів.

У межах аналізу розглянуто п'ять робіт, у яких запропоновані методи використовують щонайменше дві категорії ознак. Розглянуті підходи відрізняються рівнем складності: від поєднання метаданих із сентиментними характеристиками до багатоджерельної інтеграції метаданих, текстових представлень і графової структури (табл. 8).

Таблиця 8

Характеристика гібридних методів

Запропонована модель	Комбіновані типи ознак	Задача
Sentiment-enhanced feature pipeline; найкращий результат – AdaBoost із SMOTE	Метадані профілю та активності та сентиментні характеристики дописів	Бінарна: human / bot
SVM; додатково оцінено k-means	Часові та семантичні ознаки історії дописів	Бінарна: organic / inorganic
Stacking Ensemble: RF, LightGBM, XGBoost, SVM; метамодель - Logistic Regression	Статичні атрибути профілю та часові поведінкові ознаки	Бінарна: human / bot
Transformer-Based Multi-modal Feature Fusion	Текст, поведінка, граф та сентимент	Бінарна: human / bot
CB-MTE: DistilBERT, DeepWalk/centrality, UMAP, CatBoost	Метадані, текст, графові ознаки	Бінарна: human / bot

У моделі Sentiment-enhanced feature pipeline [15] запропоновано доповнити набір початкових характеристик акаунта сентиментними ознаками, отриманими з його дописів. Авторами використано 14 початкових ознак: кількість підписників, підписок, повідомлень, вподобань і списків, ознаки стандартного оформлення профілю, Geo_Enabled, а також Retweet_Count, Reply_Count, Num_Hashtags і Num_Urls. До них додано сім нових характеристик: кількість позитивних, нейтральних і негативних повідомлень, середній сентимент, його дисперсію, стандартне відхилення та середній час публікації. Сентимент обчислюється моделлю Bi-LSTM з Attention, після чого сформовані ознаки передаються до класичних класифікаторів. Найкращий F1-score для комбінованих ознак зі SMOTE отримав AdaBoost - 0,9957, тоді як найвищу Precision продемонстрував Random Forest – 0,9979.

У моделі SVM [16] поєднуються часові та семантичні ознаки, вилучені з історії повідомлень окремого користувача. До часових характеристик належать періодичність активності, ARIMA-ознаки та локальні максимуми періодограми. Семантичний блок включає лексичну різноманітність, середню кількість слів у повідомленні та її дисперсію, частоту хештегів, частоту найбільш уживаних слів і сентимент. Загалом автори формують 19 характеристик, із яких після відбору за Variance Inflation Factor залишають 11. Найкращий результат отримано для SVM: specificity 0,986, recall 0,979 і F1-score 0,984.

Метод Stacking Ensemble [17] комбінує статичні атрибути профілю і часові ознаки активності. Статичні характеристики включають кількість підписників, кількість підписок, вік акаунта, статус верифікації та ознаку геолокації; часові: частоту публікацій, середній рівень залучення, розподіл активності за часом, частоту ретвітів і відповідей. Для класифікації автори застосували stacking ensemble із Random Forest, LightGBM, XGBoost та SVM на першому рівні та Logistic Regression як метамоделлю. Запропонований ensemble досяг Accuracy 0,9871 та F1-score 0,9915, перевищивши всі окремі базові класифікатори в межах того самого набору ознак.

У моделі Transformer-Based Multi-modal Feature Fusion [18] запропоновано мультимодальний підхід на основі мереж перетворення, який включає текстові, поведінкові, графові та сентиментні представлення. Автори описують вилучення емоційних характеристик за допомогою моделей на кшталт VADER або TextBlob, кодування окремих модальностей спеціалізованими енкодерами та подальшу агрегацію ознак у спільний вектор для бінарної класифікації. Метод продемонстрував Accuracy від 92,7% до 95,4% на чотирьох використаних датасетах; найвищий показник отримано на Cresci-2017.

Найбільш комплексним серед розглянутих є метод CB-MTE [19]. Він формує 32-вимірний вектор метаданих, зокрема атрибути акаунта, поведінкові та соціальні показники; текст дописів кодується за допомогою DistilBERT; графова структура подається через DeepWalk і показники центральності. Після зменшення розмірності текстових і графових представлень за допомогою UMAP усі модальності об'єднуються у 67-вимірний вектор, який класифікується алгоритмом CatBoost. На п'яти збалансованих підвибірках TwiBot-22 середній F1-score методу становить 0,8084.

Розглянуті роботи демонструють поступове ускладнення гібридних підходів. У Sentiment-enhanced feature pipeline базові метадані акаунта доповнюються сентиментними показниками його повідомлень. У SVM поєднання відбувається на рівні самої історії дописів: часові характеристики доповнюються семантичними. У Stacking Ensemble статична інформація профілю об'єднується з часовою динамікою активності, що дозволяє виявляти акаунти, які виглядають правдоподібно на рівні профілю, але демонструють неприродну поведінку.

Методи Transformer-Based Multi-modal Feature Fusion і CB-MTE представляють ширшу мультимодальну групу, оскільки разом із текстом та поведінкою включають графову інформацію. У теоретичному сенсі це забезпечує найповніше представлення акаунта: модель аналізує, хто є користувач, що він публікує, як поводиться і з ким взаємодіє. Водночас саме такі підходи є найбільш залежними від наявності якісних графових наборів даних, проблема отримання яких була розглянута в попередньому підрозділі. Результати аналізу гібридних методів представлені в табл. 9.

Таблиця 9

Результати гібридних методів

Метод	Набір даних	Accuracy	Precision	Recall	F1-score
AdaBoost та SMOTE	Cresci-2017	0,9927	0,9979	0,9936	0,9957
SVM	Bot Repository: organic / inorganic users	–	–	0,9790	0,9840
Stacking Ensemble	Cresci-2017	0,9871	0,9926	0,9904	0,9915
Transformer-Based Multi-modal Feature Fusion	TweepFake	0,9480	0,9230	0,9370	–
	Botometer Dataset	0,9320	0,9010	0,9150	–
	Cresci-2017	0,9540	0,9400	0,9390	–
	PAN-2010	0,9270	0,9120	0,9050	–
CB-MTE	TwiBot-22, середнє для 5 підвибірок	0,8214	0,7924	0,8254	0,8084

Результати Sentiment-enhanced feature pipeline свідчать, що додавання сентиментних ознак до початкових характеристик акаунта може покращувати виявлення ботів на Cresci-2017. За комбінованими ознаками зі SMOTE AdaBoost досягає F1-score 0,9957. Для AdaBoost F1-score зростає з 0,9954 на початкових ознаках до 0,9957 після додавання сентименту; для Logistic Regression підвищення становить із 0,9815 до 0,9848, а для DNN – із 0,9728 до 0,9900. Водночас приріст для вже сильних деревоподібних моделей є невеликим, що свідчить не стільки про вирішальний внесок сентименту, скільки про його роль як додаткового джерела інформації.

Запропонований stacking ensemble перевищує окремі класифікатори, навчені на тому самому інтегрованому наборі статичних і часових ознак: F1-score ensemble становить 0,9915 проти 0,9855 для LightGBM, 0,9843 для XGBoost, 0,9832 для Random Forest і 0,9783 для SVM. Це підтверджує перевагу ансамблю моделей, однак не дозволяє окремо встановити, наскільки саме додавання часових ознак покращило результат порівняно з профільними ознаками без них.

Метод SVM отримав F1-score 0,984 для часових і семантичних ознак. Його перевагою є використання інтерпретованих характеристик: наприклад, автори визначають, що для SVM найбільш значущими є хештеги та кількість слів, тоді як для k-means – відносна частота слів і дисперсія довжини повідомлень. Тому цей підхід є менш складним, але більш пояснюваним порівняно з мультимодальними нейронними архітектурами.

Результати Transformer-Based Multi-modal Feature Fusion демонструють відносно стабільну якість на чотирьох різних наборах даних: Accuracy коливається від 0,927 для PAN-2010 до 0,954 для Cresci-2017. Це є перевагою порівняно з роботами, перевіреними лише на одному датасеті. Проте через стислий

опис формування ознак, валідації та складу датасетів отримані показники доцільно використовувати як ілюстрацію потенціалу мультимодальної інтеграції, а не як підставу для остаточного ранжування моделей.

Метод CB-MTE демонструє нижчі абсолютні значення F1-score, ніж у моделей, перевірених на Cresci-2017, однак оцінюється на складнішому TwiBot-22 із використанням текстових, профільних і графових даних. Найважливішим результатом цієї роботи для огляду є не лише підсумкова метрика, а й дослідження впливу компонентів: на підвибірці TwiBot_3 повна конфігурація досягає F1-score 0,847, що на 7,5 відсоткового пункту більше за варіант лише з метаданими. Додавання тексту до метаданих підвищує F1-score на 2,9 відсоткового пункту, а додавання графових ознак – на 3,9 відсоткового пункту. Отже, у цій роботі перевага поєднання модальностей підтверджена в межах контрольованого експерименту на одному наборі даних. Порівняльна оцінка гібридних методів представлена в табл. 10.

Таблиця 10

Порівняльна оцінка гібридних методів

Критерій	AdaBoost та SMOTE	SVM	Stacking Ensemble	TBMFF	CB-MTE
Кількість типів ознак	3	2	2	4	3
Складність отримання даних	Середня	Середня	Середня	Висока	Висока
Обчислювальна складність	Середня	Середня	Середня	Висока	Висока
Пояснюваність	Середня	Висока/середня	Середня	Низька	Середня/низька
Перевірка внеску різних модальностей	Часткова	Через важливість ознак	Не деталізовано	Не деталізовано	Повне дослідження впливу компонентів
Основна перевага	Просте доповнення метаданих	Інтерпретовані часові та семантичні ознаки	Висока точність і сильний ансамбль	Оцінювання на кількох датасетах	Інтеграція ознак і контрольована перевірка їх внеску
Основний недолік	Невеликий приріст від сентимент аналізу	Обмежений обсяг і склад датасету	Перевірка лише на одному датасеті	Недостатня деталізація відтворюваності	Високі вимоги до графових даних і ресурсів

Пояснюваність змішаних методів зменшується зі збільшенням кількості модальностей і складності моделі. Підхід SVM є найбільш інтерпретованим серед розглянутих, оскільки використовує явно визначені часові та семантичні ознаки й проводить оцінку їх важливості. У AdaBoost та SMOTE також можна пояснити роль сформованих сентиментних характеристик, хоча модуль Bi-LSTM з увагою, який їх обчислює, є менш прозорим.

Підхід Stacking Ensemble має середній рівень пояснюваності: використані ознаки є зрозумілими, а автори аналізують помилки класифікації. Вони зазначають, що хибні позитивні результати часто стосувалися активних реальних акаунтів, наприклад новинних сторінок, а хибні негативні – ботів із нерегулярним графіком публікацій і різноманітним контентом. Водночас архітектура накладання моделей ускладнює встановлення точного внеску кожної ознаки у фінальне рішення.

Найнижчу пояснюваність мають підходи, що поєднують трансформерні представлення з графовими вставками. У CB-MTE текст кодується через DistilBERT, граф – через DeepWalk і показники центральності, а фінальне рішення приймається CatBoost після зменшення розмірності. Така архітектура добре охоплює різні прояви бот-активності, однак інтерпретація окремого рішення потребувала б додаткових засобів. Важливою перевагою цієї роботи є те, що автори окремо проаналізували стійкість модулів до шуму: метадані виявилися найбільш стабільними, графові ознаки посіли друге місце, а текстовий модуль був найбільш чутливим до збурень.

Узагальнення результатів аналізу та проблеми сучасних методів. Проведений аналіз показує, що методи виявлення ботів у соціальних мережах істотно відрізняються за використовуваними даними, обчислювальною складністю, пояснюваністю та придатністю до виявлення різних типів ботів. Методи на основі атрибутів користувача є відносно простими й доступними для реалізації, поведінкові підходи

дозволяють виявляти неприродну активність і координацію, текстові методи аналізують зміст поширюваної інформації, а графові – структуру взаємодій між акаунтами. Змішані методи поєднують переваги декількох груп ознак і загалом є найбільш перспективними для виявлення складних ботів, однак потребують більших обсягів якісних даних і мають нижчу пояснюваність. Узагальнене порівняння проаналізованих груп методів представлено в табл. 11.

Таблиця 11

Узагальнене порівняння проаналізованих груп методів

Тип методів	Основне джерело інформації	Основна перевага	Основний недолік	Пояснюваність
На основі атрибутів користувача	Метадані профілю, кількість підписників, підписок, вік акаунта, опис профілю	Простота отримання даних і можливість аналізу окремого акаунта	Боти можуть маскувати профільні характеристики	Висока або середня
Поведінкові	Частота публікацій, часові шаблони, відповіді, регулярність активності	Виявлення автоматизованої або координованої активності	Потреба в історії активності; часто орієнтовані на групи ботів	Середня
Текстові	Зміст дописів, вставки, сентимент, трансформерні представлення	Виявлення ботів за характером поширюваного контенту	Мовна й тематична залежність	Низька або середня
Графові	Структура підписок, згадувань, відповідей і ретвітів	Виявлення бот-мереж та координованих кампаній	Складність отримання повного графа і висока ресурсомісткість	Середня або низька
Змішані/ гібридні	Комбінація профільних, поведінкових, текстових і графових ознак	Більш повне представлення акаунта та вища стійкість до маскування	Висока складність, вимоги до даних і проблеми інтерпретації	Середня або низька

Методи на основі атрибутів користувача показують, що навіть відносно доступні дані профілю можуть бути достатньо інформативними для виявлення частини ботів. Наприклад, DeeProBot використовує лише інформацію профілю, включаючи числові характеристики та текст опису акаунта, і досягає AUC до 0,97. Водночас його результативність помітно зменшується на наборах даних, де ботів складно виявити без урахування групової або ретвіт-поведінки.

Поведінкові методи є корисними для аналізу автоматизованої активності, оскільки дозволяють враховувати частоту, регулярність і часову структуру дій акаунта. Однак у наукових роботах такі ознаки часто застосовуються не як самостійний засіб класифікації окремого користувача, а для виявлення координованих груп ботів або як доповнення до інших характеристик. Це обмежує кількість чистих поведінкових підходів, придатних для прямого порівняння у задачі bot/human.

Текстові методи демонструють високу ефективність у ситуаціях, коли боти поширюють повторюваний, рекламний або політично спрямований контент. Трансформерний підхід заснований на DistilBERT досягає Accuracy 96,83% у багатокласовій задачі розпізнавання людей, спам-ботів, політичних ботів, фальшивих і Sybil-акаунтів. Водночас текстові моделі є залежними від мови, тематики та якості навчального корпусу, а розвиток генеративного штучного інтелекту ускладнює відокремлення ботів від людей лише за текстом повідомлень.

Змішані методи демонструють найбільший потенціал для подальших досліджень, оскільки дозволяють компенсувати недоліки окремих груп ознак. Наприклад, метод Stacking Ensemble поєднує статичні атрибути профілю з часовими характеристиками активності й досягає Accuracy 98,71% на Cresci-2017. Водночас автори зазначають, що помилки виникають для активних реальних акаунтів, схожих на ботів за поведінкою, та для ботів, які імітують нерегулярні людські патерни.

На основі виконаного аналізу проблеми сучасних методів виявлення ботів доцільно згрупувати у чотири основні категорії: проблеми ознак і адаптивності ботів; проблеми даних та оцінювання; проблеми практичного впровадження; проблеми пояснюваності (табл. 12).

Таблиця 12

Основні проблеми сучасних методів виявлення ботів

Група проблем	Прояв проблеми	Наслідок для методів
Адаптивність ботів і недостатність окремих ознак	Боти імітують реальні профілі, нерегулярну активність і природний текст	Одномодальні моделі втрачають ефективність для складних ботів
Неоднорідність і застарівання датасетів	Набори даних містять різні типи ботів, різні способи розмітки та різні періоди збору	Результати робіт складно безпосередньо порівнювати; моделі погано узагальнюються
Обмежений доступ до даних і масштабованість	Графові та поведінкові методи потребують великих обсягів історичних і мережевих даних	Ускладнюється формування нових повних наборів даних і застосування моделей у реальному часі
Низька пояснюваність складних моделей	Transformer-, LSTM-, GNN- і мультимодальні моделі працюють із латентними представленнями	Важко обґрунтувати, чому акаунт визначено як бот
Дисбаланс класів і неоднозначність розмітки	Людські акаунти можуть переважати; cyborg-акаунти складно однозначно класифікувати	Accuracy може бути оманливо високою, а Recall для ботів – недостатнім

Адаптивність ботів і недостатність окремих ознак виникають через те, що кожна окрема група ознак описує лише частину поведінки акаунта. Профільні методи можуть бути обійдені шляхом заповнення опису, налаштування фотографії та формування правдоподібного співвідношення підписників і підписок. Поведінкові методи стають менш ефективними, якщо бот уникає надмірно регулярної активності. Текстові методи ускладнюються через можливість автоматичного створення різноманітних і природних повідомлень. Графові методи, своєю чергою, можуть не виявити ботів, які ще не сформували вираженої мережі взаємодій. Ця проблема підтверджується аналізом помилок у роботі про модель Stacking Ensemble: хибно позитивними прикладами часто ставали активні легітимні акаунти, зокрема новинні сторінки, тоді як хибно негативними – боти з нерегулярним графіком активності та різноманітним контентом.

Однією з найбільш суттєвих проблем є відсутність єдиного універсального набору даних, який достатньо повно представляв би сучасні типи ботів. Датасети відрізняються за платформою, періодом збору, типами ботів, способом формування міток і доступними ознаками. Через це високий результат моделі на одному датасеті не гарантує аналогічної ефективності на іншому. Наприклад, моделі, що демонструють дуже високі значення метрик на Cresci-2017, можуть гірше працювати на TwiBot-22 або на наборах, побудованих за поведінкою взаємодій. У дослідженні на TwiBot-22 також наголошується, що високий показник Accuracy за незбалансованих класів може приховувати низьку ефективність саме у виявленні ботів; тому для оглядового порівняння важливішими є F1-score, Precision і Recall [20].

Іншою проблемою є отримання даних. Методи на основі профільних атрибутів потребують порівняно невеликого обсягу даних, однак поведінкові, графові та гібридні підходи вимагають збору історії дописів, відповідей, згадувань і соціальних зв’язків. Це ускладнює створення актуальних датасетів та обмежує можливість відтворення попередніх досліджень.

Крім того, складні мультимодальні архітектури важче застосовувати у режимі реального часу. Transformer- і GNN-моделі потребують більших ресурсів, ніж класичні алгоритми на табличних ознаках. У роботі CB-MTE саме зменшена модель DistilBERT використовується як компроміс між якістю семантичного аналізу та швидкістю обробки великих потоків повідомлень. Пояснюваність є важливою вимогою до систем виявлення ботів, оскільки помилкова класифікація реального користувача може призвести до необґрунтованого обмеження його діяльності. Найбільш пояснюваними є моделі, що використовують явні ознаки й класичні алгоритми, наприклад Random Forest або SVM: можна оцінити важливість віку акаунта, частоти активності, кількості підписників чи лексичної різноманітності. Натомість нейронні й мультимодальні моделі працюють із вставками та прихованими представленнями, які важко інтерпретувати без додаткових засобів. У роботі Stacking Ensemble серед рекомендованих напрямів розвитку прямо зазначено інтеграцію XAI-методів, таких як SHAP або LIME, для забезпечення прозорості прийняття рішень.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Виявлені обмеження дозволяють визначити напрями досліджень, які є найбільш актуальними для подальшого розвитку методів виявлення ботів (табл. 13).

Таблиця 13

Напрями подальших досліджень

Напрям	Проблема, яку він вирішує	Очікуваний результат
Комбінування доступних типів ознак	Недостатність лише профільних, текстових або поведінкових даних	Підвищення стійкості до маскування ботів
Міждомений датасет	Залежність моделей від конкретного датасету	Оцінювання реальної узагальнювальної здатності моделей
Оновлювані та різноманітні датасети	Еволюція ботів і застарівання вибірок	Виявлення нових типів ботів та акаунтів керованих ШІ
Пояснюваність	Непрозорість складних моделей	Можливість обґрунтувати рішення детектора
Ефективні моделі без критичної залежності від повного графа	Недоступність великих графових даних	Практична реалізація за обмеженого доступу до API
Реальний час і зниження обчислювальних витрат	Висока складність мультимодальних моделей	Можливість оперативного застосування на великих потоках даних

Найбільш обґрунтованим напрямом є розроблення та оцінювання методів, які поєднують декілька доступних типів ознак, наприклад, атрибути користувача, поведінкові характеристики й текстовий контент. Такий підхід дозволяє врахувати різні прояви автоматизованості, але не робить модель критично залежною від повного соціального графа, отримання якого є дорогим і складним.

Доцільним також є проведення експериментів не лише на одному датасеті, а на декількох вибірках із різними типами ботів. DeeProBot демонструє важливість кросс-доменного тестування, оскільки результати моделі істотно різняться залежно від характеру тестового набору даних. Дослідження на TwiBot-22 додатково показує, що в умовах складнішої та незбалансованої вибірки необхідно оцінювати не лише Accuracy, а й Precision, Recall і F1-score.

Окремої уваги потребує пояснюваність. Оскільки найбільш перспективні гібридні моделі часто поєднують глибокі текстові представлення, часові характеристики та ансамблеві алгоритми, важливо доповнювати їх засобами інтерпретації результатів. Це дозволить визначати причини класифікації акаунта як бота, виявляти помилки моделі та знижувати ризик необґрунтованого блокування реальних користувачів.

Таким чином, жодна окрема група методів не забезпечує одночасно високої точності, стійкості до сучасних ботів, доступності даних, низької обчислювальної складності та достатньої пояснюваності. Профільні методи є простими й інтерпретованими, але обмеженими для складних ботів; поведінкові – корисними для автоматизованої та координованої активності, але потребують історії дій; текстові – ефективними для аналізу змісту, проте залежними від мови й генеративних технологій; графові – перспективними для бот-мереж, але складними через вимоги до даних.

Гібридні методи можна вважати одним із найперспективніших напрямів виявлення ботів, оскільки вони зменшують залежність моделі від окремої групи ознак. Це особливо важливо для сучасних ботів, які можуть маскувати профільні характеристики, генерувати природний текст або уникати надто регулярної активності. Поєднання кількох модальностей дозволяє перевіряти акаунт з різних сторін: за профілем, поведінкою, змістом повідомлень і соціальними зв'язками.

Найбільш переконливе підтвердження перспективності гібридного підходу серед розглянутих матеріалів наведено у роботі CB-MTE, оскільки автори виконують дослідження впливу компонентів і показують, що об'єднання метаданих, тексту та графових ознак перевищує одномодальний варіант на тому самому наборі даних.

Водночас гібридні методи не слід безумовно визначати як найкращі для будь-якого практичного сценарію. Методи, які включають графові структури або складні мовні вставки, потребують більших обсягів даних, значніших обчислювальних ресурсів та складнішої процедури підтримки. Крім того, результати на різних датасетах не можна безпосередньо порівнювати: високі показники на Cresci-2017 не гарантують аналогічної якості на TwiBot-22 або на нових типах ботів.

Отримані результати можуть бути використані для формування вимог до перспективної багатовидової моделі виявлення ботів, яка поєднуватиме доступні атрибутивні, поведінкові та текстові ознаки, враховуватиме якість окремих представлень і забезпечуватиме аналіз внеску різних груп характеристик у фінальне рішення. Прикладом реалізації даних вимог може бути багатовидова модель з механізмом уваги, яка потребує експериментального підтвердження [21].

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Thales Group. (2025). AI-driven bots surpass human traffic: Bad Bot Report 2025. Thales Cloud Security Products. https://cpl.thalesgroup.com/about-us/newsroom/2025-imperva-bad-bot-report-ai-internet-traffic
2. Ng, L. H. X., & Carley, K. M. (2025). A global comparison of social media bot and human characteristics. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-96372-1
3. Wan, H., Luo, M., Ma, Z., Dai, G., & Zhao, X. (2025). How do social bots participate in misinformation spread? A comprehensive dataset and analysis. arXiv. http://arxiv.org/abs/2408.09613
4. Architecture of systems for detecting automated accounts (bots) in social networks. (2025). Security of Information Systems and Technologies, 1(9), 11-17. https://doi.org/10.17721/ISTS.2025.9.11-17
5. Hayawi, K., Mathew, S., Venugopal, N., Masud, M. M., & Ho, P.-H. (2022). DeeProBot: A hybrid deep neural network model for social bot detection based on user profile data. Social Network Analysis and Mining, 12(1). https://doi.org/10.1007/s13278-022-00869-w
6. Yang, K.-C., Varol, O., Hui, P.-M., & Menczer, F. (2020). Scalable and generalizable social bot detection through data selection. Proceedings of the AAAI Conference on Artificial Intelligence, 34(1), 1096-1103. https://doi.org/10.1609/aaai.v34i01.5460
7. Lopez-Joya, S., Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2024). Exploring social bots: A feature-based approach to improve bot detection in social networks. arXiv. https://doi.org/10.48550/arXiv.2411.06626
8. Rajendran, G., Ram, A., Vijayan, V., & Poornachandran, P. (2020). Deep temporal analysis of Twitter bots. In S. M. Thampi, L. Trajkovic, K.-C. Li, S. Das, M. Wozniak, & S. Berretti (Eds.), Machine learning and metaheuristics algorithms, and applications (pp. 38-48). Springer. https://doi.org/10.1007/978-981-15-4301-2_4
9. Chavoshi, N., Hamooni, H., & Mueen, A. (2017). Temporal patterns in bot activities. In Proceedings of the 26th International Conference Companion on World Wide Web (pp. 1601-1606). https://doi.org/10.1145/3041021.3051114
10. Chavoshi, N., Hamooni, H., & Mueen, A. (2016). Identifying correlated bots in Twitter. In Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part II (pp. 14-21). Springer. https://doi.org/10.1007/978-3-319-47874-6_2
11. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). Social fingerprinting: Detection of spambot groups through DNA-inspired behavioral modeling. IEEE Transactions on Dependable and Secure Computing, 15(4), 561-576. https://doi.org/10.1109/TDSC.2017.2681672
12. Mazza, M., Cresci, S., Avvenuti, M., Quattrociocchi, W., & Tesconi, M. (2019). RTbust: Exploiting temporal patterns for botnet detection on Twitter. arXiv. https://doi.org/10.48550/arXiv.1902.04506
13. Wei, F., & Nguyen, U. T. (2020). Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings. arXiv. https://doi.org/10.48550/arXiv.2002.01336
14. Ellaky, Z., & Benabbou, F. (2025). Advanced text-based transformer architecture for malicious social bots detection. Mathematical Modeling and Computing, 12(3), 972-981. https://doi.org/10.23939/mmc2025.03.972
15. Long, G., Lin, D., Lei, J., Guo, Z., Hu, Y., & Xia, L. (2023). A method of machine learning for social bot detection combined with sentiment analysis. In Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing (pp. 239-244). https://doi.org/10.1145/3578741.3578790
16. Ghosh, D., Boettcher, W., Johnston, R., & Lahiri, S. (2025). Bot identification in social media. arXiv. https://doi.org/10.48550/arXiv.2503.23629
17. Al-Khersan, Z. E. H., & Flayh, N. A. (2025). Enhanced Twitter bot detection via static and temporal feature integration. Journal of Al-Qadisiyah for Computer Science and Mathematics, 17(4). https://doi.org/10.29304/jqcsm.2025.17.42574
18. Chen, Q., Liao, Y., & Wang, C. (2025). Malicious social bot detection in social networks based on multi-modal feature fusion with transformer networks. In Proceedings of the 2024 2nd International Conference on Computer, Internet of Things and Smart City (pp. 173-176). https://doi.org/10.1145/3731867.3731896
19. Cheng, M., Xiao, Y., Huang, T., Lei, C., & Zhang, C. (2025). CB-MTE: Social bot detection via multi-source heterogeneous feature fusion. Sensors, 25(11), 3549. https://doi.org/10.3390/s25113549
20. Sakhabutdinov, I. (2025). Bot detection in social media: An empirical study using TwiBot-22 (Master's thesis). Luleå University of Technology. https://urn.kb.se/resolve?urn=urn:nbn:se:ltu:diva-114083
21. Buchyk, S., & Piatyhor, V. (2026). Model for identifying automated social network accounts based on a multi-species method with an attention mechanism. Cybersecurity: Education, Science, Technique, 4(32), 924-934. https://doi.org/10.28925/2663-4023.2026.32.1168

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (33), 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

METHODOLOGICAL PRINCIPLES FOR IDENTIFYING BOT ACCOUNTS ON SOCIAL MEDIA

Abstract. This article examines the methodological principles for identifying automated social media accounts. The relevance of the study is determined by the growing role of bots in the dissemination of misinformation, the artificial amplification of content, the manipulation of public opinion, and coordinated information campaigns. The emergence of generative artificial intelligence further complicates detection because automated accounts can produce natural language, vary their activity patterns, and imitate legitimate users more effectively. The reviewed approaches are systematized according to the primary source of features used for classification: user profile attributes, behavioral patterns, textual content, graph structures, and hybrid combinations of heterogeneous data. The methods are compared with respect to model type, datasets, reported evaluation metrics, data acquisition complexity, limitations, and explainability of classification decisions. Profile-based methods are shown to be scalable and relatively interpretable because they rely on accessible account metadata, but they may be insufficient for detecting sophisticated bots that maintain credible profiles. Behavioral approaches can reveal abnormal posting rhythms, repetitive activity, and coordination, although they require a sufficiently long activity history and are often applied to groups of accounts rather than individual users. Text-based methods can achieve strong classification results by analyzing message content through recurrent neural networks or transformer representations; however, they are sensitive to language, topic, dataset composition, and the increasing naturalness of AI-generated text. Graph-based characteristics are valuable for identifying coordinated amplification and bot networks: the reviewed evidence indicates differences between bot and human interaction structures, including denser bot ego-networks and a greater proportion of bot-to-bot links. At the same time, graph-based analysis is constrained by the cost, incompleteness, and limited reproducibility of large-scale interaction data acquisition. The analysis of hybrid approaches shows that combining profile, behavioral, textual, and, where available, graph features provides a more comprehensive representation of an account and can reduce dependence on a single weak modality. Nevertheless, complex multi-source models usually require more computational resources and have lower inherent explainability. The review identifies key unresolved problems: bot adaptability, heterogeneous and aging datasets, class imbalance, limited access to platform data, non-comparability of results obtained on different benchmarks, and insufficient interpretation of complex models. The results support the development of explainable hybrid detection models that combine accessible profile, behavioral, and textual features without critical dependence on a complete social graph.

Keywords: social networks; automated accounts; social bots; bot detection; machine learning; hybrid methods; explainability.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Thales Group. (2025). AI-driven bots surpass human traffic: Bad Bot Report 2025. Thales Cloud Security Products. https://cpl.thalesgroup.com/about-us/newsroom/2025-imperva-bad-bot-report-ai-internet-traffic
2. Ng, L. H. X., & Carley, K. M. (2025). A global comparison of social media bot and human characteristics. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-96372-1
3. Wan, H., Luo, M., Ma, Z., Dai, G., & Zhao, X. (2025). How do social bots participate in misinformation spread? A comprehensive dataset and analysis. arXiv. http://arxiv.org/abs/2408.09613

4. Architecture of systems for detecting automated accounts (bots) in social networks. (2025). Security of Information Systems and Technologies, 1(9), 11-17. https://doi.org/10.17721/ISTS.2025.9.11-17
5. Hayawi, K., Mathew, S., Venugopal, N., Masud, M. M., & Ho, P.-H. (2022). DeeProBot: A hybrid deep neural network model for social bot detection based on user profile data. Social Network Analysis and Mining, 12(1). https://doi.org/10.1007/s13278-022-00869-w
6. Yang, K.-C., Varol, O., Hui, P.-M., & Menczer, F. (2020). Scalable and generalizable social bot detection through data selection. Proceedings of the AAAI Conference on Artificial Intelligence, 34(1), 1096-1103. https://doi.org/10.1609/aaai.v34i01.5460
7. Lopez-Joya, S., Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2024). Exploring social bots: A feature-based approach to improve bot detection in social networks. arXiv. https://doi.org/10.48550/arXiv.2411.06626
8. Rajendran, G., Ram, A., Vijayan, V., & Poornachandran, P. (2020). Deep temporal analysis of Twitter bots. In S. M. Thampi, L. Trajkovic, K.-C. Li, S. Das, M. Wozniak, & S. Berretti (Eds.), Machine learning and metaheuristics algorithms, and applications (pp. 38-48). Springer. https://doi.org/10.1007/978-981-15-4301-2_4
9. Chavoshi, N., Hamooni, H., & Mueen, A. (2017). Temporal patterns in bot activities. In Proceedings of the 26th International Conference Companion on World Wide Web (pp. 1601-1606). https://doi.org/10.1145/3041021.3051114
10. Chavoshi, N., Hamooni, H., & Mueen, A. (2016). Identifying correlated bots in Twitter. In Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part II (pp. 14-21). Springer. https://doi.org/10.1007/978-3-319-47874-6_2
11. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). Social fingerprinting: Detection of spambot groups through DNA-inspired behavioral modeling. IEEE Transactions on Dependable and Secure Computing, 15(4), 561-576. https://doi.org/10.1109/TDSC.2017.2681672
12. Mazza, M., Cresci, S., Avvenuti, M., Quattrociocchi, W., & Tesconi, M. (2019). RTbust: Exploiting temporal patterns for botnet detection on Twitter. arXiv. https://doi.org/10.48550/arXiv.1902.04506
13. Wei, F., & Nguyen, U. T. (2020). Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings. arXiv. https://doi.org/10.48550/arXiv.2002.01336
14. Ellaky, Z., & Benabbou, F. (2025). Advanced text-based transformer architecture for malicious social bots detection. Mathematical Modeling and Computing, 12(3), 972-981. https://doi.org/10.23939/mmc2025.03.972
15. Long, G., Lin, D., Lei, J., Guo, Z., Hu, Y., & Xia, L. (2023). A method of machine learning for social bot detection combined with sentiment analysis. In Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing (pp. 239-244). https://doi.org/10.1145/3578741.3578790
16. Ghosh, D., Boettcher, W., Johnston, R., & Lahiri, S. (2025). Bot identification in social media. arXiv. https://doi.org/10.48550/arXiv.2503.23629
17. Al-Khersan, Z. E. H., & Flayh, N. A. (2025). Enhanced Twitter bot detection via static and temporal feature integration. Journal of Al-Qadisiyah for Computer Science and Mathematics, 17(4). https://doi.org/10.29304/jqcsm.2025.17.42574
18. Chen, Q., Liao, Y., & Wang, C. (2025). Malicious social bot detection in social networks based on multi-modal feature fusion with transformer networks. In Proceedings of the 2024 2nd International Conference on Computer, Internet of Things and Smart City (pp. 173-176). https://doi.org/10.1145/3731867.3731896
19. Cheng, M., Xiao, Y., Huang, T., Lei, C., & Zhang, C. (2025). CB-MTE: Social bot detection via multi-source heterogeneous feature fusion. Sensors, 25(11), 3549. https://doi.org/10.3390/s25113549
20. Sakhabutdinov, I. (2025). Bot detection in social media: An empirical study using TwiBot-22 (Master's thesis). Luleå University of Technology. https://urn.kb.se/resolve?urn=urn:nbn:se:ltu:diva-114083
21. Buchyk, S., & Piatyhor, V. (2026). Model for identifying automated social network accounts based on a multi-species method with an attention mechanism. Cybersecurity: Education, Science, Technique, 4(32), 924-934. https://doi.org/10.28925/2663-4023.2026.32.1168

Отримано редакцією журналу / Received: 28.02.26
Прорецензовано / Revised: 03.03.26
Схвалено до друку / Accepted: 25.06.26