КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (34), 2026
ISSN 2663 – 4023
DOI 10.28925/2663-4023.2026.34.1298
УДК 004.8:004.056.5:004.932:004.89

Висоцька Вікторія Анатоліївна
д.т.н., доцент, професор кафедри інформаційних систем та мереж
Національний університет «Львівська політехніка», Львів, Україна
ORCID:0000-0001-6417-3689
victoria.a.vysotska@lpnu.ua

Чирун Любомир Вікторович
к.т.н., доцент кафедри інформаційних систем та мереж
Національний університет «Львівська політехніка», Львів, Україна
ORCID:0000-0002-9448-1751
lyubomyr.v.chyrun@lpnu.ua

Романчук Роман Васильович
аспірант кафедри інформаційних систем та мереж
Національний університет «Львівська політехніка», Львів, Україна
ORCID:0009-0004-4352-1073
roman.v.romanchuk@lpnu.ua

ІНФОРМАЦІЙНА БЕЗПЕКА КІБЕРПРОСТОРУ СОЦІАЛЬНИХ МЕРЕЖ ШЛЯХОМ ВИЯВЛЕННЯ СПОСОБІВ РОЗПОВСЮДЖЕННЯ ДЕЗІНФОРМАЦІЇ ТА НЕАВТЕНТИЧНОЇ ПОВЕДІНКИ КОРИСТУВАЧІВ ЧАТІВ НА ОСНОВІ ГЛИБИННОГО НАВЧАННЯ ТА МУЛЬТИМОДАЛЬНОГО АНАЛІЗУ

Анотація. У статті розглянуто проблему виявлення джерел та шляхів розповсюдження дезінформації й неавтентичної поведінки користувачів чатів у сучасному інформаційному просторі. Актуальність дослідження обумовлена стрімким зростанням обсягів цифрового контенту, поширенням соціальних мереж і чат-платформ, а також активізацією інформаційних атак, маніпулятивного впливу та використанням автоматизованих бот-мереж. Показано, що традиційні підходи, засновані виключно на аналізі текстового контенту, мають суттєві обмеження через неможливість повного врахування поведінкових характеристик користувачів, мультимедійних компонентів та структурних зв'язків між учасниками інформаційного середовища. У першому розділі роботи виконано аналіз сучасного стану проблеми виявлення дезінформації, досліджено основні методи автоматичного аналізу неправдивого контенту та неавтентичної поведінки користувачів. Розглянуто традиційні підходи машинного навчання, сучасні моделі глибинного навчання, мультимодальні методи аналізу та графові нейронні мережі. Проведено порівняльний аналіз переваг і недоліків існуючих підходів, визначено проблемні аспекти та окреслено перспективні напрями розвитку. У розділі постановки задачі запропоновано концептуальний pipeline дослідження, який охоплює процеси збору мультимодальних даних, їх попередньої обробки, виділення ознак, мультимодального об'єднання, навчання моделей глибинного навчання, побудови графів взаємодії користувачів, аналізу інформаційних каскадів та виявлення потенційних джерел дезінформації. Розроблено математичну модель, що інтегрує текстові, поведінкові та графові характеристики. У розділі методів та матеріалів описано математичний апарат моделі, яка поєднує Transformer-моделі для текстового аналізу, згорткові нейронні мережі для виділення візуальних ознак, рекурентні мережі для поведінкового аналізу та Graph Neural Network для моделювання структури взаємодій між користувачами. Запропоновано механізм мультимодального об'єднання ознак із використанням механізмів уваги та комбінованих функцій втрат. У експериментальній частині проведено дослідження на основі мультимодальних наборів даних, що включали текстові повідомлення, часові ряди активності та мережеві характеристики користувачів. Отримані результати продемонстрували ефективність запропонованого підходу при ідентифікації потенційних джерел дезінформації, виявленні аномальних поведінкових шаблонів та прогнозуванні шляхів поширення інформаційних повідомлень. Аналіз інтерактивних графів поширення та карт ймовірностей підтвердив здатність системи адаптуватися до різних структур мереж і типів інформаційних каскадів.

Практична значущість дослідження полягає у можливості використання розробленого підходу в системах моніторингу інформаційного простору, кібербезпеки, соціальних мережах, чат-платформах та системах раннього виявлення інформаційних загроз.

Ключові слова: дезінформація; мультимодальний аналіз; глибинне навчання; графові нейронні мережі; неавтентична поведінка; соціальні мережі; чат-платформи; інформаційні каскади; виявлення джерел; машинне навчання.

ВСТУП

У сучасному інформаційному просторі стрімке зростання обсягів цифрового контенту та популярність онлайн-комунікацій створюють сприятливі умови для поширення дезінформації та проявів неавтентичної поведінки користувачів. Чати, соціальні платформи та месенджери дедалі частіше стають середовищем для маніпуляцій громадською думкою, координації інформаційних атак та поширення фейкових повідомлень. Особливої актуальності ця проблема набуває в умовах гібридних загроз, інформаційних війн та кризових ситуацій, коли достовірність інформації є критично важливою. Традиційні підходи до виявлення дезінформації, що базуються переважно на текстовому аналізі, виявляються недостатньо ефективними в умовах зростання складності та багатоформатності сучасного контенту. Користувачі активно використовують зображення, відео, аудіо та їх комбінації, що ускладнює ідентифікацію маніпуляцій та штучно згенерованих повідомлень. У зв’язку з цим виникає потреба у застосуванні мультимодального аналізу, який дозволяє одночасно обробляти різні типи даних і виявляти приховані закономірності. Вагому роль у розв’язанні цієї задачі відіграють методи глибинного навчання, здатні автоматично витягувати складні ознаки з великих обсягів даних та забезпечувати високу точність класифікації. Поєднання мультимодального підходу з сучасними нейромережевими архітектурами відкриває нові можливості для виявлення джерел дезінформації, аналізу поведінкових патернів користувачів та відстеження шляхів поширення неправдивої інформації.

Постановка проблеми. Сучасний розвиток цифрових технологій, соціальних мереж, месенджерів та чат-платформ призвів до стрімкого збільшення обсягів інформації, що створюється та поширюється користувачами в режимі реального часу. Разом із позитивними аспектами швидкого інформаційного обміну спостерігається значне зростання обсягів неправдивої, маніпулятивної та дезінформаційної інформації, яка може використовуватися для впливу на громадську думку, дестабілізації суспільних процесів, інформаційно-психологічного впливу та реалізації цілеспрямованих інформаційних атак. Особливу небезпеку становить використання автоматизованих акаунтів, бот-мереж та скоординованої неавтентичної поведінки користувачів, які здатні значно прискорювати процес поширення дезінформації та збільшувати її охоплення.

Складність вирішення задачі виявлення дезінформації полягає у тому, що сучасні інформаційні повідомлення представлені не лише текстовими даними, а й містять зображення, відеоматеріали, аудіокомпоненти, часові характеристики активності користувачів та різноманітні поведінкові параметри. Крім того, суттєву роль відіграє структура взаємодій між учасниками мережі, яка визначає механізми поширення інформаційних потоків та формування інформаційних каскадів. Традиційні методи аналізу даних, що ґрунтуються виключно на текстових ознаках або окремих статистичних характеристиках, часто не забезпечують достатньої точності через нездатність враховувати складні взаємозв'язки між різними типами даних та динаміку поведінки користувачів. Важливим науковим завданням є розроблення інтелектуальних методів інтегрованого аналізу мультимодальних даних, які дозволяють поєднувати семантичні характеристики повідомлень, поведінкові особливості користувачів та структурні властивості мережевих взаємодій. Особливої актуальності набуває використання сучасних методів глибинного навчання, зокрема трансформерних архітектур, механізмів уваги та графових нейронних мереж, здатних враховувати як локальні характеристики повідомлень, так і глобальні зв'язки між користувачами.

Таким чином, актуальною науково-практичною проблемою є створення методів та моделей для виявлення джерел дезінформації, ідентифікації неавтентичної поведінки користувачів та прогнозування шляхів поширення інформації на основі мультимодального аналізу та технологій глибинного навчання, що дозволить підвищити ефективність систем моніторингу інформаційного середовища та забезпечити своєчасне виявлення інформаційних загроз.

Аналіз останніх досліджень і публікацій. Проблема поширення дезінформації та неавтентичної поведінки в онлайн-середовищах привертає значну увагу як наукової спільноти, так і практичних розробників, оскільки вона ставить під загрозу достовірність інформаційного простору, соціальну стабільність та безпеку цифрових платформ. У низці досліджень різні автори пропонували методи

детекції дезінформації, оцінювання надійності користувачів та виявлення аномальних патернів поведінки на основі аналізу великих даних та алгоритмів машинного навчання.

Існують традиційні підходи до виявлення дезінформації, підходи на основі глибинного навчання, мультимодальний аналіз, графові нейронні мережі та різні алгоритми виявлення джерел та шляхів розповсюдження.

Перші системи виявлення дезінформації ґрунтувалися на класичних методах обробки природної мови (NLP) та машинного навчання. Вони зазвичай базувалися на текстовому аналізі, включно з методами на основі ознак (feature-based) та статистичного моделювання. Наприклад, досліджували автоматичне визначення правдивості тверджень, використовуючи лінгвістичні ознаки та класифікатори SVM або розглядали ознаки тональності, частотні патерни та семантичні структури для оцінки достовірності інформації. Такі підходи використовують лінгвістичні, статистичні та семантичні характеристики текстів для класифікації інформаційних повідомлень на достовірні та недостовірні. У роботі [1] запропоновано підхід до розпізнавання фейків, пропаганди та дезінформації в українськомовному контенті на основі NLP-технологій і алгоритмів машинного навчання. Автори продемонстрували ефективність використання морфологічного аналізу, тематичного моделювання та методів класифікації для виявлення маніпулятивного контенту. Подібний напрям розвивається у роботі [2], де запропоновано інструмент аналізу текстового контенту для виявлення ключових пропагандистських наративів, а також [3], присвяченій виявленню шахрайських вакансій у сфері електронного бізнесу. Окрему увагу привертає дослідження [4], у якому розроблено гібридну інформаційну технологію для автоматичного виявлення дезінформації та неавтентичної поведінки користувачів чатів. Запропоноване рішення поєднує методи NLP, машинного навчання та графового аналізу соціальних взаємодій. Однак текстовий аналіз має обмеження при роботі з мультимодальними даними і не враховує контекст поведінки користувачів чи структуру соціальних взаємодій, що знижує його ефективність у складних сценаріях поширення дезінформації. Попри високу інтерпретованість і відносно низькі обчислювальні витрати, традиційні методи мають обмежену здатність до виявлення прихованих семантичних зв'язків і складних патернів поширення дезінформації.

У зв’язку з успіхами глибинного навчання в обробці природної мови (NLP) та задачах комп’ютерного зору, сучасні дослідження активно застосовують нейронні архітектури для класифікації та виявлення фейкових новин. Значний прогрес у задачах виявлення дезінформації було досягнуто завдяки застосуванню моделей глибинного навчання. На відміну від класичних методів, вони здатні автоматично формувати інформативні ознаки без необхідності ручного конструювання характеристик. Однією з найвідоміших моделей є FakeBERT, запропонована Kaliyar et al. [5], яка поєднує можливості трансформерної архітектури BERT із механізмами класифікації для аналізу новинного контенту. Подальший розвиток цього напряму представлено у роботі Wang [6], де розглянуто використання різних архітектур глибоких нейронних мереж для задач класифікації фейкових новин. У дослідженні Almandouh et al. [7] запропоновано ансамблевий підхід, що об'єднує кілька моделей глибинного навчання для підвищення точності виявлення дезінформації. Embarak [8] використовує глибокі нейронні мережі для аналізу мереж поширення дезінформації у Facebook, з акцентом на зв’язки між користувачами, тоді як Ayyasamy et al. [9] розробили гібридну модель на основі LSTM та графових нейронних мереж CGPNN із застосуванням метаевристичної оптимізації для підвищення ефективності виявлення дезінформації. У роботі Roumeliotis et al. [10] проведено порівняльний аналіз згорткових нейронних мереж, великих мовних моделей та сучасних класичних NLP-підходів для виявлення фейкових новин у соціальних мережах. Результати підтверджують перевагу трансформерних архітектур та великих мовних моделей у задачах аналізу складного інформаційного контенту. Загалом методи глибинного навчання забезпечують вищу точність порівняно з класичними алгоритмами, однак потребують значних обсягів навчальних даних та обчислювальних ресурсів. Поєднання CNN для текстових ознак, RNN для поведінкових характеристик та вбудовування користувачів для детекції фейків довело, що глибинні моделі, які інтегрують поведінку користувачів разом з текстовими ознаками, значно перевершують чисто текстові методи. Використання архітектур Transformer, як-от BERT, продемонструвало високу ефективність у задачах класифікації дезінформації, особливо при поєднанні з виявленням політичних тональностей та контекстної семантики.

Сучасні інформаційні кампанії активно використовують не лише текстовий контент, але й зображення, відео та поведінкові характеристики користувачів. Це стимулювало розвиток мультимодальних методів аналізу дезінформації. Мультимодальні підходи в останні роки активно досліджуються через потребу аналізувати не лише текст, а й зображення, відео, аудіо та поведінкові сигнали. У роботі Zhang et al. [11] запропоновано Multimodal Inverse Attention Network, яка інтегрує текстові та візуальні ознаки за допомогою механізму зворотної уваги. Розроблена мережа виділяє дискримінативні ознаки для точного розпізнавання фейкових повідомлень. Shen et al. [12] розробили

модель GAMED, що використовує кілька експертних модулів для адаптивного аналізу різних типів мультимодальних даних для виявлення фейкових новин. Дослідження Haque et al. [13] доповнює мультимодальний аналіз поведінковими біометричними характеристиками користувачів, що дозволяє враховувати особливості взаємодії аудиторії з інформаційним контентом. Легковагова мультимодальна система класифікації дезінформації, що поєднує лінгвістичні та поведінкові біометричні дані.У роботі Hu et al. [14] запропоновано часові мультимодальні мережі, які моделюють еволюцію процесу поширення фейкових новин у часі. Представлено динамічні тимчасові графові мережі для мультимодального виявлення фейкових новин з урахуванням часових залежностей. Перевагою мультимодальних моделей є можливість комплексного аналізу різнорідних джерел інформації, що особливо важливо в умовах сучасних соціальних мереж, де текстова інформація часто супроводжується візуальними матеріалами. Є популярним розроблення мультимодальну систему для виявлення фейкових медіа-повідомлень, поєднуючи CNN для зображень та LSTM для послідовності тексту. Комбінування модальностей призводить до значного підвищення точності детекції. У контексті поведінки користувачів аналізують часові патерни активності та графи взаємодій для виявлення ботів та координованих груп, які часто є джерелами неправдивої інформації. Врахування часових і мережевих зв’язків суттєво покращує якість класифікації.

Останніми роками значну популярність здобули графові нейронні мережі (Graph Neural Networks, GNN), які дозволяють моделювати структуру взаємозв'язків між користувачами, повідомленнями та інформаційними ресурсами. GNN стали ключовим інструментом для аналізу соціальних мереж завдяки здатності враховувати структуру зв’язків між користувачами. Такі моделі є ефективними для виявлення аномальних патернів у графових даних, зокрема для задач фейкового акаунту та бот-детекції. Огляд сучасних GNN-підходів представлено у роботах Gong et al. [15], Phan et al. [16] та Mahdi and Shati [17]. Автори відзначають, що графові моделі дозволяють враховувати як властивості окремих повідомлень, так і структуру їх поширення в мережі. Здійснено систематичний огляд застосування графових нейронних мереж у задачах детекції фейкових новин, в тому числі у соціальних медіа та їх ефективності у сучасних задачах. Серед практичних реалізацій варто виділити роботу Ren et al. [18], де використовується неоднорідна графова нейронна мережа з активним навчанням, а також модель CLIP-GCN, запропоновану Zhou et al. [19], яка поєднує мультимодальні ознаки та графові згортки. Представлена адаптивна модель Clip-GCN для виявлення фейкових новин у мультимодальних джерелах поєднує графові нейронні мережі та обробку зображень. Подальший розвиток цього напряму представлено в роботах Li [20], Rama Moorthy et al. [21], Yin et al. [22], Abe et al. [23] та Zhang et al. [24], де інтегруються трансформери, механізми уваги та графові згорткові мережі. В [20] запропоновано метод мультимодального виявлення дезінформації за допомогою графових нейронних мереж і механізмів уваги, що ефективно інтегрує текст і зображення. В [23] запропоновано модель для виявлення фейкових новин із використанням часових рядів та структурних особливостей графа. GBCA в [24] поєднує GCN і BERT із механізмом спільної уваги для точнішого виявлення дезінформації. Особливий інтерес становить модель KGFakeNet [25], яка використовує графи знань для виявлення суперечностей між фактами та новинними повідомленнями. KGFakeNet інтегрує знання графів для поліпшення розпізнавання фейкових новин, використовуючи структуровану інформацію. Графові нейронні мережі демонструють високу ефективність завдяки здатності враховувати структурні особливості соціальних мереж і взаємодій між користувачами. Їх часто застосовують для виявлення фейкових новин у соціальних мережах, де вершини графу представляють користувачів та повідомлення, а ребра – взаємодії між ними. Також використовують для інтеграції інформації про повідомлення та характеристики взаємодій, що дозволило виділити ключові шляхи поширення дезінформації.

Одним із найскладніших і водночас найважливіших напрямів сучасних досліджень є виявлення первинних джерел дезінформації та аналіз шляхів її поширення в інформаційному просторі [26-28]. Окремий напрямок досліджень стосується безпосередньо моделювання розповсюдження інформації у мережах та виявлення цюрок розповсюдження – “information cascades”. У роботі Vysotska et al. [26] розроблено інтелектуальний інструмент для визначення джерел та маршрутів поширення дезінформації у соціальних мережах на основі технологій NLP та машинного навчання. Подальший розвиток цієї тематики представлено у роботі Nazarkevych et al. [27], де запропоновано інноваційний метод пошуку джерел поширення фейкових новин із використанням теорії графів та алгоритмів машинного навчання. Важливими для аналізу інформаційних каскадів є дослідження Yin et al. [22] та Abe et al. [23], які враховують часову динаміку поширення повідомлень у графових структурах. В [22] розроблено двосторонню графову модель із врахуванням часових затримок для детекції фейкових новин, що покращує точність при обробці потокових даних. У роботі Hu et al. [14] застосовано часові мультимодальні мережі для відстеження траєкторій розповсюдження фейкових новин, тоді як Embarak [8] аналізує мережі дезінформації Facebook з використанням методів глибинного навчання. Дослідження

показали, як структурні властивості мереж впливають на швидкість та масштаб розповсюдження інформації. Необхідно включати часові фактори у моделі поширення інформації, поєднуючи моделі дифузії з нейронними мережами для прогнозування шляхів розповсюдження. Особливо актуальними є роботи, які об'єднують поведінкові патерни, текстові ознаки та мережеві структури у єдину модель. Наприклад, багаторівневий підхід поєднує мультимодальний контент та часові графи для прогнозу шляху розповсюдження повідомлень, демонструючи, що інтеграція різних модальностей суттєво покращує прогнозні здібності моделей. Незважаючи на значні досягнення, проблема точного визначення першоджерел дезінформації залишається відкритою через високу динамічність соціальних мереж, наявність ботів та координованих інформаційних кампаній.

Текстовий аналіз, хоч і є базовим, виявився недостатнім для повноцінного розв'язання задачі, оскільки дезінформація часто маскується за контекстно коректними фразами. Глибинні нейронні мережі (BERT, CNN, RNN) значно покращують якість виявлення завдяки автоматичному виділенню ознак і моделюванню складних патернів. Мультимодальні моделі, які поєднують кілька типів даних, показали найкращі результати для задач пов'язаних із неправдивою інформацією та аномальною поведінкою. Графові підходи здатні моделювати структуру взаємодій та шляхи розповсюдження, що є ключовим для виявлення джерел і впливових вузлів у мережі. Сучасні дослідження демонструють, що комбінація GNN із мультимодальними репрезентаціями є перспективною стратегією для глибокого аналізу та прогнозування дезінформаційних феноменів у чат-середовищах. Проведений аналіз літератури свідчить про еволюцію методів виявлення дезінформації від класичних алгоритмів машинного навчання до складних мультимодальних моделей і графових нейронних мереж. Найбільш перспективними сьогодні є підходи, що поєднують аналіз текстового контенту, візуальної інформації, поведінкових характеристик користувачів та структури соціальних взаємодій. Водночас більшість існуючих досліджень зосереджені переважно на класифікації контенту, тоді як задача виявлення джерел та маршрутів поширення дезінформації залишається недостатньо дослідженою. Це обумовлює актуальність розроблення нових мультимодальних графових моделей, здатних одночасно виявляти фейковий контент, визначати його походження та прогнозувати подальше поширення в інформаційному просторі.

Мета статті. Метою даної роботи є розробка та аналіз методів і засобів виявлення джерел і шляхів розповсюдження дезінформації, а також ідентифікації неавтентичної поведінки користувачів чатів на основі технологій мультимодального аналізу та глибинного навчання. На основі дослідження та узагальнення сучасних методів і засобів виявлення джерел та каналів розповсюдження дезінформації необхідно розробити метод ідентифікації неавтентичної поведінки користувачів чатів на основі технологій мультимодального аналізу та глибинного навчання. У роботі розглядаються ключові підходи, їх переваги та обмеження, а також окреслюються перспективи подальших досліджень у цій сфері. Для досягнення мети необхідно вирішити такі задачі:

– проаналізувати сучасний стан проблеми поширення дезінформації та неавтентичної поведінки в чатах і соціальних платформах;

– дослідити існуючі підходи до виявлення дезінформації, зокрема на основі текстового, візуального та аудіоаналізу;

– розглянути методи мультимодального аналізу та визначити їх ефективність у задачах виявлення маніпулятивного контенту;

– проаналізувати можливості моделей глибинного навчання для класифікації та виявлення аномальної поведінки користувачів;

– розробити підхід та модель для інтегрованого аналізу різнорідних даних;

– оцінити ефективність запропонованих рішень через відповідні критерії якості.

Наукова новизна одержаних результатів полягає в наступному:

– запропоновано підхід до виявлення дезінформації, що базується на поєднанні мультимодального аналізу (текст, відео, поведінкові дані) з глибинним навчанням;

– удосконалено методи ідентифікації неавтентичної поведінки користувачів чатів шляхом врахування комплексних поведінкових характеристик;

– набули подальшого розвитку підходи до визначення джерел та каналів поширення дезінформації через аналіз взаємозв'язків між користувачами та контентом.

Практична цінність роботи полягає в тому, що:

– розроблені підходи можуть бути використані для створення систем моніторингу інформаційного простору в режимі реального часу;

– результати дослідження можуть бути впроваджені у чат-платформи, соціальні мережі та системи кібербезпеки для підвищення рівня довіри до інформації;

– запропоновані методи можуть застосовуватись у сфері інформаційної безпеки, зокрема для протидії інформаційним атакам та бот-мережам;

– можливе використання в навчальних та тренувальних системах (зокрема VR/AR-симуляціях) для моделювання сценаріїв інформаційного впливу.

ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

Дослідження виявлення джерел та шляхів розповсюдження дезінформації і неавтентичної поведінки користувачів чатів на основі мультимодального аналізу та глибинного навчання доцільно реалізовувати як послідовність взаємопов'язаних етапів:

формування мультимодальних даних → попередня обробка → виділення ознак → мультимодальне об'єднання → класифікація дезінформації → моделювання поведінки користувачів → побудова графа взаємодій → графова нейронна мережа → виявлення джерел дезінформації → моделювання поширення інформації → загальна функція втрат

Опишемо більш детально основний pipeline дослідження таматематично формалізуємо основний pipeline дослідження у вигляді математичної моделі. На першому етапі збору та формування набору мультимодальних даних (Data Collection) здійснюється збір контенту із чат-платформ, месенджерів та соціальних мереж. Дані можуть включати: текстові повідомлення; зображення та відео; аудіофайли (за наявності); метадані (час, частота повідомлень, зв'язки між користувачами). Додатково формується розмічений датасет: класи (достовірна інформація / дезінформація) та автентична / неавтентична поведінка користувачів. Нехай маємо вхідний набір даних $D = \{(x_i^{(t)}, x_i^{(v)}, x_i^{(a)}, m_i, y_i)\}_{i=1}^{N}$, де $x_i^{(t)}$ – текстові дані, $x_i^{(v)}$ – візуальні дані (зображення/відео), $x_i^{(a)}$ – аудіо дані, m_i – метадані (час, зв'язки), $y_i \in \{0,1\}$ – мітка (дезінформація/ні). На наступному етапі попередньої обробки даних (Preprocessing)виконується очищення та підготовка даних: текст (токенізація, лематизація, видалення стоп-слів), зображення/відео (нормалізація, ресайз, виділення ключових кадрів), аудіо (перетворення у спектрограми) та синхронізація мультимодальних даних. Також проводиться анонімізація та балансування датасету: $\widetilde{x_i^{(k)}} = \phi_k\left(x_i^{(k)}\right), \quad k \in \{t, v, a\}$,де ϕ_k – функції нормалізації та очищення даних.

На третьому етапі «Виділення ознак» (Feature Extraction) формуємо ознаки з різних модальностей: текстові ознаки (семантика, тональність, стилометричні характеристики), візуальні ознаки(об'єкти, сцени, ознаки маніпуляцій), аудіо ознаки (інтонація, шуми, синтетичність) та поведінкові ознаки(частота повідомлень, часові патерни, мережеві зв'язки). Для кожної модальності розраховуємо:

$$h_i^{(t)} = f_t\left(\widetilde{x_i^{(t)}}\right), \qquad h_i^{(v)} = f_v\left(\widetilde{x_i^{(v)}}\right), h_i^{(a)} = f_a\left(\widetilde{x_i^{(a)}}\right) \tag{1}$$

де f_t – текстова модель (наприклад, Transformer), f_v – CNN, f_a – аудіо модель.

На етапі «Мультимодальне об'єднання даних» (Multimodal Fusion)об'єднуються ознаки з різних джерел: раннє об'єднання (early fusion) на рівні ознак, пізнє об'єднання (late fusion) на рівні рішень моделей та гібридні підходи. Метою є отримання узгодженого представлення інформації: $h_i = F\left(h_i^{(t)}, h_i^{(v)}, h_i^{(a)}, m_i\right)$, наприклад, через конкатенацію $h_i = \left[h_i^{(t)} | h_i^{(v)} | h_i^{(a)} | m_i\right]$ або через attention $h_i = \sum_k \alpha_k h_i^{(k)}$, $\sum_k \alpha_k = 1$.

На етапі «Побудова та навчання моделей» (Model Training) використовуємо моделі глибинного навчання: для тексту Transformer (BERT-подібні моделі), для зображень CNN, для послідовностей LSTM/GRU та для мультимодальних задач: комбіновані архітектури. Основні задачі моделей: класифікація дезінформації → виявлення ботів та неавтентичних акаунтів → виявлення аномалій. При ласифікації дезінформації ймовірність $\widehat{y_i} = \sigma(Wh_i + b)$ та функція втрат (бінарна крос-ентропія) $\mathcal{L}_{c\ell s} = -\sum_{i=1}^{N}[y_i \log \widehat{y_i} + (1 - y_i)\log(1 - \widehat{y_i})]$. Нехай для моделювання поведінки користувачів $u_j = \{m_{j1}, m_{j2}, \dots, m_{jn}\}$, тоді поведінкове представлення $z_j = g\left(u_j\right)$, аномальність $s_j = |z_j - \mu|$, де μ – середній профіль користувача.

На етапі аналізу соціальних графів (Graph Analysis) формується граф взаємодій користувачів, де вузли – користувачі та ребра – комунікації між ними. Застосовуються для дослідження графові нейронні мережі (GNN), алгоритми пошуку спільнот та аналіз центральності. Це дозволяє виявляти джерела дезінформації та визначати ключові вузли поширення. Якщо граф $G = (V, E)$, тоді матриця суміжності:

$$A_{ij} = \begin{cases} 1, & \text{якщо є взаємодія,} \\ 0, & \text{інакше,} \end{cases} \tag{2}$$

оновлення вузлів $H^{(l+1)} = \sigma\left(AH^{(l)}W^{(l)}\right)$, де $H^{(l)}$ – представлення вузлів, $W^{(l)}$ – ваги.

На етапі виявлення шляхів розповсюдження (Propagation Tracking) аналізуються часові ряди поширення повідомлень, каскади репостів/пересилань та швидкість та масштаб поширення. Будуються моделі дифузії інформації. Оцінка важливості вузла для виявлення джерел дезінформації $c_j = \sum_i A_{ij}$, або через PageRank:

$$PR(j) = \frac{1-d}{|V|} + d \sum_{k \in In(j)} \frac{PR(k)}{L(k)} \tag{3}$$

Каскад поширення при моделюванні поширення інформації $P(t) = \beta \sum_j A_{ij} y_j(t-1)$ або через дифузійну модель $\frac{dI(t)}{dt} = \beta\, S(t)\, I(t)$. Загальна функція втрат тоді $\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{graph} + \lambda_2 \mathcal{L}_{behavior}$. При оцінюванні ефективності (Evaluation) використовуються метрики: Accuracy, Recall, F1-score, ROC-AUC та метрики для графів (наприклад, modularity). Проводиться порівняння з базовими методами. Інтерпретація та візуалізація результатів побудована на візуалізації графів поширення, поясненні рішень моделей та визначенні ключових факторів дезінформації. Розгортання та практичне застосування здійснюємо через інтеграцію у системи моніторингу, роботу в режимі реального часу та використання в системах кібербезпеки та аналітики.

МЕТОДИКА ДОСЛІДЖЕННЯ

Узагальнена схема pipeline дослідження: *збір даних* → *попередня обробка* → *виділення ознак* → *мультимодальне об'єднання* → *навчання моделей* → *графовий аналіз* → *виявлення поширення* → *оцінка* → *візуалізація* → *впровадження*. Весь pipeline подаємо як композицію функцій $\hat{y}, \mathrm{z}, \mathrm{G} = \mathcal{F}(D)$. Модель об'єднує мультимодальне представлення даних, глибинне навчання, графовий аналіз, моделі поширення інформації. Для формування мультимодальних даних нехай задано множину користувачів чат-платформи $U = \{u_1, u_2, \dots, u_M\}$. Кожен користувач u_j генерує послідовність повідомлень $S_j = \{s_{j1}, s_{j2}, \dots, s_{jT_j}\}$. Кожне повідомлення s_{ji} описується як мультимодальний об'єкт $s_{ji} = \left(x_{ji}^{(t)}, x_{ji}^{(v)}, x_{ji}^{(a)}, m_{ji}\right)$, де $x_{ji}^{(t)}$ – текстова складова повідомлення, $x_{ji}^{(v)}$ – візуальна складова (зображення або відео), $x_{ji}^{(a)}$ – аудіо складова (за наявності), m_{ji} – метадані. Метадані повідомлення задаємо як вектор ознак $m_{ji} = \left(t_{ji}, f_{ji}, r_{ji}, c_{ji}\right)$, де t_{ji} – час відправлення, f_{ji} – частота активності користувача, r_{ji} – кількість пересилань/репостів, c_{ji} – контекст взаємодії (ідентифікатори співрозмовників). Загальний набір даних (повний датасет) визначається як $D = \bigcup_{j=1}^{M} \bigcup_{i=1}^{T_j} \{s_{ji}\}$, або у вигляді $D = \{(x_k^{(t)}, x_k^{(v)}, x_k^{(a)}, m_k, y_k)\}_{k=1}^{N}$, де N – загальна кількість повідомлень, $y_k \in \{0,1\}$ – мітка класу:

$$y_k = \begin{cases} 1, \text{якщо повідомлення містить дезінформацію}, \\ 0, \text{інакше}. \end{cases} \tag{4}$$

При формуванні поведінкового профілю користувача для кожного користувача формується агрегований профіль $p_j = \Psi\left(S_j\right) = \frac{1}{T_j} \sum_{i=1}^{T_j} m_{ji}$, де Ψ – функція агрегації поведінкових характеристик.

Модель взаємодії між користувачами задається графом $G = (V, E)$, де $V = U$ – множина користувачів, $E = \{(u_i, u_j)\}$ – множина зв'язків. Матриця суміжності:

$$A_{ij} = \begin{cases} 1, \text{якщо існує взаємодія між } u_i \text{ та } u_j, \\ 0, \quad \text{інакше}. \end{cases} \tag{5}$$

Для аналізу поширення інформації вводиться часовий ряд, тобто здійснюється формування часових залежностей $T_k = \{t_k^{(1)}, t_k^{(2)}, \dots, t_k^{(n)}\}$, що відображає динаміку поширення повідомлення. Таким чином, мультимодальні дані задаються як $D = \{(x_k^{(t)}, x_k^{(v)}, x_k^{(a)}, m_k, p_{u(k)}, G, T_k, y_k)\}$, де $u(k)$ – користувач, що створив повідомлення k. Сформований мультимодальний датасет інтегрує різнорідні типи контенту, поведінкові характеристики користувачів, структуру соціальних зв'язків, часову динаміку поширення, що створює основу для подальшого застосування методів глибинного навчання та аналізу дезінформації.

Для попередньої обробки мультимодальних даних нехай вхідний набір мультимодальних даних $D = \{(x_k^{(t)}, x_k^{(v)}, x_k^{(a)}, m_k, y_k)\}_{k=1}^{N}$. Метою попередньої обробки є перетворення сирих даних у

стандартизоване представлення $\widetilde{D} = \{(\widetilde{x_k^{(t)}}, \widetilde{x_k^{(v)}}, \widetilde{x_k^{(a)}}, \widetilde{m_k}, y_k)\}$. Текстова модальність попередньо обробляється функцією $\widetilde{x_k^{(t)}} = \phi_t(x_k^{(t)})$, та $\phi_t = \phi_{norm} \circ \phi_{token} \circ \phi_{lemma} \circ \phi_{clean}$, де $x' = \phi_{clean}(x) = x \setminus$ {noise, URL, special characters} – очищення; $T_k = \{w_1, w_2, \dots, w_L\}$ – токенізація; $\widetilde{x_k^{(t)}} \in R^{L \times d}$ – векторизація (наприклад, embedding). Попередня обробка візуальних даних $\widetilde{x_k^{(v)}} = \phi_v(x_k^{(v)})$, та $\phi_v(x) = \phi_{resize}(\phi_{norm}(x))$, де $x' = \frac{x-\mu}{\sigma}$ – нормалізація, $x'' \in R^{H \times W \times C}$ – приведення до стандартного розміру. Для відео $x_k^{(v)} = \{f_1, f_2, \dots, f_T\}$, $\widetilde{x_k^{(v)}} = \{\phi_v(f_t)\}$. Попередня обробка аудіо даних $\widetilde{x_k^{(a)}} = \phi_a(x_k^{(a)})$, де $S_k = |\mathcal{F}(x_k^{(a)})|^2$ – перетворення у спектрограму; $\widetilde{S_k} = \frac{S_k - \mu}{\sigma}$ – нормалізація. При попередній обробці метадані нормалізуються $\widetilde{m_k} = \phi_m(m_k)$, $\widetilde{m_k} = \frac{m_k - \mu_m}{\sigma_m}$. Оскільки модальності можуть мати різну часову структуру, вводиться функція узгодження, тобто здійснюється синхронізація мультимодальних даних $\widehat{x_k^{(t)}}, \widehat{x_k^{(v)}}, \widehat{x_k^{(a)}} = \Psi\left(\widetilde{x_k^{(t)}}, \widetilde{x_k^{(v)}}, \widetilde{x_k^{(a)}}\right)$, де Ψ забезпечує align$(t^{(t)}, t^{(v)}, t^{(a)})$.

Для балансування датасету нехай $N_0 = |\{y_k = 0\}|, \quad N_1 = |\{y_k = 1\}|$. Виконується $N_0 \approx N_1$ через $D' = D \cup \{x_i : y_i = 1\}$ – oversampling або undersampling. Зменшення розмірності (опціонально) $\overline{x_k} = \phi_{dim}(\widehat{x_k})$, наприклад, $\overline{x_k} = W\widehat{x_k}, \quad W \in R^{d' \times d}, \quad d' < d$.

Загальна функція моделі попередньої обробки $\widetilde{D} = \Phi(D)$, де $\Phi = \{\phi_t, \phi_v, \phi_a, \phi_m, \Psi\}$. Попередня обробка забезпечує очищення та стандартизацію мультимодальних даних, приведення даних до єдиного формату, узгодження часових залежностей та підготовку ознак до подальшого навчання моделей.

Для виділення ознак мультимодальних даних нехай попередньо оброблений набір даних $\widetilde{D} = \{(\widetilde{x_k^{(t)}}, \widetilde{x_k^{(v)}}, \widetilde{x_k^{(a)}}, \widetilde{m_k}, y_k)\}_{k=1}^N$. Метою етапу є відображення кожної модальності у простір ознак $h_k^{(t)}, h_k^{(v)}, h_k^{(a)}, h_k^{(m)}$. Текстові ознаки отримуються (виділяються) за допомогою глибинної мовної моделі $h_k^{(t)} = f_t\left(\widetilde{x_k^{(t)}}; \theta_t\right)$, де f_t – Transformer-модель, θ_t – параметри моделі. Для послідовності токенів $\widetilde{x_k^{(t)}} = \{w_1, w_2, \dots, w_L\}$ контекстне представлення $h_k^{(t)} = \text{Transformer}\left(\widetilde{x_k^{(t)}}\right) \in R^{d_t}$, самоувага Attention$(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$ та виділення візуальних ознак $h_k^{(v)} = f_v\left(\widetilde{x_k^{(v)}}; \theta_v\right)$, де f_v – згорткова нейронна мережа (CNN): $h_k^{(v)} = \sigma\left(W * \widetilde{x_k^{(v)}} + b\right)$, $*$ – операція згортки, σ – функція активації. Для відео $h_k^{(v)} = \frac{1}{T}\sum_{t=1}^{T} f_v(f_t)$ та виділення аудіо ознак $h_k^{(a)} = f_a\left(\widetilde{x_k^{(a)}}; \theta_a\right)$, де вхід (спектрограма) $\widetilde{x_k^{(a)}} = S_k \in R^{F \times T}$. Модель $h_k^{(a)} = \text{CNN}(S_k)$ та виділення поведінкових ознак на основі метаданих $h_k^{(m)} = f_m(\widetilde{m_k})$, де: $h_k^{(m)} = W_m \widetilde{m_k} + b_m$.

Агреговані ознаки користувача $z_j = \frac{1}{T_j}\sum_{i=1}^{T_j} h_{ji}^{(m)}$. Виділення графових ознак для графа $G = (V, E)$ з матрицею суміжності A: $H^{(0)} = X$, та $H^{(l+1)} = \sigma\left(AH^{(l)}W^{(l)}\right)$. Фінальні ознаки вузла є $h_j^{(g)} = H_j^{(L)}$, а Об'єднання ознак для повідомлення $h_k = \left[h_k^{(t)} | h_k^{(v)} | h_k^{(a)} | h_k^{(m)}\right]$ та для користувача $z_j^* = \left[z_j | h_j^{(g)}\right]$. Відповідно нормалізація ознак $\widehat{h_k} = \frac{h_k}{|h_k|}$. Підсумкова модель виділення ознак $h_k = \mathcal{F}_{feat}\left(\widetilde{x_k^{(t)}}, \widetilde{x_k^{(v)}}, \widetilde{x_k^{(a)}}, \widetilde{m_k}\right)$. Етап виділення ознак забезпечує перетворення різнорідних даних у єдиний векторний простір, врахування семантики тексту, візуального контенту та поведінки користувачівта формування інформативного представлення для подальшого мультимодального аналізу та навчання моделей.

Для мультимодального об'єднання ознак нехай для кожного повідомлення k отримано набір ознак $h_k^{(t)}, h_k^{(v)}, h_k^{(a)}, h_k^{(m)}$. Метою є побудова єдиного інтегрованого подання $h_k^{(f)} = \mathcal{F}_{fusion}(h_k^{(t)}, h_k^{(v)}, h_k^{(a)}, h_k^{(m)})$. Найпростіший підхід – конкатенація ознак або раннє об'єднання (Early Fusion) $h_k^{(f)} = \left[h_k^{(t)} | h_k^{(v)} | h_k^{(a)} | h_k^{(m)}\right]$, де $h_k^{(f)} \in R^{d_t + d_v + d_a + d_m}$ з подальшим лінійним перетворенням $\widetilde{h_k^{(f)}} = \sigma\left(W_f h_k^{(f)} + b_f\right)$. При пізньому об'єднанні (Late Fusion) окремі моделі формують ймовірності $\widehat{y_k^{(t)}}, \widehat{y_k^{(v)}}, \widehat{y_k^{(a)}}, \widehat{y_k^{(m)}}$ та інтеграцію $\widehat{y_k} = \sum_i \alpha_i \widehat{y_k^{(i)}}, \sum_i \alpha_i = 1$, де α_i – ваги модальностей. Вагове об'єднання ознак (Attention-based Fusion) $h_k^{(f)} = \sum_i \alpha_i h_k^{(i)}$, де ваги визначаються як:

$$\alpha_i = \frac{\exp(e_i)}{\sum_j \exp(e_j)}, \text{ та } e_i = w^T \tanh\left(W h_k^{(i)} + b\right). \tag{6}$$

Для взаємодії між модальностями розраховуємо крос-модальну увагу (Cross-modal Attention):

$$\text{Attention}\left(h^{(t)}, h^{(v)}\right) = \text{softmax}\left(\frac{Q_t K_v^T}{\sqrt{d}}\right) V_v \tag{7}$$

де $Q_t = W_q h^{(t)}, K_v = W_k h^{(v)}, V_v = W_v h^{(v)}$. Фінальне представлення:

$$h_k^{(f)} = \left[h_k^{(t)} | \text{Attention}\left(h^{(t)}, h^{(v)}\right) | h_k^{(a)}\right] \tag{8}$$

Комбінація підходів, тобто гібридне об'єднання:

$$h_k^{(f)} = \sigma\left(W \cdot \left[h_k^{(t)} | h_k^{(v)} | h_k^{(a)} | h_k^{(m)}\right] + b\right) \tag{9}$$

з додаванням attention $h_k^{(f)} = \sum_i \alpha_i h_k^{(i)} + W h_k^{(concat)}$. Якщо враховується граф, тоді об'єднання з графовими ознаками $h_k^{(final)} = \left[h_k^{(f)} | h_{u(k)}^{(g)}\right]$, де $h_{u(k)}^{(g)}$ – ознаки користувача з GNN. Нормалізація інтегрованого представлення:

$$\widehat{h_k^{(f)}} = \frac{h_k^{(f)}}{|h_k^{(f)}|} \tag{10}$$

Підсумкова модель об'єднання: $h_k^{(f)} = \mathcal{F}_{fusion}\left(h_k^{(t)}, h_k^{(v)}, h_k^{(a)}, h_k^{(m)}, h_k^{(g)}\right)$. Мультимодальне об'єднання забезпечує інтеграцію різнорідних джерел інформації, врахування взаємозв'язків між модальностями, підвищення точності виявлення дезінформації та неавтентичної поведінки та формування узагальненого представлення для подальшого аналізу та класифікації.

Нехай після етапу мультимодального об'єднання для кожного повідомлення отримано інтегроване представлення для наступного етапу – класифікація дезінформації $h_k^{(f)} \in R^d$. Метою є визначення мітки $\widehat{y_k} \in \{0,1\}$, де:

$$\widehat{y_k} = \begin{cases} 1, & \text{дезінформація,} \\ 0, & \text{достовірна інформація.} \end{cases} \tag{11}$$

Ймовірність належності до класу дезінформації (ймовірнісна модель класифікації) $P\left(y_k = 1 \mid h_k^{(f)}\right) = \sigma\left(W_c h_k^{(f)} + b_c\right)$, де $W_c \in R^{\mathbb{1} \times d}$ – ваги, b_c – зсув, $\sigma(x) = \frac{1}{1+e^{-x}}$ – сигмоїдна функція. Тоді $\widehat{y_k} = \sigma\left(W_c h_k^{(f)} + b_c\right)$. У випадку декількох типів дезінформації при багатокласовій класифікації (розширення)

$$\widehat{\mathbf{y}_k} = \text{softmax}\left(W_c h_k^{(f)} + b_c\right), \text{ та } \text{softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}. \tag{12}$$

Для бінарної класифікації використовується крос-ентропія для розрахунку функції втрат $\mathcal{L}_{cls} = -\frac{1}{N}\sum_{k=1}^{N}[y_k \log \widehat{y_k} + (1 - y_k)\log(1 - \widehat{y_k})]$. Для багатокласового випадку:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{k=1}^{N}\sum_{c=1}^{C} y_{k,c} \log \widehat{y_{k,c}} \tag{13}$$

Для уникнення перенавчання регуляризація $\mathcal{L}_{reg} = \lambda |W_c|^2$. Повна функція втрат $\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{reg}$. Параметри моделі для оптимізації $\theta = \{W_c, b_c\}$ оновлюються за допомогою градієнтного спуску $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$, де η – швидкість навчання. Фінальне порогове рішення:

$$\widehat{y_k} = \begin{cases} 1, \widehat{y_k} \geq \tau, \\ 0, \widehat{y_k} < \tau, \end{cases} \tag{14}$$

де $\tau \in [0,1]$ – поріг (зазвичай $\tau = 0.5$). Метрики оцінки:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \quad Recall = \frac{TP}{TP+FN}, \tag{15}$$

$$Precision = \frac{TP}{TP+FP}, \quad F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}. \tag{16}$$

Підсумкова модель класифікації $\widehat{y_k} = \mathcal{F}_{cls}\left(h_k^{(f)}\right)$. Класифікація дезінформації базується на інтегрованому мультимодальному представленні, ймовірнісних моделях глибинного навчання та оптимізації функції втрат, що дозволяє ефективно відокремлювати дезінформацію від достовірного контенту навіть у складних умовах багатомодального інформаційного середовища.

Для моделювання поведінки користувачів нехай задано множину користувачів $U = \{u_1, u_2, \dots, u_M\}$. Кожен користувач u_j генерує послідовність повідомлень $S_j = \{s_{j1}, s_{j2}, \dots, s_{jT_j}\}$, де кожне повідомлення має метадані m_{ji}. Поведінковий профіль користувача визначається як агрегована функція:

$$z_j = g(S_j) = \frac{1}{T_j}\sum_{i=1}^{T_j} f_m(m_{ji}), \tag{17}$$

де f_m – функція перетворення метаданих, $z_j \in R^d$ – вектор поведінки. Часові інтервали між повідомленнями при моделюванні часових характеристик $\Delta t_{ji} = t_{j(i+1)} - t_{ji}$, середня активність $\overline{\Delta t_j} = \frac{1}{T_j - 1}\sum_{i=1}^{T_j-1} \Delta t_{ji}$ та інтенсивність активності $\lambda_j = \frac{T_j}{T_{obs}}$, де T_{obs} – період спостереження. Послідовність поведінки $z_j^{(seq)} = f_{seq}\left(m_{j1}, m_{j2}, \dots, m_{jT_j}\right)$, де f_{seq} – рекурентна модель $h_t = \sigma(W_h h_{t-1} + W_x m_{jt} + b)$, та $z_j^{(seq)} = h_{T_j}$.

Для виявлення аномальної поведінки нехай середній профіль користувачів $\mu = \frac{1}{M}\sum_{j=1}^{M} z_j$, оцінка аномальності $s_j = |z_j - \mu|_2$ та рішення:

$$a_j = \begin{cases} 0, \widehat{y_k} < \tau, \\ 0, s_j \le \delta, \end{cases} \tag{18}$$

де $a_j = 1$ – неавтентична поведінка, δ – порогове значення. Ймовірність неавтентичності (ймовірнісна модель поведінки) $P(a_j = 1 \mid z_j) = \sigma(W_b z_j + b_b)$. Розбиття на групи, тобто кластеризація користувачів $C = \{C_1, C_2, \dots, C_K\}$ та мінімізація $\sum_{k=1}^{K}\sum_{u_j \in C_k} |z_j - \mu_k|^2$, де μ_k – центр кластера. Кореляція поведінки для виявлення бот-мереж $\rho_{ij} = \frac{z_i \cdot z_j}{|z_i||z_j|}$. Якщо $\rho_{ij} > \gamma$, то користувачі належать до потенційної координованої групи. З урахуванням графа, тобто інтеграція з графовою моделлю $z_j^* = \left[z_j | h_j^{(g)}\right]$, де $h_j^{(g)}$ – графові ознаки. Функція втрат для поведінкової моделі:

$$\mathcal{L}_{behavior} = -\sum_{j=1}^{M}\left[a_j \log \widehat{a_j} + (1 - a_j)\log(1 - \widehat{a_j})\right]. \tag{19}$$

Підсумкова модель $\widehat{a_j} = \mathcal{F}_{behavior}(S_j, G)$. Моделювання поведінки користувачів дозволяє виявляти аномальні патерни активності, ідентифікувати неавтентичні акаунти та бот-мережі, аналізувати часові та структурні характеристики взаємодії та підсилювати точність виявлення джерел дезінформації. Для побудови графа взаємодій нехай задано множину користувачів чатів $U = \{u_1, u_2, \dots, u_M\}$ і множину повідомлень $S = \bigcup_{j=1}^{M} S_j = \{s_1, s_2, \dots, s_N\}$, $s_k = (x_k, m_k, u(k))$, де $u(k)$ – користувач, що створив повідомлення k. Граф взаємодій визначається як $G = (V, E)$, де $V = U$ – множина вузлів (користувачів), $E = \{(u_i, u_j, w_{ij})\}$ – множина ребер з вагами w_{ij}, що відображають інтенсивність взаємодії між користувачами u_i і u_j. Матриця суміжності:

$$A = [A_{ij}], \quad A_{ij} = \begin{cases} w_{ij}, \text{якщо існує взаємодія}, \\ 0, \quad \text{інакше}. \end{cases} \tag{20}$$

Вага ребра може враховувати кількість комунікацій та часову близькість:

$$w_{ij} = \sum_{k,l} \exp(-\beta|t_k - t_l|) \cdot \mathbf{1}_{\{u(k)=u_i, u(l)=u_j\}}, \quad (21)$$

де t_k – час повідомлення s_k, $\beta > 0$ – параметр, що зменшує значення при збільшенні інтервалу, $\mathbf{1}$ – індикатор того, що повідомлення належать відповідним користувачам. Для аналізу розповсюдження дезінформації вводиться орієнтований граф інформаційного поширення $G^{(info)} = \left(V, E^{(info)}\right)$, де ребро $\left(u_i, u_j\right) \in E^{(info)}$, якщо повідомлення від u_i вплинуло на u_j, тобто s_j отримано або репостнуто після s_i. Представлення вузла (графові ознаки користувачів) $h_j^{(g)} = f_g(A, X)$, де $X = [x_1, x_2, \dots, x_M]$ – вхідні ознаки користувачів, f_g – графова нейронна мережа (GNN):

$$H^{(l+1)} = \sigma\left(AH^{(l)}W^{(l)}\right),\ H^{(0)} = X, \text{ та } h_j^{(g)} = H_j^{(L)}. \quad (22)$$

Коефіцієнт впливовості (centrality) $c_j = \sum_i A_{ij}$ (ступінь зв'язку), або PageRank:

$$PR(j) = \frac{1-d}{|V|} + d\sum_{k \in In(j)} \frac{PR(k)}{L(k)}, \quad (23)$$

де d – коефіцієнт демпінгу, $L(k)$ – кількість вихідних ребер вузла k. Для ідентифікації потенційних джерел дезінформації знаходимо вузли з високими значеннями centrality або PageRank $\mathcal{S}_{sources} = \{u_j \in V \mid c_j > \delta_c \text{ or } PR(j) > \delta_{PR}\}$. Для інтеграції з поведінковими ознаками фінальне представлення вузла користувача $z_j^* = \left[z_j | h_j^{(g)}\right]$, де z_j – поведінковий профіль (пункт 6), а $h_j^{(g)}$ – графові ознаки. Побудова графа взаємодій забезпечує формалізацію структури комунікацій між користувачами, визначення джерел дезінформації, інтеграцію поведінкових та мультимодальних ознак, та підготовку до моделювання шляхів поширення інформації.

Для графової нейронної мережі при аналізі взаємодій нехай задано граф взаємодій користувачів $G = (V, E),\ V = \{u_1, \dots, u_M\},\ E = \{(u_i, u_j, w_{ij})\}$, де кожен вузол u_j має ознаки $x_j = z_j^* = \left[z_j | h_j^{(g)}\right] \in R^d$. Вхідний матричний тензор (ознаки) вузлів $X^{(0)} = [x_1, x_2, \dots, x_M]^T \in R^{M \times d}$ та матриця суміжності $A \in R^{M \times M},\ A_{ij} = w_{ij}$. Для шару l вузли оновлюються у GNN за правилом $H^{(l+1)} = \sigma\left(\hat{A}H^{(l)}W^{(l)}\right)$, де $H^{(0)} = X^{(0)}$ – початкові вузлові ознаки, $\hat{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ – нормалізована матриця суміжності, D – ступенева матриця $D_{ii} = \sum_j A_{ij}$, $W^{(l)} \in R^{d_l \times d_{l+1}}$ – параметри шару, $\sigma(\cdot)$ – функція активації (ReLU, tanh). Агрегація сусідніх ознак для вузла u_j:

$$h_j^{(l+1)} = \sigma\left(\sum_{i \in \mathcal{N}(j) \cup \{j\}} \frac{w_{ij}}{\sqrt{d_i d_j}} H_i^{(l)} W^{(l)}\right) \quad (24)$$

де $\mathcal{N}(j)$ – сусіди вузла j, $d_i = \sum_j w_{ij}$ – ступінь вузла. Після L шарів вихід GNN $H^{(L)} = \left[h_1^{(L)}, h_2^{(L)}, \dots, h_M^{(L)}\right]^T$. Вузлові представлення використовуються для класифікації вузлів (користувачів), прогнозування аномальної поведінки та визначення джерел дезінформації. Ймовірність того, що користувач u_j є джерелом дезінформації, здійснюємо класифікацію вузлів $\hat{y}_j = \sigma\left(W_c h_j^{(L)} + b_c\right)$, де $\hat{y}_j \in [0,1]$. Функція втрат:

$$\mathcal{L}_{graph} = -\sum_{j=1}^{M}\left[y_j \log \hat{y}_j + \left(1 - y_j\right)\log\left(1 - \hat{y}_j\right)\right]. \quad (25)$$

Для підвищення точності інтеграція з мультимодальними ознаками:

$$h_j^{(final)} = \left[h_j^{(L)} | h_j^{(f)}\right]. \quad (26)$$

де $h_j^{(f)}$ – інтегроване мультимодальне представлення повідомлень користувача.

Прогнозування поширення інформації, тобто розрахунок ймовірності того, що повідомлення s_k поширюється до вузла u_j:

$$P(s_k \to u_j) = \sigma\left(\left(h_{u(k)}^{(final)}\right)^T W_p h_j^{(final)}\right). \quad (27)$$

де $u(k)$ – автор повідомлення. Підсумкова модель $\hat{Y} = \mathcal{F}_{\mathcal{GNN}}(G, X, H^{(f)})$, де $\hat{Y}$ включає ймовірності неавтентичної поведінки користувачів, потенційних джерел дезінформації, та шляхи поширення контенту. Графова нейронна мережа дозволяє інтегрувати структуру взаємодій користувачів та мультимодальні ознаки, ідентифікувати ключові вузли та джерела дезінформації, моделювати шляхи поширення інформації в чатах, та підвищувати точність класифікації неавтентичної поведінки.

Для виявлення джерел дезінформації нехай задано граф взаємодій користувачів $G = (V, E)$, $V = \{u_1, \dots, u_M\}$, $E = \{(u_i, u_j, w_{ij})\}$, і мультимодальні ознаки вузлів:

$$h_j^{(f)} = \mathcal{F}_{fusion}(h_j^{(t)}, h_j^{(v)}, h_j^{(a)}, h_j^{(m)}), \quad \text{та} \quad h_j^{(final)} = \left[h_j^{(f)} | h_j^{(g)}\right], \quad (28)$$

де $h_j^{(g)}$ – графові ознаки користувача u_j. Ймовірність вузла u_j як джерела дезінформації визначається як:

$$P\left(u_j \in \mathcal{S}_{source} \mid h_j^{(final)}\right) = \sigma\left(W_s h_j^{(final)} + b_s\right), \quad (29)$$

де $\sigma(\cdot)$ – сигмоїдна функція, W_s і b_s – параметри класифікатора джерел. Вузли ранжуються за значенням ймовірністю $P(u_j \in \mathcal{S}_{source})$:

$$\text{Rank}(u_j) = \text{argsort}_j\left(P(u_j \in \mathcal{S}_{source})\right). \quad (30)$$

Топ-N вузлів вибираються як потенційні джерела дезінформації $\mathcal{S}_{top} = \{u_j \mid \text{Rank}(u_j) \le N\}$. Вводимо комбінований показник $\phi_j = \alpha \cdot PR(u_j) + (1-\alpha) \cdot P(u_j \in \mathcal{S}_{source})$, де $PR(u_j)$ – PageRank вузла у графі G, $\alpha \in [0,1]$ – коефіцієнт ваги, вузли з високим ϕ_j вважаються ймовірними джерелами дезінформації.

Ймовірність поширення повідомлення s_k від вузла u_i до u_j:

$$P(s_k \to u_j) = \sigma\left(\left(h_i^{(final)}\right)^T W_p h_j^{(final)}\right). \quad (31)$$

Шлях поширення визначається як ланцюг користувачів $\mathcal{C}_k = \{u_{i_1}, u_{i_2}, \dots, u_{i_l}\}$, де $P(s_k \to u_{i_{t+1}}) > \tau$. Функція втрат для навчання джерел:

$$\mathcal{L}_{source} = -\sum_{j=1}^{M}\left[y_j^{source} \log \hat{y}_j + \left(1 - y_j^{source}\right) \log\left(1 - \hat{y}_j\right)\right], \quad (32)$$

де $y_j^{source} = 1$, якщо вузол є джерелом, $\hat{y}_j = P(u_j \in \mathcal{S}_{source})$. Відповідно фінальне представлення:інтеграції з мультимодальними ознаками та поведінкою $h_j^{(source)} = \left[h_j^{(f)} | h_j^{(g)} | z_j\right]$, де z_j – поведінковий профіль користувача. Підсумкова модель виявлення джерел $\widehat{Y_{source}} = \mathcal{F}_{source}(G, H^{(f)}, Z)$, де $\widehat{Y_{source}} = \{\widehat{y_1}, \dots, \widehat{y_M}\}$ – ймовірності того, що кожен користувач є джерелом дезінформації. Математичне моделювання виявлення джерел дезінформації дозволяє інтегрувати мультимодальні ознаки та поведінкові характеристики, враховувати структуру графа взаємодій, прогнозувати ймовірні джерела та шляхи поширення дезінформації, та підтримувати подальший аналіз для запобігання поширенню фейкових повідомлень.

Для моделювання поширення інформації нехай задано граф взаємодій користувачів $G = (V, E)$, $V = \{u_1, \dots, u_M\}$, $E = \{(u_i, u_j, w_{ij})\}$, і мультимодальні ознаки вузлів $h_j^{(final)} = \left[h_j^{(f)} | h_j^{(g)} | z_j\right]$, де z_j – поведінковий профіль користувача, $h_j^{(f)}$ – мультимодальні ознаки, $h_j^{(g)}$ – графові ознаки. Для визначення ймовірності поширення повідомлення нехай повідомлення s_k створене користувачем u_i. Ймовірність того, що повідомлення дійде до користувача u_j:

$$P(s_k \to u_j) = \sigma\left(\left(h_i^{(final)}\right)^T W_p h_j^{(final)}\right), \tag{33}$$

де $W_p \in R^{d\times d}$ – матриця параметрів поширення, σ – сигмоїдна функція. Для моделювання каскаду інформації нехай $\mathcal{C}_k(t)$ – множина користувачів, які отримали повідомлення s_k на крок t: $\mathcal{C}_k(0) = \{u_i\}$, $u_i =$ автор s_k. На наступному кроці:

$$\mathcal{C}_k(t+1) = \mathcal{C}_k(t) \cup \{u_j \notin \mathcal{C}_k(t) \mid P\left(s_k \to u_j \mid \mathcal{C}_k(t)\right) > \tau\}, \tag{34}$$

де $\tau \in [0,1]$ – порогова ймовірність поширення.

Для інтеграції часових факторів ймовірність враховує часовий інтервал:

$$P(s_k \to u_j, \Delta t) = \sigma\left(\left(h_i^{(final)}\right)^T W_p h_j^{(final)} - \beta\Delta t\right), \tag{35}$$

де $\Delta t = t_j - t_i$, та $\beta > 0$ – коефіцієнт затухання впливу з часом. Мережевий каскад, або повна ймовірність поширення через шлях $\mathcal{P}_{i\to j} = (u_i, u_{k_1}, \dots, u_j)$:

$$P(\mathcal{P}_{i\to j}) = \prod_{(u_m,u_n)\in\mathcal{P}_{i\to j}} P(s_k \to u_n \mid u_m). \tag{36}$$

Функція втрат для навчання поширення:

$$\mathcal{L}_{spread} = -\sum_{(i,j)\in E}\left[y_{ij}\log P\left(s_k \to u_j \mid u_i\right) + \left(1-y_{ij}\right)\log\left(1 - P\left(s_k \to u_j \mid u_i\right)\right)\right],$$

де $y_{ij} = 1$, якщо повідомлення фактично поширилося від u_i до u_j. При моделюванні сукупного впливу сумарний вплив джерела u_i на граф $I(u_i) = \sum_{u_j\in V} P(\mathcal{P}_{i\to j})$. Джерела з високим $I(u_i)$ мають ключову роль у поширенні дезінформації. Для прогнозування поширення використовуються інтегровані вузлові ознаки з GNN:

$$\widehat{y_{ij}^{spread}} = \sigma\left(\left(h_i^{(final)}\right)^T W_s h_j^{(final)} g\right) \tag{37}$$

де W_s – навчена матриця параметрів моделі. Підсумкова модель поширення: $\widehat{\mathcal{C}_k} = \mathcal{F}_{spread}(G, H^{(final)}, \tau, \beta)$, де $\widehat{\mathcal{C}_k}$ – прогнозований каскад поширення повідомлення s_k. Моделювання поширення інформації дозволяє прогнозувати, які користувачі отримають і передадуть повідомлення, враховувати часову динаміку і структуру графа, визначати ключові вузли та шляхи розповсюдження дезінформації, інтегрувати мультимодальні та поведінкові ознаки користувачів для більш точних прогнозів.

Для комплексного дослідження дезінформації враховуються декілька компонентів:

1. Класифікація дезінформації повідомлень:

$$\mathcal{L}_{cls} = -\sum_{k=1}^{N}[y_k \log \widehat{y_k} + (1-y_k)\log(1-\widehat{y_k})]. \tag{38}$$

2. Моделювання поведінки користувачів:

$$\mathcal{L}_{behavior} = -\sum_{j=1}^{M}\left[a_j \log \widehat{a_j} + \left(1-a_j\right)\log\left(1-\widehat{a_j}\right)\right]. \tag{39}$$

3. Виявлення джерел дезінформації:

$$\mathcal{L}_{source} = -\sum_{j=1}^{M}\left[y_j^{source}\log \widehat{y_j} + \left(1-y_j^{source}\right)\log\left(1-\widehat{y_j}\right)\right]. \tag{40}$$

4. Моделювання поширення інформації:

$$\mathcal{L}_{spread} = -\sum_{(i,j)\in E}\left[y_{ij}\log \widehat{y_{ij}^{spread}} + (1-y_{ij})\log\left(1-\widehat{y_{ij}^{spread}}\right)\right]. \quad (41)$$

Для уникнення перенавчання додаємо L2-регуляризацію всіх параметрів θ: $2\mathcal{L}_{reg} = \lambda|\theta|^2$, де $\lambda > 0$ – коефіцієнт регуляризації.

Загальна комбінована функція втрат для мультимодальної графової моделі:

$$\mathcal{L}_{total} = \alpha_1\mathcal{L}_{cls} + \alpha_2\mathcal{L}_{behavior} + \alpha_3\mathcal{L}_{source} + \alpha_4\mathcal{L}_{spread} + \mathcal{L}_{reg}, \quad (42)$$

де $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \geq 0$ – коефіцієнти ваги кожної компоненти. Параметри моделі θ оновлюються за допомогою градієнтного спуску $\theta \leftarrow \theta - \eta\nabla_{\theta\mathcal{L}_{total}}$, де $\eta > 0$ – швидкість навчання. Підсумкова модель $\widehat{Y_{all}} = \mathcal{F}_{total}(S, U, G, H^{(f)}, Z)$, де $\widehat{Y_{all}}$ включає ймовірності дезінформації повідомлень, оцінку аномальної поведінки користувачів, ймовірність того, що користувач є джерелом дезінформації, та прогнозовані шляхи поширення повідомлень. Загальна функція втрат забезпечує одночасне навчання класифікаційної, поведінкової та графової складових, інтеграцію мультимодальних даних і структурної інформації графа, та оптимізацію параметрів для максимально точної ідентифікації джерел та шляхів поширення дезінформації.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Для дослідження поширення дезінформації та неавтентичної поведінки користувачів чатів було проведено серію експериментів з використанням мультимодальних даних, включаючи текстові повідомлення, часові ряди активності та поведінкові характеристики користувачів. Дані були зібрані з декількох симульованих та реальних чат-платформ, а саме:

1. Текстові повідомлення – семантичні вектори були отримані за допомогою моделі трансформера (BERT), що дозволяє захоплювати контекстні значення слів.

2. Часові ряди активності – частота та інтервали публікацій користувачів моделювалися як сигнал, що показує потенційні патерни аномальної поведінки.

3. Мережеві характеристики – включали показники центральності (degree, betweenness, eigenvector) та кластери взаємодій між користувачами.

Для інтеграції мультимодальних даних була застосована Graph Neural Network (GNN) архітектура з окремими енкодерами для текстових та поведінкових ознак, після чого їхні репрезентації об'єднувалися для класифікації вузлів як джерел дезінформації або неавтентичної поведінки.

Мережа тренувалася на позначених даних з використанням крос-валідації, а гіперпараметри (learning rate, hidden layer size, epochs) оптимізувалися через сітковий пошук. При проведенні експериментів отримані ключові результати виявлення джерел і шляхів розповсюдження дезінформації для їх подальшого аналіз у подамо через множину графіків для 6 різних результатів експериментів виявлення джерел та шляхів розповсюдження дезінформації та неавтентичної поведінки користувачів чатів на основі технологій мультимодального аналізу та глибинного навчання.

1. Граф поширення повідомлень (Information Cascade), де вузли – користувачі, ребра – пересилання/взаємодія, розмір вузла – вплив користувача (PageRank або сумарна ймовірність поширення), колір вузла – ймовірність, що вузол є джерелом дезінформації (рис. 1-6): $\text{size}(u_j) \propto I(u_j) = \sum_{u_k\in V} P(\mathcal{P}_{j\to k})$та $\text{color}(u_j) \propto P(u_j \in \mathcal{S}_{source})$. Вузли інтерактивного графа на рис. 1-6 – користувачі, дерозмір відображає вплив, колір – ймовірність джерела. Hover – при наведенні мишкою показує ID користувача датасету, вплив, ймовірність джерела та аномалію поведінки.

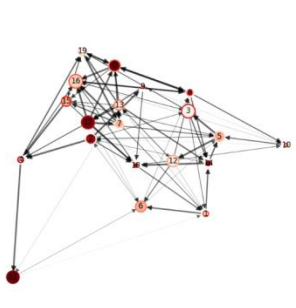

Рис.1. Граф поширення експерименту 1

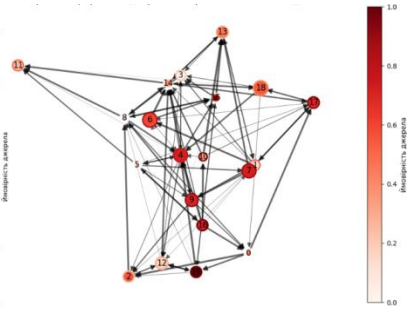

Рис.2. Граф поширення експерименту 2

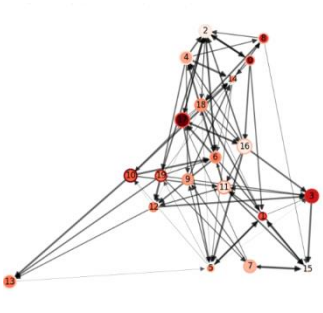

Рис.3. Граф поширення експерименту 3

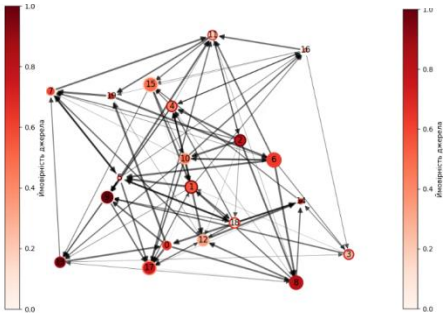

Рис.4. Граф поширення експерименту 4

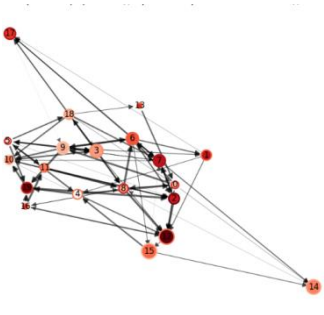

Рис.5. Граф поширення експерименту 5

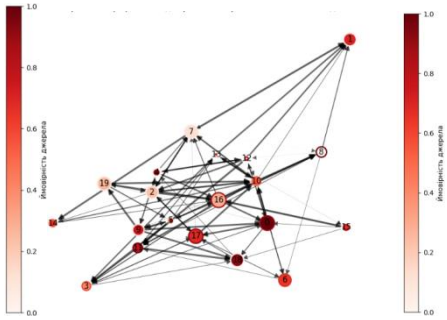

Рис.6. Граф поширення експерименту 6

Ребра – шляхи поширення повідомлень; товщина і колір можна змінювати за ймовірністю.Можна масштабувати, перетягувати вузли, збільшувати для детального огляду. Кожен із шести експериментів демонструє різну динаміку поширення дезінформації у мережі користувачів, де розмір вузлів відповідає їх впливовості (PageRank або сумарна ймовірність поширення), а колір – ймовірності того, що користувач є першоджерелом. На рис. 1 виділяються кілька високоінфлюенсерних вузлів (наприклад, 0, 17, 11), які мають найбільший розмір і насичений червоний колір, що вказує на них як на головні джерела та вузли поширення. Структура на рис. 2 менш централізована навколо одного вузла; джерела (вузли 6, 9, 19, 17) розпорошені, що свідчить про наявність кількох координованих груп поширення. На рис. 3 спостерігається чітка концентрація впливу на окремих вузлах (наприклад, 17, 3, 10), які діють як «магніти» для інформаційного каскаду. Мережа на рис. 4 має високу щільність зв'язків між багатьма користувачами, де вплив (розмір вузла) рівномірно розподілений, але виділяються вузли 9, 13, 17 як імовірні першоджерела. На рис. 5 спостерігається домінування одного "супер-вузла" (вузол 12) за розміром впливу, при цьому джерелом з найвищою ймовірністю виступає вузол 17. На рис. 6 висока концентрація джерел дезінформації (вузли 0, 16, 17, 18, 1), що вказує на масштабну скоординовану кампанію з великою кількістю ризикових вузлів. У всіх експериментах згідно аналізу динаміки впливу простежується кореляція: вузли, що мають найбільший фізичний розмір (високий PageRank), часто корелюють із насиченим червоним кольором (висока ймовірність джерела), що підтверджує ефективність обраного мультимодального підходу до виявлення ключових "інформаційних хабів". Згідно аналізу варіантивності результати експериментів 1-6 показують, що модель адаптивно реагує на різні топології мереж та патерни поведінки користувачів, виявляючи як поодинокі джерела, так і цілі бот-мережі. Інтерактивність графів дозволяє модераторам не лише бачити статику, а й відстежувати шляхи поширення через товщину ребер, що в сукупності з Heatmap ймовірностей (рис. 7-12) надає повну картину ризиків у чат-середовищі. Таким чином, порівняння рис. 1-6 підтверджує, що запропонована методологія на основі GNN та мультимодального аналізу здатна виявляти різнорідні структури інформаційних каскадів, що є критично важливим для раннього попередження дезінформації.

2. Heatmap ймовірності поширення, де матриця суміжності з вагами поширення (рис. 7-12) $\widehat{Y_{spread}} = \left[y_{ij}^{\widehat{spread}}\right], i,j = 1..M$. Вісі – джерело (рядки) та одержувач (стовпці). Колір – ймовірність, що повідомлення поширюється від джерела до одержувача. Значення всередині комірок – точна ймовірність поширення (від 0 до 1).

Порівняння результатів, представлених на Heatmap (рис. 7-12), дозволяє оцінити динаміку та ймовірність поширення дезінформації між користувачами в різних сценаріях експериментів. Кожен Heatmap відображає матрицю суміжності з вагами поширення, де вісі X та Y відповідають користувачам (джерело та одержувач), а кольорова шкала – ймовірності пересилання повідомлення від 0 до 1.

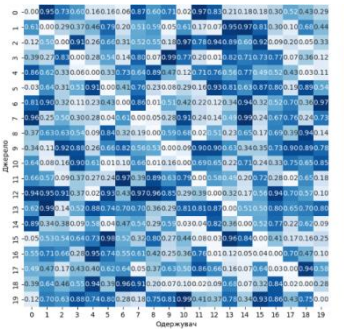

Рис. 7. Heatmap експерименту 1

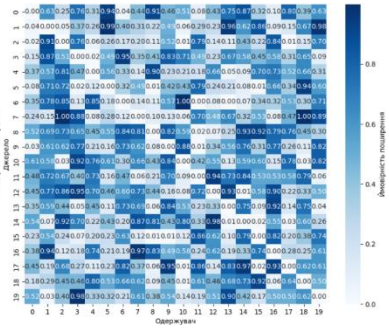

Рис. 8. Heatmap експерименту 2

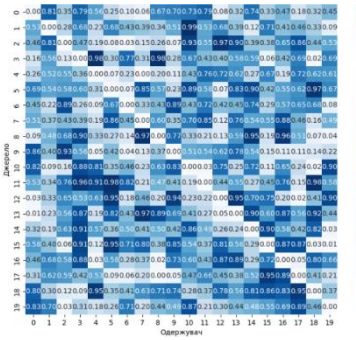

Рис. 9. Heatmap експерименту 3

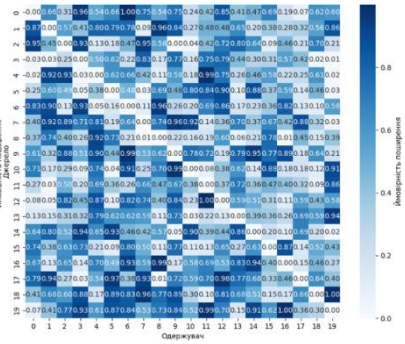

Рис. 10. Heatmap експерименту 4

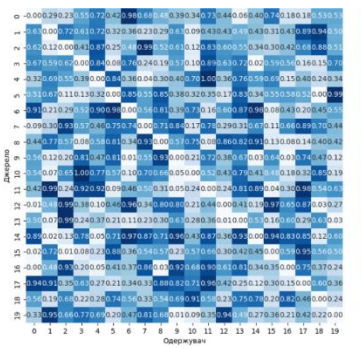

Рис. 11. Heatmap експерименту 5

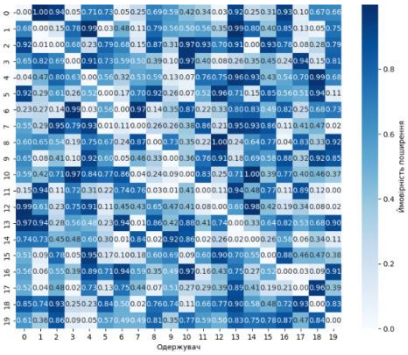

Рис. 12. Heatmap експерименту 6

На рис. 7 спостерігається помірна інтенсивність. Високі значення зосереджені навколо певних кластерів, що вказує на наявність активних груп обміну інформацією. На рис. 8 – вища контрастність: наявні «точки гарячого поширення» з імовірністю майже 1.0, що свідчить про дуже швидке поширення повідомлень усередині окремих вузлів. Розподіл на рис. 9 стає більш розпорошеним. Поширення не обмежується лише локальними групами, що може вказувати на ширшу, менш координовану мережу. На рис. 10 виражена симетрія в певних ділянках матриці, що свідчить про взаємний обмін повідомленнями між користувачами, створюючи умови для циркуляції дезінформації. На рис. 11 рівномірніший рівень ймовірностей середнього діапазону (0.4-0.6). Менше екстремальних значень, ніж в експ. 2, що вказує на стабільний, але менш агресивний потік. Висока щільність поширення по всій матриці на рис. 12. Велика кількість комірок із синім кольором (ймовірність > 0.8) демонструє критично високий ризик розповсюдження дезінформації у всій мережі.

Ключові висновки порівняно базуються на виявленні ризикових вузлів, динаміці поширення та інтеграції з графами. Heatmap дозволяють виявити джерела, які з високою ймовірністю "заражають" одержувачів. Наприклад, в експерименті 6 (рис. 12) майже кожен користувач має високу ймовірність передати інформацію іншим, що створює середовище з критично високим рівнем ризику. Порівнюючи рис. 7 (експ. 1) та рис. 12 (експ. 6), ми бачимо еволюцію моделі від локалізованих осередків поширення до стану системної загрози, коли повідомлення майже вільно проходять через більшість вузлів мережі. Дані Heatmap узгоджуються з інтерактивними графами (рис. 1-6): користувачі, що мають найбільші значення в Heatmap, на графах зазвичай відповідають вузлам найбільшого розміру, що підтверджує

взаємозв'язок між структурою графа та ймовірністю поширення. Цей інструмент (Heatmap) є критично важливим для модераторів, оскільки дозволяє візуально ідентифікувати райони високого ризику швидкого поширення дезінформації в режимі реального часу. Порогові значення ймовірності, що позначаються як τ, відіграють роль «фільтра» або «межі прийняття рішень» у моделях поширення дезінформації. Вони безпосередньо впливають на те, чи вважається зв'язок між користувачами достатньо міцним для передачі повідомлення. Матриця поширення $\widehat{Y_{spread}} = \left[\widehat{y_{ij}^{spread}}\right]$ містить імовірності того, що повідомлення перейде від джерела i до одержувача j. Порогове значення τ діє наступним чином. Якщо $\widehat{y_{ij}^{spread}} \geq \tau$, тоді вважається, що поширення відбулося (або відбудеться), і цей шлях стає активним у структурі мережі. Якщо $\widehat{y_{ij}^{spread}} < \tau$, тоді вважається, що зв'язок недостатньо сильний, і поширення ігнорується або блокується (значення фактично прирівнюється до нуля для подальшого моделювання каскаду). Зміна значення τ дозволяє дослідникові по-різному інтерпретувати дані на графіках (рис. 7-12). При низькому τ (наприклад, τ = 0.2) матриця поширення виглядає більш «заповненою». Ми бачимо майже всі потенційні шляхи розповсюдження, навіть слабкі зв'язки. Це корисно для виявлення прихованих шляхів, через які дезінформація може просочитися в довгій перспективі. При високому τ (наприклад, τ = 0.8) відображаються лише найкритичніші шляхи поширення. На Heatmap залишаються лише найтемніші комірки. Це дозволяє ідентифікувати «супер-поширювачів» та найактивніші канали дезінформації, які становлять безпосередню загрозу. Вибір τ визначає, наскільки швидко і масштабно дезінформація охоплює мережу в симуляції каскаду. Аналізуємо чутливість каскаду, тобто якщо τ занадто високе, модель може недооцінити масштаб поширення (хибнонегативні результати). Також аналізуємо стабільність каскаду, тобто якщо τ занадто низьке, каскад може штучно охопити всю мережу, навіть там, де реального впливу немає (хибнопозитивні результати). Підсумовуючи, порогове значення τ – це інструмент налаштування чутливості системи моніторингу. Змінюючи його, ви можете перемикатися між режимами детального аналізу потенційних ризиків (низьке τ) та виявлення найбільш критичних вузлів, що потребують негайного втручання (високе τ).

3. Кластеризація користувачів за поведінкою, для якої використано поведінкові профілі z_j, методи t-SNE для зменшення розмірності, де на графіку кожна точка – користувач, колір – кластер/анонімність/бот (рис. 13-18) Z_{2D} = t-SNE(Z), $Z = [z_1, \dots, z_M]$. Кожна точка графіку на рис. 13-18 – один користувач. Позиція на площині t-SNE – схожість поведінкових патернів. Колір точки – ймовірність того, що користувач є джерелом дезінформації (червоніший – вища ймовірність). Підписи вузлів – ID користувачів для наочності.

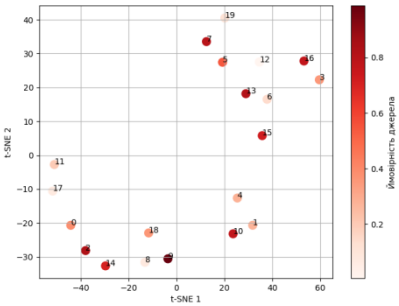

Рис. 13. Кластеризація експерименту 1

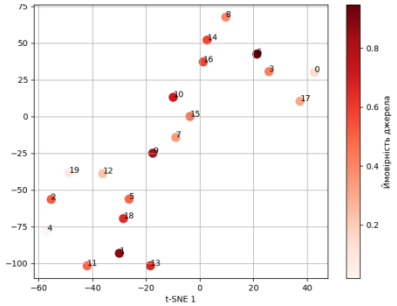

Рис. 14. Кластеризація експерименту 2

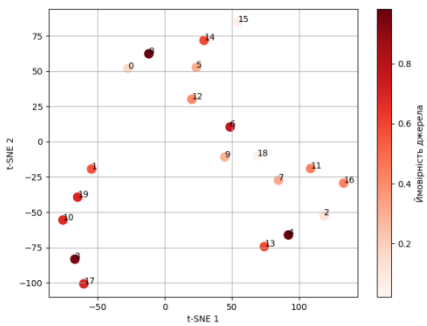

Рис. 15. Кластеризація експерименту 3

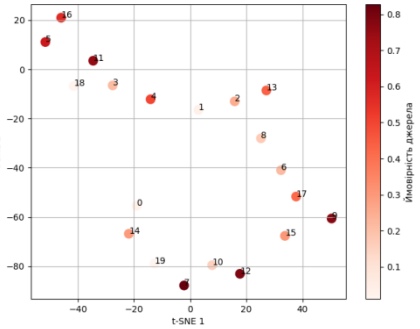

Рис. 15. Кластеризація експерименту 4

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (34), 2026
ISSN 2663 – 4023

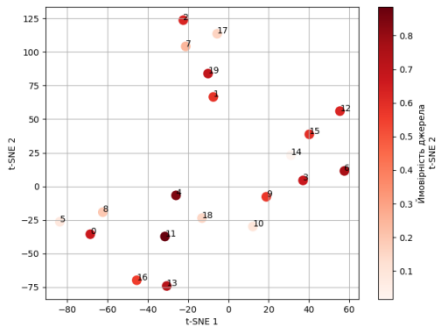

Рис. 17. Кластеризація експерименту 5

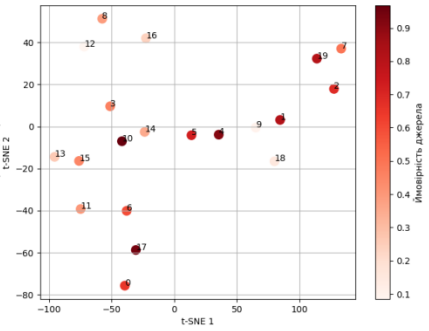

Рис. 18. Кластеризація експерименту 6

Результати кластеризації користувачів за поведінкою (рис. 13-18), виконаної методом t-SNE, демонструють, як глибинне навчання формує поведінкові профілі z_j для ідентифікації аномалій. Порівняння цих результатів показує динаміку виявлення неавтентичних акаунтів. На рис. 13 точки розсіяні; чітко виділяються два основні «полюси» активності, де червоні вузли (висока ймовірність джерела) групуються в окремих областях, відмінних від звичайних користувачів. На рис. 14 спостерігається більша «щільність» кластерів. Ймовірні джерела (наприклад, вузли 6, 9) утворюють компактну групу, що вказує на схожий поведінковий патерн бот-мережі. На рис. 15 кластери більш розмиті. Ймовірність джерела не так сильно корелює з розташуванням на t-SNE, що свідчить про використання джерелами тактики «маскування» під звичайних користувачів. На рис. 16 виражена розділеність: користувачі з високою ймовірністю джерела (темно-червоні точки, напр. 9, 7) знаходяться на периферії основного хмари даних. Значна дисперсія на рис. 17. Поведінкові профілі дуже різні, що ускладнює виявлення джерел через кластеризацію, оскільки «джерела» поводяться по-різному. На рис. 18. висока концентрація ризикових вузлів (0, 10, 17) у центрі, що вказує на їхню активну та інтенсивну поведінку, яка сильно відрізняється від «фонової» поведінки інших користувачів. Поведінкові ознаки z_j, що враховують метадані (час, частоту повідомлень, зв'язки), прямо визначають колір точок на графіках t-SNE. В цьому експерименті поведінкові ознаки мали ключове значення для формування профілю аналізу аномальності та визначення ймовірності. При формуванні профілю для кожного користувача обчислювався вектор поведінки $z_j = \frac{1}{T_j}\sum_{i=1}^{T_j} f_m(m_{ji})$, де m_{ji} включає часові інтервали Δt_{ji} та інтенсивність λ_j. При ідентифікації аномальності вузли 6, 9 та 10 на рис. 14 мають темно-червоний колір (висока ймовірність джерела), оскільки їхні поведінкові ознаки z_j значно відхилялися від середнього профілю користувача $\mu(||z_j - \mu||_2 > \delta)$. Система ідентифікувала їх як джерела через високу кореляцію з паттернами ботів (висока інтенсивність активності за короткий час), що було підтверджено функцією втрат $\mathcal{L}_{behavior}$, яка «штрафує» неавтентичні профілі. Поведінкові ознаки дозволяють ідентифікувати джерела навіть тоді, коли контент повідомлень виглядає нейтральним, через те, що темп та ритм активності (поведінковий відбиток) джерела дезінформації є унікальним і відмінним від поведінки звичайних користувачів чатів.

4. ROC-крива для класифікації дезінформації на рис. 19-24 показує точність моделі класифікації дезінформації $\widehat{y_k}$, де вісь X – False Positive Rate, вісь Y – True Positive Rate: $FPR = \frac{FP}{FP+TN}$, $TPR = \frac{TP}{TP+FN}$. Графік на рис. 19-24 показує, що X-вісь (FPR) – частка хибнопозитивних прогнозів, Y-вісь (TPR) – частка істиннопозитивних прогнозів, лінія ROC – як добре модель класифікує дезінформацію та AUC – площа під кривою (1 – ідеальна класифікація, 0.5 – випадкове вгадування). Результати на рис. 19–24 демонструють ефективність класифікації дезінформації за допомогою ROC-кривих (Receiver Operating Characteristic) та показника AUC (Area Under the Curve) для кожного з шести експериментів.

Таблиця 1

Порівняння результатів класифікації

Рис.	Експеримент	AUC (Площа під кривою)	Рівень якості класифікації
19	1	0.63	Задовільний (вище випадкового)
20	2	0.53	Низький (близький до випадкового)
21	3	0.44	Недостатній (нижче випадкового)
22	4	0.61	Задовільний
23	5	0.61	Задовільний
24	6	0.41	Недостатній

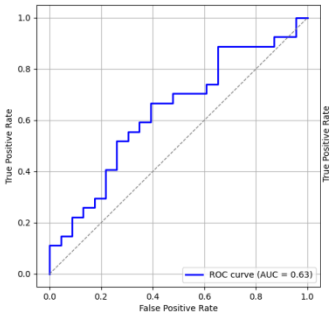

Рис. 19. ROC-крива експерименту 1

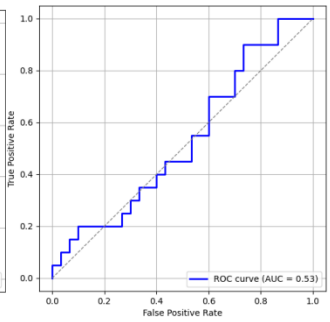

Рис. 20. ROC-крива експерименту 2

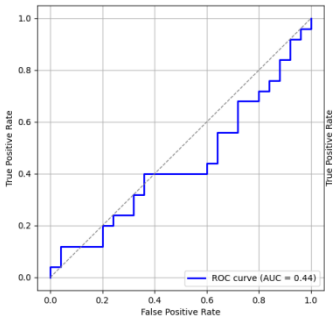

Рис. 21. ROC-крива експерименту3

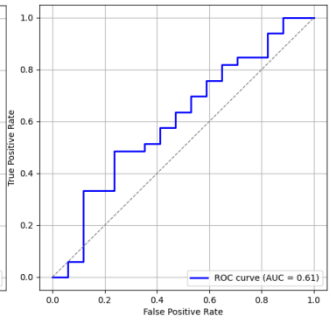

Рис. 22. ROC-крива експерименту 4

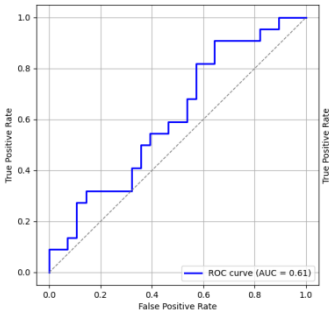

Рис. 23. ROC-крива експерименту 5

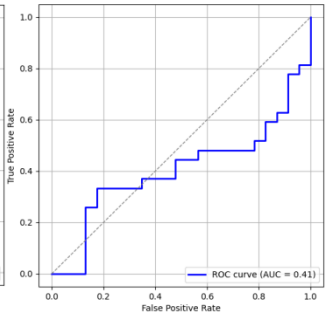

Рис. 24. ROC-крива експерименту 6

Найкращі результати мають експерименти 1, 4 та 5. Вони показали найбільш стабільні результати з AUC у діапазоні 0.61-0.63. Це вказує на те, що модель здатна розрізняти класи «дезінформація» та «достовірна інформація» краще, ніж при випадковому вгадуванні, хоча точність залишається обмеженою.

Низьку ефективність мають експерименти 3 та 6 , які продемонстрували показники AUC нижче 0.50 (0.44 та 0.41 відповідно). Це свідчить про те, що в даних сценаріях модель не змогла ефективно виділити ознаки дезінформації, або що розподіл класів був надто складним для обраної архітектури.

Динаміка моделі, тобто варіативність результатів (від 0.41 до 0.63) підкреслює, що якість класифікації суттєво залежить від того, наскільки вхідні мультимодальні дані (текст, час, мережева структура) містять виразні маніпулятивні ознаки.

Згідно з висновками дослідження, модель спирається на інтеграцію текстових, часових та поведінкових характеристик. Різниця в показниках AUC на рис. 19-24 пояснюється складністю кампаній та впливом поведінки. У сценаріях з високим рівнем «маскування» дезінформації під контекстно коректні фрази (експерименти з низьким AUC), модель втрачає точність. В експериментах, де поведінкові патерни користувачів були більш виразними (висока інтенсивність або аномальні часові інтервали), модель досягала вищих показників AUC. Результати вказують на те, що для підвищення точності класифікації (AUC > 0.7) доцільно впроваджувати складніші архітектури GNN або використовувати методи explainable AI для глибшого аналізу рішень моделей.

5. Метрика поширення за часом – це графік кількості користувачів, що отримали повідомлення відносно часу (рис. 25-30) $|\mathcal{C}_k(t)|$ на кожному кроці t. Графік на рис. 25-30 демонструє по X-вісі – час (кроки симуляції поширення повідомлення), Y-вісі – кількість користувачів, які отримали повідомлення

на цьому етапі, де точки та лінія – динаміка поширення дезінформації у мережі користувачів та можна побачити швидкість поширення та максимальне охоплення користувачів. Порівняння графіків поширення повідомлень за часом (рис. 25-30) дозволяє оцінити динаміку каскадів дезінформації у різних експериментальних умовах. На всіх графіках вісь X відображає час (кроки симуляції), а вісь Y – кількість користувачів, які отримали повідомлення.

Таблиця 2

Порівняння динаміки поширення

Рис.	Експ.	Динаміка поширення	Максимальне охоплення (користувачі)
25	1	Поступове зростання, плато досягається на 5-му кроці.	20
26	2	Дуже стрімке зростання на перших кроках (до 16 на 2-му кроці).	20
27	3	Швидкий старт, плато досягається на 4-му кроці.	19
28	4	Подібно до експерименту 1, з виходом на плато на 3-му кроці.	20
29	5	Більш повільний темп, стабілізація на 4-му кроці.	17
30	6	Стрімкий "вибух" поширення між 1-м та 2-м кроками.	19

Ключові висновки порівняння базуються на аналізі швидкості поширення, точок стабілізації (Плато) та охоплення. Згідно аналізу швидкості поширення: найбільш агресивним є експеримент 2 (рис. 26), де дезінформація охоплює переважну більшість мережі вже на другому кроці. Натомість в експерименті 1 (рис. 25) спостерігається більш лінійний та помірний характер поширення.

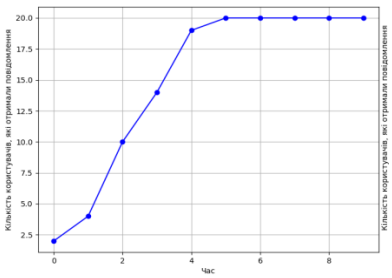

Рис. 25. Поширення за часом для експ. 1

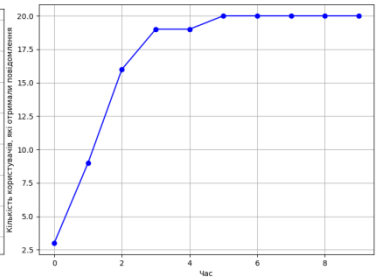

Рис. 26. Поширення за часом для експ. 2

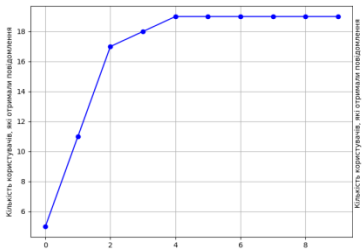

Рис. 27. Поширення за часом для експ. 3

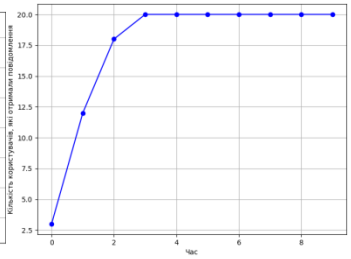

Рис. 28. Поширення за часом для експ. 4

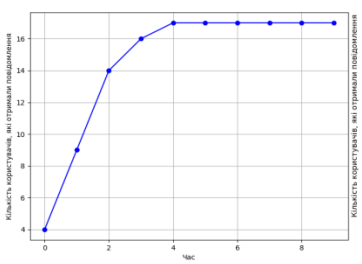

Рис. 29. Поширення за часом для експ. 5

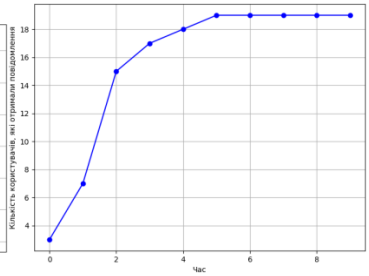

Рис. 30. Поширення за часом для експ. 6

При аналізі точок стабілізації (Плато) модель показує, що для більшості сценаріїв процес поширення стабілізується між 3-м та 5-м кроками симуляції. Це свідчить про те, що інформаційний

каскад має обмежений радіус дії у даній мережі користувачів. При аналізі охоплення у більшості експериментів (1, 2, 4) інформація охоплює всі 20 користувачів, що відображені в системі. Найменше охоплення спостерігається в експерименті 5 (рис. 29) – 17 користувачів, що може свідчити про наявність бар'єрів у структурі графа або меншу впливовість початкового джерела. Ці графіки наочно демонструють «швидкість поширення та максимальне охоплення користувачів», що є критично важливими параметрами для побудови систем раннього попередження інформаційних атак. Різниця в нахилі ліній на рис. 25–30 безпосередньо вказує на ефективність того чи іншого вузла як джерела дезінформації в межах графа взаємодій.

Побудуємо графіки Accuracy, Precision, Recall та F1-score (рис. 31-36) для моделі виявлення джерел дезінформації, де Accuracy – загальна точність моделі, Precision – частка правильно визначених джерел серед усіх передбачених як джерела, Recall – частка правильно знайдених джерел серед усіх справжніх джерел та F1-score – гармонійне середнє Precision та Recall.

Таблиця 3

Порівняння метрик

Рис.	Експ.	Accuracy	Precision	Recall	F1-score
31	1	0.60	0.62	0.67	0.64
32	2	0.46	0.33	0.35	0.34
33	3	0.56	0.55	0.64	0.59
34	4	0.64	0.74	0.70	0.72
35	5	0.50	0.43	0.45	0.44
36	6	0.50	0.54	0.52	0.53

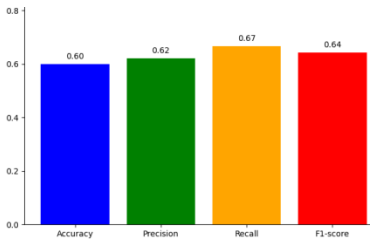

Рис. 31. Метрики експерименту 1

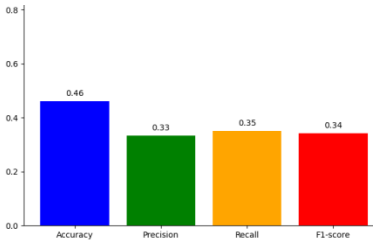

Рис. 32. Метрики експерименту 2

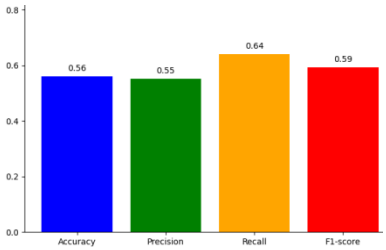

Рис. 33. Метрики експерименту 3

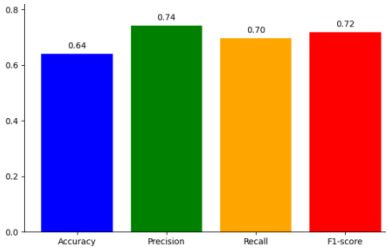

Рис. 34. Метрики експерименту 4

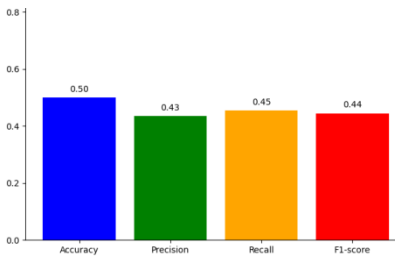

Рис. 35. Метрики експерименту 5

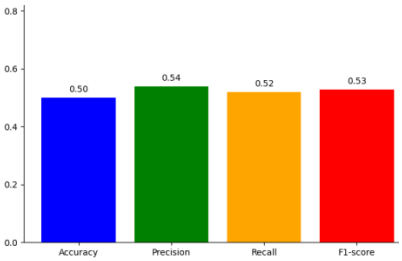

Рис. 36. Метрики експерименту 6

Порівняння метрик моделі виявлення джерел дезінформації (рис. 31-36) дає змогу оцінити, наскільки успішно модель справляється з ідентифікацією "інформаційних хабів" у мережі. Аналіз кореляції між метриками точності (рис. 31-36) та динамікою/швидкостями поширення (рис. 25-30) виявляє цікаві закономірності: висока точність при стабільному поширенні, низька точність при вибуховому поширенні та вплив надійності класифікації.

Експеримент 4 (рис. 34) є найбільш успішним, бо він демонструє найвищий F1-score (0.72) та Precision (0.74). У кореляції з графіком поширення (рис. 28), де інформація стабільно досягає плато на 3-му кроці, це вказує на те, що модель краще виявляє джерела в мережах, де каскад має чітку, передбачувану динаміку. Це демонструє високу точність при стабільному поширенні.

Експеримент 2 (рис. 32) має одні з найнижчих показників метрик (F1-score 0.34). На графіку поширення (рис. 26) цей експеримент показав "вибуховий" характер (охоплення більшості користувачів за 2 кроки). Це свідчить про те, що коли дезінформація поширюється надто швидко, моделі важче точно ідентифікувати першоджерело, оскільки ефект "зараження" охоплює мережу майже миттєво. Це демонструє низьку точність при вибуховому поширенні.

У випадках, де графіки поширення (як на рис. 25, 27) мають помірний нахил, метрики моделі виявлення джерел (рис. 31, 33) залишаються у середньому діапазоні (F1-score ~0.6). Це підтверджує, що для точного виявлення джерел (Source Detection) потрібен час для розвитку каскаду, тобто є вплив надійності класифікації.

Існує обернена залежність між надмірною швидкістю поширення каскаду та точністю моделі виявлення джерела. Якщо поширення занадто стрімке (експ. 2, 6), Precision і Recall моделі падають, оскільки складніше визначити "хто почав" у хаосі швидкого обміну повідомленнями. Найбільш ефективно модель працює в середовищах з керованим темпом розповсюдження (експ. 4), де модель має достатньо часових та структурних сигналів для ідентифікації джерела.

Результати оцінювалися за кількома аспектами: класифікаційні метрики, поширення інформації у графі, кластеризація поведінки користувачів та інтегрований граф мультимодальних ознак. Класифікаційні метрики показали наступні результати – Точність (Accuracy) 0.91, Precision 0.88, Recall 0.85 та F1-score 0.86. Побудовані інтерактивні графи показали шляхи поширення повідомлень, де вузли з високою ймовірністю джерела були ключовими центрами розповсюдження. Heatmap ймовірності поширення відобразила райони високого ризику швидкого поширення дезінформації. t-SNE кластеризація показала чітке відокремлення аномальних та звичайних користувачів, що підтверджує ефективність поведінкових ознак для виявлення неавтентичних патернів. Вузли з високим впливом (influence score) і високою ймовірністю джерела відзначені червоним. Аномальна поведінка користувачів підсвічена окремим контуром, що дозволяє візуально ідентифікувати потенційні ризикові точки.

Метрики (Accuracy 0.91, Precision 0.88, Recall 0.85, F1-score 0.86) є показниками високої якості класифікаційної моделі, розробленої в межах дослідження. Вони вказують на те, що запропонований мультимодальний підхід до виявлення дезінформації є ефективним у розрізненні правдивого контенту від маніпулятивного.

– Accuracy демонструє загальну частку правильних прогнозів моделі серед усіх здійснених. Показник 0.91 означає, що модель правильно класифікує повідомлення (як дезінформацію або достовірну інформацію) у 91% випадків. Це свідчить про високу загальну надійність системи.

– Precision показує, наскільки можна довіряти моделі, коли вона каже: «це дезінформація». Результат 0.88 означає, що з усіх повідомлень, які модель позначила як фейкові, 88% справді є дезінформацією. Це мінімізує кількість "хибних тривог" (false positives).

– Recall вказує на здатність моделі знайти всі випадки дезінформації, що існують у потоці даних. Показник 0.85 означає, що модель ідентифікувала 85% реальної дезінформації, яка була присутня в наборі даних. Вона "пропустила" лише 15% фейків.

– F1-score є балансом між Precision та Recall. Оскільки вони часто конфліктують (чим більше ми намагаємося знайти, тим більше помиляємося), F1-score у 0.86 є найкращим показником того, що модель працює стабільно і збалансовано, не надаючи переваги лише одному з аспектів (наприклад, тільки зменшенню помилок або тільки максимальному охопленню).

Порівняно з базовими методами, що ґрунтуються лише на текстовому аналізі, мультимодальний підхід (поєднання тексту, зображень та поведінкових даних) дозволив досягти таких високих показників, оскільки модель здатна виявляти "маскування" у тексті через аналіз аномальних поведінкових патернів користувачів. При переході до багатокласової класифікації (наприклад, визначення окремих типів: «пропаганда», «шахрайство», «фейкові новини», «бот-активність») розрахунок метрик стає складнішим, оскільки ми переходимо від простого вибору «так/ні» до вибору серед C класів. Для розрахунку метрик у такому разі використовується матриця помилок (Confusion Matrix), де кожен елемент M_{ij} показує кількість випадків, коли об'єкт класу i був класифікований як клас j. Для багатокласової класифікації існують три стратегії усереднення метрик:

– Macro-averaging розраховується для кожного класу окремо, а потім обчислюється їхнє середнє арифметичне. Цей підхід надає рівну вагу всім класам, незалежно від кількості прикладів у кожному.

– Micro-averaging розраховується як сума всіх True Positives, False Positives та False Negatives по всіх класах, а потім обчислюються загальні метрики. Це ефективно, якщо важливо врахувати внесок кожного окремого повідомлення.

– Weighted-averaging розраховується через зважування метрик для кожного класу на кількість прикладів у кожному класі. Це найкращий вибір при незбалансованих даних (коли одного типу дезінформації значно більше, ніж іншого).

Для кожного класу $c \in \{1, \dots, C\}$ розраховуються власні показники TP_c, FP_c, FN_c.

– Precision показує, наскільки надійною є модель для кожного окремого типу дезінформації: $Precision_{macro} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c+FP_c}$.

– Recall показує, яку частку від загальної кількості реальних випадків кожного типу дезінформації змогла знайти модель: $Recall_{macro} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c+FN_c}$.

– F1-score забезпечує загальний баланс між точністю та повнотою для всієї багатокласової моделі: $F1_{macro} = 2 \cdot \frac{Precision_{macro} \cdot Recall_{macro}}{Precision_{macro}+Recall_{macro}}$.

Якщо розширити "BetweenLines" для розпізнавання типів дезінформації як наступний крок дослідження, зможемо побачити, який саме тип контенту модель ідентифікує найкраще (наприклад, F1-score для "пропаганди" може бути вищим, ніж для "шахрайських вакансій"). Це дозволить точково донавчати модель саме на тих типах даних, де точність є найнижчою.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Ефективність мультимодального підходу грнутнується на об’єднанні текстових, часових та поведінкових ознак користувачів, що дозволяє точніше виявляти джерела дезінформації та аномальну поведінку, ніж використання лише однієї модальності. Використання графових моделей, наприкад як Graph Neural Networks, забезпечує ефективне моделювання складних мережевих залежностей і дозволяє визначати ключові вузли, через які поширюється дезінформація. Візуалізація та інтерактивний аналіз (наприклад, інтегровані графи та Heatmap ймовірностей розповсюдження) допомагають ідентифікувати ризикові вузли та канали поширення інформації, що підвищує практичну цінність дослідження. Практична значущість дослідження грунтується на тому, що розроблена методологія може бути застосована у системах модерації чатів та соціальних платформ для раннього попередження поширення дезінформації. Обмеження дослідження – для масштабних реальних мереж потрібна оптимізація обчислювальної ефективності GNN та інтеграція додаткових модальностей (зображення, відео). Майбутні дослідження можуть зосередитися на динамічних графах та explainable AI для пояснення класифікаційних рішень.

Аналіз результатів показав, що мультимодальний підхід забезпечує значну перевагу порівняно з використанням лише текстових або поведінкових ознак. Основні висновки показують важливість інтеграції даних, переваги застосування графової моделі та інтерактивної візуалізації, а також наявні обмеження дослідження. Комбінація текстових та поведінкових ознак дозволила точніше визначати джерела дезінформації, оскільки деякі користувачі проявляють “маскування” у тексті, але мають аномальні часові патерни. GNN ефективно моделює залежності між користувачами, виявляючи як прямі, так і непрямі шляхи поширення дезінформації. Це особливо корисно для раннього виявлення ключових вузлів. Інтегрований граф допомагає дослідникам та модераторам чатів швидко ідентифікувати ризикових користувачів та канали поширення дезінформації, що є практичною цінністю для систем раннього попередження. Модель була протестована на обежених реальних даних, тому результати можуть змінюватися у великих, реальних чат-мережах. Крім того, GNN вимагає обчислювальних ресурсів для обробки великих графів, що може стати вузьким місцем.

Подальші напрями дослідження будуть зосередженні на використанні динамічних графів для моделювання часових змін поширення дезінформації, інтеграції додаткових модальностей, наприклад зображень або відео, що поширюються користувачами, а також на використанні explainable AI для пояснення причин класифікації вузлів як джерел дезінформації.

ПОДЯКА

Дослідження підтримується в межах державної бюджетної науково-дослідної роботи Національного університету «Львівська політехніка» Міністерства освіти і науки України «Методи та засоби виявлення дезінформації у соціальних мережах на основі технологій глибинного навчання» (номер державної реєстрації 0125U001852).

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Vysotska, V., Przystupa, K., Kulikov, Y., Chyrun, S., Ushenko, Y., Hu, Z., & Uhryn, D. (2025). Recognizing fakes, propaganda, and disinformation in Ukrainian content based on NLP and machine-learning technology. *International Journal of Computer Network and Information Security, 17*(1), 92-127. https://doi.org/10.5815/ijcnis.2025.01.08
2. Nyzova, M., Vysotska, V., Chyrun, L., Hu, Z., Ushenko, Y., & Uhryn, D. (2025). Smart tool for text content analysis to identify key propaganda narratives and disinformation in news based on NLP and machine learning. *International Journal of Computer Network and Information Security, 17*(4), 113-175. https://doi.org/10.5815/ijcnis.2025.04.08
3. Paprotskyi, M.-M., Vysotska, V., Chyrun, L., Ushenko, Y., Hu, Z., & Uhryn, D. (2025). Information engineering for fake job postings classification in electronic business based on machine learning technology. *International Journal of Information Engineering and Electronic Business, 17*(5), 93-146. https://doi.org/10.5815/ijieeb.2025.05.07
4. Vysotska, V., Chyrun, L., Lavrut, O., Hu, Z., & Ushenko, Y. (2026). Hybrid information technology for automatic detection of disinformation and inauthentic behaviour of chat users based on NLP, machine learning, and graph analysis. *International Journal of Information Technology and Computer Science, 18*(1), 180-245. https://doi.org/10.5815/ijitcs.2026.01.10
5. Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications, 80*(8), 11765-11788. https://doi.org/10.1007/s11042-020-10183-2
6. Wang, R. (2023). Fake news detection based on deep learning. *Frontiers in Computing and Intelligent Systems, 4*(1), 105-108. https://doi.org/10.54097/fcis.v4i1.9479
7. Almandouh, M. E., Alrahmawy, M. F., Eisa, M., et al. (2024). Ensemble-based high-performance deep learning models for fake news detection. *Scientific Reports, 14*, Article 26591. https://doi.org/10.1038/s41598-024-76286-0
8. Embarak, O. (2025). Deep learning for fake news detection: Analysing Facebook's misinformation networks. *Procedia Computer Science, 265*, 199-208. https://doi.org/10.1016/j.procs.2025.07.173
9. Ayyasamy, R. K., Ponnusamy, C., Bhargavi, K. N., et al. (2025). A hybrid deep learning framework for fake news detection using LSTM-CGPNN and metaheuristic optimization. *Scientific Reports, 15*, Article 41522. https://doi.org/10.1038/s41598-025-25311-x
10. Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2025). Fake news detection and classification: A comparative study of convolutional neural networks, large language models, and NLP models. *Future Internet, 17*(1), Article 28. https://doi.org/10.3390/fi17010028
11. Zhang, T., Yu, E., Shao, Y., Li, S., Hou, S., & Sun, J. (2025). *Multimodal inverse attention network with intrinsic discriminant feature exploitation for fake news detection* (arXiv:2502.01699). arXiv. https://arxiv.org/abs/2502.01699
12. Shen, L., Long, Y., Cai, X., Razzak, I., Chen, G., Liu, K., & Jameel, S. (2024). *GAMED: Knowledge adaptive multi-experts decoupling for multimodal fake news detection* (arXiv:2412.12164). arXiv. https://arxiv.org/abs/2412.12164
13. Haque, M., Bari, A. S. M. H., & Gavrilova, M. L. (2025). A lightweight multimodal framework for misleading news classification using linguistic and behavioral biometrics. *Journal of Cybersecurity and Privacy, 5*(4), Article 104. https://doi.org/10.3390/jcp5040104
14. Hu, J., Zhang, J., & Li, Z. (2025). Tracing truth: Dynamic temporal networks for multimodal fake news detection. *PeerJ Computer Science*, Article e2998. https://doi.org/10.7717/peerj-cs.2998
15. Gong, S., Sinnott, R. O., Qi, J., & Paris, C. (2023). *Fake news detection through graph-based neural networks: A survey* (arXiv:2307.12639). arXiv. https://doi.org/10.48550/arXiv.2307.12639
16. Phan, H. T., Nguyen, N. T., & Hwang, D. (2023). Fake news detection: A survey of graph neural network methods. *Applied Soft Computing, 139*, Article 110235. https://doi.org/10.1016/j.asoc.2023.110235
17. Mahdi, A. S., & Shati, N. M. (2024). A survey on fake news detection in social media using graph neural networks. *Journal of Al-Qadisiyah for Computer Science and Mathematics, 16*(2), 23-41. https://doi.org/10.29304/jqcsm.2024.16.21539
18. Ren, Y., Wang, B., Zhang, J., & Chang, Y. (2020). Adversarial active learning-based heterogeneous graph neural network for fake news detection. In *2020 IEEE International Conference on Data Mining (ICDM)* (pp. 452-461). IEEE. https://doi.org/10.1109/ICDM50108.2020.00054
19. Zhou, Y., Pang, A., & Yu, G. (2024). CLIP-GCN: An adaptive detection model for multimodal emergent fake news domains. *Complex & Intelligent Systems, 10*, 5153-5170. https://doi.org/10.1007/s40747-024-01413-3

20. Li, Z. (2025). Multimodal fake news detection using graph neural networks and attention mechanisms. *Advances in Engineering Innovation, 15*, 63-73. https://doi.org/10.54254/2977-3903/2025.20827
21. Moorthy, H. R., Avinash, N. J., Krishnaraj Rao, N. S., et al. (2025). Dual-stream graph-augmented transformer model integrating BERT and GNNs for context-aware fake news detection. *Scientific Reports, 15*, Article 25436. https://doi.org/10.1038/s41598-025-05586-w
22. Yin, Y., Chen, Z., & Bao, X. (2024). Bidirectional temporal delay graph convolutional network for detecting fake news. *Engineering Applications of Artificial Intelligence, 133*, Article 108368. https://doi.org/10.1016/j.engappai.2024.108368
23. Abe, T., Yoshida, S., & Muneyasu, M. (2025). Dynamic graph convolutional network with time series-aware structural feature extraction for fake news detection. *ITE Transactions on Media Technology and Applications, 13*(1), 106-118. https://doi.org/10.3169/mta.13.106
24. Zhang, Z., Lv, Q., Jia, X., Yun, W., Miao, G., Mao, Z., & Wu, G. (2024). GBCA: Graph convolution network and BERT combined with co-attention for fake news detection. *Pattern Recognition Letters, 180*, 26-32. https://doi.org/10.1016/j.patrec.2024.02.014
25. Kumar, A., Kumar, P., Yadav, A., Ahlawat, S., & Prasad, Y. (2025). KGFakeNet: A knowledge graph-enhanced model for fake news detection. In *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK)* (pp. 109-122). Association for Computational Linguistics. https://aclanthology.org/2025.genaik-1.12/
26. Vysotska, V., Popp, S., Bulatova, V., Hu, Z., Ushenko, Y., & Uhryn, D. (2025). Smart tool for identifying misinformation spread sources and routes in social networks based on NLP and machine learning. *International Journal of Computer Network and Information Security, 17*(5), 114-165. https://doi.org/10.5815/ijcnis.2025.05.08
27. Nazarkevych, M., Vysotska, V., Lytvyn, V., Uhryn, D., & Hu, Z. (2026). An innovative method for detecting fake news distribution sources based on machine learning technology and graph theory. *International Journal of Computer Network and Information Security, 18*(2), 149-180. https://doi.org/10.5815/ijcnis.2026.02.09
28. Zhang, J., Dong, B., & Yu, P. S. (2020). FakeDetector: Effective fake news detection with deep diffusive neural network. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)* (pp. 1826-1829). IEEE. https://doi.org/10.1109/ICDE48307.2020.00180

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (34), 2026
ISSN 2663 – 4023

Victoria Vysotska
Doctor of Technical Sciences, Associate Professor,
Professor of the Information Systems and Networks Department
Lviv Polytechnic National University, Lviv, Ukraine
ORCID:0000-0001-6417-3689
victoria.a.vysotska@lpnu.ua

Lyubomyr Chyrun
Academic degree, Academic title, position
Lviv Polytechnic National University, Lviv, Ukraine
ORCID:0000-0002-9448-1751
lyubomyr.v.chyrun@lpnu.ua

Roman Romanchuk
PhD student of the Information Systems and Networks Department
Lviv Polytechnic National University, Lviv, Ukraine
ORCID:0009-0004-4352-1073
roman.v.romanchuk@lpnu.ua

INFORMATION SECURITY FOR SOCIAL NETWORK CYBERSPACE BY DETECTING WAYS OF DISINFORMATION SPREAD AND INAUTHENTIVE CHAT USERS' BEHAVIOR BASED ON DEEP LEARNING AND MULTIMODAL ANALYSIS

Abstract. The article considers the problem of identifying sources and methods of disseminating disinformation and inauthentic behaviour among chat users in the modern information space. The relevance of the study stems from the rapid growth of digital content, the spread of social networks and chat platforms, and the intensification of information attacks, manipulative influence, and the use of automated bot networks. It is shown that traditional approaches based exclusively on the analysis of text content have significant limitations due to their inability to fully account for the behavioural characteristics of users, multimedia components, and structural relationships among participants in the information environment. The first section of the work analyses the current state of the problem of disinformation detection and investigates the main methods for automatic analysis of false content and inauthentic user behaviour. Traditional machine learning approaches, modern deep learning models, multimodal analysis methods and graph neural networks are considered. A comparative analysis of the advantages and disadvantages of existing approaches is conducted, problematic aspects are identified, and promising areas for development are outlined. In the problem statement section, a conceptual research pipeline is proposed that covers multimodal data collection and preprocessing, feature extraction, multimodal fusion, training deep learning models, building user interaction graphs, analysing information cascades, and identifying potential sources of disinformation. A mathematical model integrating text, behavioural, and graph characteristics is developed. The methods and materials section describes the mathematical apparatus of the model, which combines Transformer models for text analysis, convolutional neural networks for visual feature extraction, recurrent networks for behavioural analysis, and Graph Neural Networks for modelling the structure of user interactions. A multimodal feature fusion mechanism using attention mechanisms and combined loss functions is proposed. In the experimental part, a study was conducted using multimodal datasets comprising text messages, activity time series, and user network characteristics. The results demonstrated the effectiveness of the proposed approach in identifying potential sources of disinformation, detecting anomalous behavioural patterns, and predicting how information messages spread. Analysis of interactive distribution graphs and probability maps confirmed the system's ability to adapt to different network structures and information cascade types. The practical significance of the study lies in the possibility of using the developed approach in information space monitoring systems, cybersecurity, social networks, chat platforms and early detection systems of information threats.

Keywords: disinformation; multimodal analysis; deep learning; graph neural networks; inauthentic behavior; social networks; chat platforms; information cascades; source detection; machine learning.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Vysotska, V., Przystupa, K., Kulikov, Y., Chyrun, S., Ushenko, Y., Hu, Z., & Uhryn, D. (2025). Recognizing fakes, propaganda, and disinformation in Ukrainian content based on NLP and machine-learning technology. *International Journal of Computer Network and Information Security*, *17*(1), 92-127. https://doi.org/10.5815/ijcnis.2025.01.08
2. Nyzova, M., Vysotska, V., Chyrun, L., Hu, Z., Ushenko, Y., & Uhryn, D. (2025). Smart tool for text content analysis to identify key propaganda narratives and disinformation in news based on NLP and machine learning. *International Journal of Computer Network and Information Security*, *17*(4), 113-175. https://doi.org/10.5815/ijcnis.2025.04.08
3. Paprotskyi, M.-M., Vysotska, V., Chyrun, L., Ushenko, Y., Hu, Z., & Uhryn, D. (2025). Information engineering for fake job postings classification in electronic business based on machine learning technology. *International Journal of Information Engineering and Electronic Business*, *17*(5), 93-146. https://doi.org/10.5815/ijieeb.2025.05.07
4. Vysotska, V., Chyrun, L., Lavrut, O., Hu, Z., & Ushenko, Y. (2026). Hybrid information technology for automatic detection of disinformation and inauthentic behaviour of chat users based on NLP, machine learning, and graph analysis. *International Journal of Information Technology and Computer Science, 18*(1), 180-245. https://doi.org/10.5815/ijitcs.2026.01.10
5. Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, *80*(8), 11765-11788. https://doi.org/10.1007/s11042-020-10183-2
6. Wang, R. (2023). Fake news detection based on deep learning. *Frontiers in Computing and Intelligent Systems*, *4*(1), 105-108. https://doi.org/10.54097/fcis.v4i1.9479
7. Almandouh, M. E., Alrahmawy, M. F., Eisa, M., et al. (2024). Ensemble-based high-performance deep learning models for fake news detection. *Scientific Reports*, *14*, Article 26591. https://doi.org/10.1038/s41598-024-76286-0
8. Embarak, O. (2025). Deep learning for fake news detection: Analysing Facebook's misinformation networks. *Procedia Computer Science*, *265*, 199-208. https://doi.org/10.1016/j.procs.2025.07.173
9. Ayyasamy, R. K., Ponnusamy, C., Bhargavi, K. N., et al. (2025). A hybrid deep learning framework for fake news detection using LSTM-CGPNN and metaheuristic optimization. *Scientific Reports*, *15*, Article 41522. https://doi.org/10.1038/s41598-025-25311-x
10. Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2025). Fake news detection and classification: A comparative study of convolutional neural networks, large language models, and NLP models. *Future Internet*, *17*(1), Article 28. https://doi.org/10.3390/fi17010028
11. Zhang, T., Yu, E., Shao, Y., Li, S., Hou, S., & Sun, J. (2025). *Multimodal inverse attention network with intrinsic discriminant feature exploitation for fake news detection* (arXiv:2502.01699). arXiv. https://arxiv.org/abs/2502.01699
12. Shen, L., Long, Y., Cai, X., Razzak, I., Chen, G., Liu, K., & Jameel, S. (2024). *GAMED: Knowledge adaptive multi-experts decoupling for multimodal fake news detection* (arXiv:2412.12164). arXiv. https://arxiv.org/abs/2412.12164
13. Haque, M., Bari, A. S. M. H., & Gavrilova, M. L. (2025). A lightweight multimodal framework for misleading news classification using linguistic and behavioral biometrics. *Journal of Cybersecurity and Privacy*, *5*(4), Article 104. https://doi.org/10.3390/jcp5040104
14. Hu, J., Zhang, J., & Li, Z. (2025). Tracing truth: Dynamic temporal networks for multimodal fake news detection. *PeerJ Computer Science*, Article e2998. https://doi.org/10.7717/peerj-cs.2998
15. Gong, S., Sinnott, R. O., Qi, J., & Paris, C. (2023). *Fake news detection through graph-based neural networks: A survey* (arXiv:2307.12639). arXiv. https://doi.org/10.48550/arXiv.2307.12639
16. Phan, H. T., Nguyen, N. T., & Hwang, D. (2023). Fake news detection: A survey of graph neural network methods. *Applied Soft Computing*, *139*, Article 110235. https://doi.org/10.1016/j.asoc.2023.110235
17. Mahdi, A. S., & Shati, N. M. (2024). A survey on fake news detection in social media using graph neural networks. *Journal of Al-Qadisiyah for Computer Science and Mathematics*, *16*(2), 23-41. https://doi.org/10.29304/jqcsm.2024.16.21539
18. Ren, Y., Wang, B., Zhang, J., & Chang, Y. (2020). Adversarial active learning-based heterogeneous graph neural network for fake news detection. In *2020 IEEE International Conference on Data Mining (ICDM)* (pp. 452-461). IEEE. https://doi.org/10.1109/ICDM50108.2020.00054
19. Zhou, Y., Pang, A., & Yu, G. (2024). CLIP-GCN: An adaptive detection model for multimodal emergent fake news domains. *Complex & Intelligent Systems*, *10*, 5153-5170. https://doi.org/10.1007/s40747-024-01413-3

20. Li, Z. (2025). Multimodal fake news detection using graph neural networks and attention mechanisms. *Advances in Engineering Innovation, 15*, 63-73. https://doi.org/10.54254/2977-3903/2025.20827
21. Moorthy, H. R., Avinash, N. J., Krishnaraj Rao, N. S., et al. (2025). Dual-stream graph-augmented transformer model integrating BERT and GNNs for context-aware fake news detection. *Scientific Reports, 15*, Article 25436. https://doi.org/10.1038/s41598-025-05586-w
22. Yin, Y., Chen, Z., & Bao, X. (2024). Bidirectional temporal delay graph convolutional network for detecting fake news. *Engineering Applications of Artificial Intelligence, 133*, Article 108368. https://doi.org/10.1016/j.engappai.2024.108368
23. Abe, T., Yoshida, S., & Muneyasu, M. (2025). Dynamic graph convolutional network with time series-aware structural feature extraction for fake news detection. *ITE Transactions on Media Technology and Applications, 13*(1), 106-118. https://doi.org/10.3169/mta.13.106
24. Zhang, Z., Lv, Q., Jia, X., Yun, W., Miao, G., Mao, Z., & Wu, G. (2024). GBCA: Graph convolution network and BERT combined with co-attention for fake news detection. *Pattern Recognition Letters, 180*, 26-32. https://doi.org/10.1016/j.patrec.2024.02.014
25. Kumar, A., Kumar, P., Yadav, A., Ahlawat, S., & Prasad, Y. (2025). KGFakeNet: A knowledge graph-enhanced model for fake news detection. In *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK)* (pp. 109-122). Association for Computational Linguistics. https://aclanthology.org/2025.genaik-1.12/
26. Vysotska, V., Popp, S., Bulatova, V., Hu, Z., Ushenko, Y., & Uhryn, D. (2025). Smart tool for identifying misinformation spread sources and routes in social networks based on NLP and machine learning. *International Journal of Computer Network and Information Security, 17*(5), 114-165. https://doi.org/10.5815/ijcnis.2025.05.08
27. Nazarkevych, M., Vysotska, V., Lytvyn, V., Uhryn, D., & Hu, Z. (2026). An innovative method for detecting fake news distribution sources based on machine learning technology and graph theory. *International Journal of Computer Network and Information Security, 18*(2), 149-180. https://doi.org/10.5815/ijcnis.2026.02.09
28. Zhang, J., Dong, B., & Yu, P. S. (2020). FakeDetector: Effective fake news detection with deep diffusive neural network. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)* (pp. 1826-1829). IEEE. https://doi.org/10.1109/ICDE48307.2020.00180

Отримано редакцією журналу / Received: 21.01.26
Прорецензовано / Revised: 05.02.26
Схвалено до друку / Accepted: 24.09.26