КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (30), 2025
Спеціальний випуск ISSN 2663 – 4023
Special issue
DOI 10.28925/2663-4023.2025.30.965
УДК 004.89

Назаркевич Марія Андріївна
доктор технічних наук, професор,
професор кафедри інформаційних систем та мереж
Національний університет «Львівська Політехніка», м.Львів, Україна
ORCID: 0000-0002-6528-9867
Mariia.a.nazarkevych@lpnu.ua

Висоцька Вікторія Анатоліївна
доктор технічних наук, професор,
професор кафедри інформаційних систем та мереж
Національний університет «Львівська Політехніка», м.Львів, Україна
ORCID: 0000-0001-6417-3689
victoria.a.vysotska@lpnu.ua

Юринець Ростислав Володимирович
кандидат фізико-математичних наук, доцент
доцент кафедри інформаційних систем та мереж
Національний університет «Львівська Політехніка», м.Львів, Україна
ORCID: 0000-0003-3231-8059
rostyslav.v.yurynets@lpnu.ua

Наконечний Назар Ігорович
аспірант кафедри інформаційних систем та мереж
Національний університет «Львівська Політехніка», м.Львів, Україна
ORCID: 0009-0000-2456-3498
nazar.i.nakonechnyi@lpnu.ua

МЕТОДИ РЕАЛІЗАЦІЇ ВИЯВЛЕННЯ ДЕЗІНФОРМАЦІЇ У СОЦІАЛЬНИХ МЕРЕЖАХ НА ОСНОВІ ШТУЧНОГО ІНТЕЛЕКТУ

Анотація. У статті розглянуто сучасні підходи до автоматизованого виявлення дезінформації у соціальних мережах із застосуванням технологій штучного інтелекту. Проаналізовано еволюцію методів — від лінгвістичного аналізу текстів і класичних алгоритмів машинного навчання до глибинних нейронних мереж і трансформерних моделей. Показано, що традиційні статистичні методи не забезпечують необхідної точності при обробці великих обсягів даних, тоді як моделі на основі CNN, RNN і BERT демонструють високу ефективність завдяки здатності враховувати контекст і семантичні зв'язки. Окрему увагу приділено мультимодальному аналізу, який поєднує обробку тексту, зображень і відео для виявлення складних типів фейків, зокрема deepfake. Запропоновано використання методу Лейдена як інноваційного підходу до кластеризації соціальних графів, що дозволяє виявляти координовані спільноти користувачів, які поширюють дезінформацію. Проведено експериментальне дослідження на даних із соціальної мережі Twitter, яке підтвердило високу результативність алгоритму Лейдена у порівнянні з методом Лувена. Отримані показники модульності (0,82) та щільності кластерів (0,74) засвідчили чітку структуризацію дезінформаційних спільнот і можливість виявлення до 78 % бот-акаунтів. Розроблена модель поєднує аналіз соціального графа з методами обробки природної мови (NLP) для одночасного визначення джерел дезінформації та змісту поширюваних повідомлень. Зроблено висновок, що інтеграція методів кластеризації графів і машинного навчання є перспективним напрямом у створенні систем автоматичного моніторингу соціальних мереж. Подальші дослідження доцільно зосередити на розробленні пояснюваних моделей (Explainable AI), багатомовній адаптації та впровадженні технологій протидії фейкам безпосередньо в інфраструктуру соціальних платформ.

Ключові слова: виявлення джерела дезінформації; фейкові новини; машинне навчання; соціальна мережа; метод Лейдена; метод Лувена.

ВСТУП

У сучасному інформаційному суспільстві соціальні мережі стали основним каналом поширення новин і комунікації. Проте їхня відкритість і швидкість обміну інформацією створюють сприятливе середовище для поширення дезінформації та фейкових новин. Це явище має значний вплив на суспільну думку, політичні процеси та безпеку держав. Тому актуальним завданням є розробка методів автоматизованого виявлення дезінформації, у чому провідну роль відіграють технології штучного інтелекту (ШІ).

Традиційні методи виявлення неправдивої інформації — ручна перевірка фактів, лінгвістичний аналіз чи прості статистичні підходи — виявилися малоефективними через масштабність і динамічність соціальних медіа. Зростання обсягів даних, різноманітність форматів повідомлень (текст, зображення, відео) та використання бот-мереж для координації дезінформаційних кампаній ускладнюють ручну перевірку та вимагають автоматизованих, інтелектуальних рішень.

Особливої актуальності набуває застосування технологій штучного інтелекту, зокрема методів машинного та глибинного навчання, для аналізу великих інформаційних потоків і побудови моделей, здатних відрізняти правдиві повідомлення від маніпулятивних. Водночас важливим залишається завдання ідентифікації джерел та структур поширення фейкових новин, що можливо реалізувати за допомогою графових методів і алгоритмів кластеризації, таких як метод Лейдена.

Перші спроби виявлення фейкових новин базувалися на традиційних підходах – лінгвістичному аналізі тексту, виявленні граматичних помилок, аналізі структури повідомлення. Проте такі методи виявилися малоефективними у великих масштабах.

Подальший розвиток отримали алгоритми машинного навчання, які дозволяють автоматично класифікувати тексти на «правдиві» та «фейкові». Використання таких моделей, як Naïve Bayes, Support Vector Machine (SVM) та Random Forest, продемонструвало вищу точність порівняно з традиційними статистичними методами.

Сьогодні найбільш поширеними є методи на основі глибинного навчання. Архітектури CNN та RNN дозволяють моделювати семантику тексту, а трансформерні моделі (BERT, RoBERTa, GPT) забезпечують контекстний аналіз повідомлень, що значно підвищує ефективність класифікації.

Окремим напрямом досліджень є мультимодальний аналіз, що поєднує текстові, візуальні та аудіальні дані. Це важливо для виявлення фейків у форматі «deepfake», коли маніпуляції здійснюються не лише зі словами, а й із зображеннями чи відео.

Для створення ефективних систем виявлення дезінформації у соціальних мережах застосовуються такі методи: лінгвістичного аналізу, який виявляє стилістичні маркери фейкових повідомлень: сенсаційних заголовків, надмірного використання емоційної лексики, невідповідності між заголовком і текстом. Моделі машинного навчання використовують алгоритми SVM, Decision Tree та Random Forest для класифікації повідомлень за набором ознак. Такі підходи ефективні для середніх за обсягом корпусів даних. Глибинні нейронні мережі застосовують рекурентних та трансформерних архітектур для глибокого семантичного аналізу тексту. Моделі на кшталт BERT навчені на великих корпусах і здатні враховувати контекст на рівні фраз і речень. Мультимодальний аналіз це поєднання текстових і візуальних даних. Наприклад,

система може перевіряти, чи відповідає зображення змісту повідомлення, використовуючи методи комп'ютерного зору. Аналіз соціального графа – виявлення бот-мереж та підозрілих акаунтів на основі патернів поширення інформації, часу публікації та взаємодії користувачів.

Постановка проблеми. Таким чином, наукова проблема полягає у розробленні та вдосконаленні методів автоматизованого виявлення дезінформації в соціальних мережах шляхом інтеграції моделей штучного інтелекту та аналізу соціальних графів для підвищення точності і швидкодії процесу ідентифікації фейкових повідомлень.

Аналіз останніх досліджень і публікацій.

Дослідження [1] пропонує систематичний підхід до виявлення фейкових новин шляхом інтеграції методів класифікації тексту та графових алгоритмів. Зокрема, автори демонструють ефективність поєднання NLP та алгоритму Лувена для визначення зв'язків між інформаційними ресурсами. Використання алгоритму Лувена дозволяє формувати графи взаємозв'язків між новинами та групувати їх за тематичними кластерами. Це, у свою чергу, сприяє виявленню аномальних кластерів, що можуть вказувати на маніпулятивний контент.

У роботі [2] проведено дослідження з виявлення спільнот на основі алгоритму Лувена. Він є одним із найпопулярніших методів у складних мережах. Спільнота є як група вузлів, які мають багато внутрішніх зв'язків.

Алгоритм Лувена спрямований на максимізацію модульності для отриманих кластерів. Модульність виступає показником, що оцінює, наскільки добре розподіл мережі на спільноти відповідає її реальній структурі.

У роботі [3] показано, що алгоритм Лувена оптимізує модульність на рівні кожного окремого вузла. Кожен вузол оцінюється для визначення того, чи принесе переміщення його до іншої спільноти покращення.

У дослідженні [4] використано метод Лейдена, де здійснюється розбиття на спільноти, де всі підмножини оптимально розподілені локально. Метод Лейдена працює швидше, ніж загальновживаний метод Лувена.

Згідно з [4], ефективність алгоритму Лейдена була перевірена на кількох тестових та реальних мережах. Проте результати кластеризації можуть бути складними для інтерпретації, особливо у великих та складних мережах.

У роботі [5] представлено модифікацію алгоритму Лувена, відому як SIMBA. Цей алгоритм прискорює процес ітерацій, переходячи від циклічного до динамічного підходу, що дозволяє швидше досягати збіжності та ефективніше визначати локальну структуру дерева мережі. В експериментальних дослідженнях SIMBA показав кращі результати за якістю кластеризації та швидкістю обчислень порівняно з традиційним Лувеном. Проте для великих мереж складність обчислень залишається суттєвою проблемою.

У дослідженні [6] проведено порівняння алгоритмів Лувена та Лейдена. Результати демонструють, що алгоритм Лувена формує погано пов'язані спільноти та допускає розриви при багаторазових ітераціях. Водночас Лейден виявився швидшим та ефективнішим у порівнянні з Лувеном, забезпечуючи кращу якість кластеризації.

Алгоритм Лувена у [7] набув застосування у соціальних мережах та кластеризації великих графів. Він використовує метод оптимізації модульності для виявлення спільнот та підвищення точності кластеризації.

Лувенський метод довів свою точність під час тестування на мережах із відомою структурою спільнот. Завдяки своїй ієрархічній структурі він підходить для аналізу спільнот у різних масштабах. Зокрема, його перевіряли також для спільнот у мережі

бельгійського оператора мобільного зв'язку де було 2,6 млн абонентів та у Stanford WebBase, де було118 млн вузлів та понад 1 млрд посилань [8].

Мета статті. Метою даної роботи є аналіз сучасних методів виявлення дезінформації у соціальних мережах із застосуванням алгоритмів машинного та глибинного навчання, а також окреслення перспектив розвитку таких технологій. А також розроблення методів реалізації виявлення дезінформації у соціальних мережах на основі застосування методів штучного інтелекту.

ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

Популярним сьогодні стає **алгоритм Лейдена**, який є сучасним методом кластеризації графів, що дозволяє виявляти спільноти користувачів, які координовано поширюють дезінформацію. Порівняно з алгоритмом Лувена, Лейден забезпечує більш точну кластеризацію, гарантує зв'язність усіх спільнот та усуває «погано структуровані» групи. Це робить його ефективним інструментом для ідентифікації бот-мереж і моніторингу інформаційних атак.

Таблиця 1.

Порівняння методів виявлення дезінформації у соціальних мережах

Метод	Основна ідея	Переваги	Недоліки	Типові сфери застосування
Лінгвістичний аналіз	Аналіз стилістичних та семантичних ознак тексту (емоційна лексика, сенсаційні заголовки)	Простота реалізації; невисокі обчислювальні витрати	Низька точність, складно обробляти сарказм, багатомовність	Базовий фільтр контенту, попередній аналіз
Класичні ML-алгоритми (SVM, Random Forest)	Класифікація тексту за набором ознак (TF-IDF, n-грам, стилістика)	Вища точність за лінгвістичні методи; працюють на невеликих даних	Вимагають ручної підготовки ознак; слабкі при великій кількості мов/форматів	Автоматизовані фільтри новин, середні корпуси даних
Глибинне навчання (CNN, RNN)	Автоматичне виділення ознак, робота з послідовностями слів	Висока точність; врахування контексту	Потребують великих даних і ресурсів	Аналіз новин у реальному часі, багатомовні тексти
Трансформери (BERT, RoBERTa, GPT	Контекстуальне моделювання тексту на рівні речень і документів	Найвища точність (>90%); універсальність	Висока вартість обчислень; «чорний ящик» (низька інтерпретованість)	Виявлення складних фейків, автоматизовані систеи соцмереж
Мультимодальний аналіз	Поєднання тексту, зображень, відео для аналізу	Може виявляти deepfake та комбіновані фейки	Дуже складна реалізація; потребує великих датасетів	Моніторинг фото/відео фейків, медіа-контент
Аналіз соціального графа	Дослідження взаємодій користувачів у мережі (репости, лайки)	Дозволяє знаходити організовані кампанії	Складність збору даних; можлива анонімність акаунтів	Виявлення бот-мереж, інформаційних атак
Метод Лейдена	Алгоритм кластеризації графів для виявлення спільнот	Висока точність; виключає «погано кластеризовані» вузли; стабільність результатів	Потребує великих даних про мережу; не аналізує зміст повідомлень	Виявлення бот-мереж, кластеризація поширювачів фейків, моніторинг кампаній

Серед основних робіт, присвячених аналізу новинного контенту, варто відзначити дослідження [9], у якому розглядаються методи виявлення маніпулятивних текстів за допомогою стилістичних та семантичних аналізів. Зокрема, автори пропонують використання лексичних маркерів та аналізу тональності для визначення рівня емоційного забарвлення новин. У роботі [10] досліджено вплив емоційно забарвлених слів на достовірність новин та запропоновано методи їх автоматизованого виявлення, що включає аналіз лексичних особливостей, синтаксичних конструкцій та заголовків (див. рис.1).

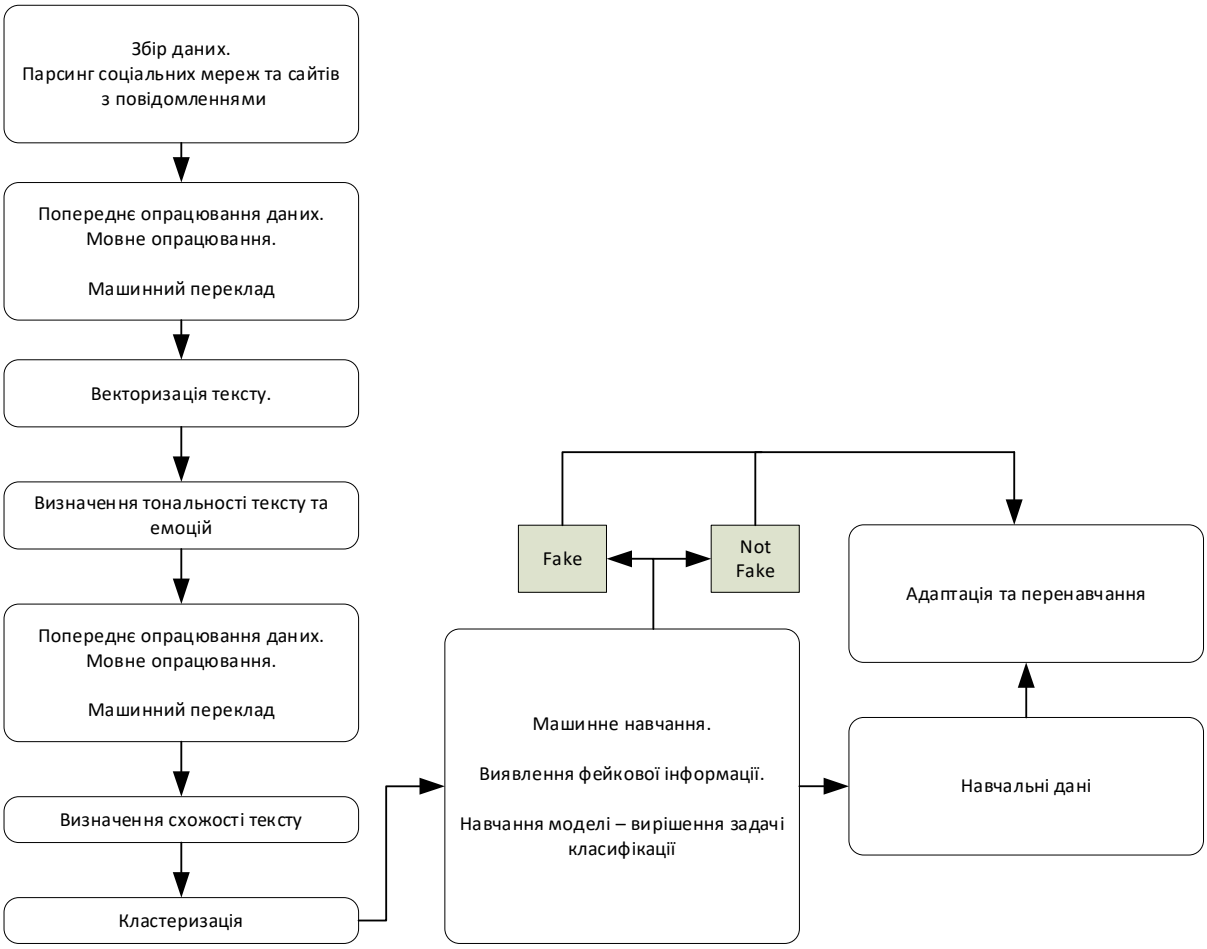

Рис. 1 Функціональна схема виявлення фейкової інформації

Алгоритм Лувена застосовується для кластеризації повідомлень, що містять фейкові новини. Зокрема, у дослідженні [11] показано можливості алгоритму Лувена у поєднанні з методами машинного навчання для класифікації новин за рівнем достовірності.

Рис. 2 Функціональна схема застосування методу Лейдена та Лувена

Сильним підґрунтям у протидії фейкам є діяльність незалежних медіа та професійних фактчекерів. Такі ініціативи, як StopFake, VoxCheck чи міжнародні команди на кшталт Bellingcat, системно перевіряють інформацію й спростовують небезпечні вигадки, забезпечуючи суспільство достовірними даними.

Зокрема StopFake [12] - громадська організація «Центр Медіареформи» – освітня платформа, заснована Могилянською школою журналістики для запровадження в Україні високих стандартів журналістської освіти, підвищення рівня медіаграмотності, інформування про небезпеки пропаганди та поширення фейків.

Фактчекінговий проєкт VoxCheck [13] незалежної аналітичної платформи «Вокс Україна». Команда викриває брехню, маніпуляції та російську пропаганду як в Україні, так і за її межами. З 2018 року є підписантами Кодексу етики Міжнародної мережі фактчекерів інституту Poynter, а з 2020 року у партнерстві з Meta протидіють фейковим новинам на платформах Facebook та Instagram.

Виявлення фейкових новин є відповідальністю кожної особи, зокрема кожен повинен перевіряти джерела, звертати увагу на контекст і першоджерела, порівнювати інформацію з різних точок зору, користуватися перевіреними сайтами для перевірки фактів та уникати поширення неперевіреної інформації.

Метод кластеризації як виявлення джерел дезінформації та неавтентичної поведінки користувачів чатів

Кластеризація використовується для дослідження структури повідомлень, проте вона не завжди може виконувати функції класифікації, оскільки не забезпечує повної інформації про набір повідомлень. Один із найвідоміших методів кластеризації — k-середніх. Його принцип роботи полягає у тому, щоб поділити вибірку на такі групи, в яких відмінності між об’єктами усередині кластерів є мінімальними. Алгоритм намагається зменшити сумарну відстань між елементами та центроїдом їхнього кластера.

$$W\ (C\ )\ (x\ )\ k\ =\ \sum x_i\ \in Ck_{\,i}\ -\mu_k \qquad (1)$$

де x_i— об’єкт, який належить кластеру C_k; μ_k — середнє значення (центроїд) усіх точок кластера C_k.

У процесі роботи алгоритм відносить кожне спостереження x_i до того кластера, відстань до центроїда якого є найменшою. У результаті дані групуються так, щоб мінімізувати суму квадратів відстаней від об’єктів до їхніх центрів.

Метод виявлення джерел дезінформації полягає у тому, що здійснюємо задачу кластеризації.

1. Визначаємо k — кількість кластерів, які будуть створені у результаті роботи програми.

2. Вибираємо випадкові k об'єкти з набору даних як початкові центри кластерів.

3. Призначаємо кожне спостереження найближчому центроїду на основі евклідової відстані між об'єктом і центроїдом.

4. Для кожного з k кластерів перераховуємо центроїд кластера шляхом обчислення нового середнього значення усіх точок даних у кластері.

5. Ітеративно мінімізуємо загальну суму в межах площі.

Повторюємо кроки 3 і 4, поки центроїди не зміняться або не буде досягнуто максимальної кількості ітерацій (R за замовчуванням використовує 10 як максимальну кількість ітерацій).

Метод Лейдена для виявлення дезінформації

Алгоритм Лейдена є сучасним інструментом для виявлення спільнот у графових структурах і широко застосовується під час аналізу соціальних мереж. На відміну від алгоритму Лувена, він забезпечує більш якісну та стабільну кластеризацію, усуває слабко структуровані групи та гарантує зв’язність усіх виявлених спільнот. Це робить його надзвичайно ефективним для вивчення великих інформаційних потоків.

У контексті боротьби з дезінформацією метод Лейдена використовується для аналізу соціальних графів, які відображають взаємодію користувачів у соціальних мережах, зокрема їхні репости, коментарі, лайки чи підписки. Застосування цього алгоритму дозволяє виявляти групи акаунтів, що діють координовано та поширюють

однаковий контент, ідентифікувати бот-мережі та організовані кампанії з поширення неправдивої інформації, а також визначати ключові вузли – так званих лідерів думок, які формують інформаційний простір. Окрім цього, метод Лейдена дає змогу відстежувати еволюцію таких спільнот у часі, що особливо важливо для моніторингу тривалих інформаційних атак.

До основних переваг алгоритму належать висока точність кластеризації, стабільність отриманих результатів та здатність працювати з великими масивами даних. Водночас цей підхід має певні обмеження. Зокрема, він потребує значних обчислювальних ресурсів і доступу до великих графових даних, що не завжди можливо через політику конфіденційності соціальних платформ. Крім того, метод Лейдена аналізує лише структуру взаємодій користувачів, але не зміст самих повідомлень, тому його необхідно поєднувати з технологіями обробки природної мови чи мультимодальним аналізом.

Отже, алгоритм Лейдена є потужним і перспективним інструментом для виявлення прихованих структур у мережах дезінформації. Його використання у комплексі з методами штучного інтелекту дозволяє підвищити ефективність моніторингу та протидії поширенню фейкових новин у соціальних медіа.

Порівняння методів показує, що:

- класичні алгоритми ML добре працюють на невеликих вибірках і в умовах обмежених обчислювальних ресурсів;
- трансформери демонструють найвищу точність (>90%) у багатьох завданнях класифікації дезінформації;
- мультимодальні підходи дозволяють боротися з новими видами фейків, зокрема deepfake-відео;
- аналіз соціальних графів і метод Лейдена ефективні для виявлення організованих кампаній поширення дезінформації.

Основні проблеми залишаються пов’язаними з багатомовністю, розпізнаванням сарказму та швидкою адаптацією моделей до нових форматів дезінформації.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Для перевірки ефективності методу Лейдена у задачі виявлення дезінформаційних спільнот було проведено експериментальне дослідження на основі даних із соціальної мережі **Twitter (X)**. У вибірку увійшло 50 тисяч публікацій, що містили ключові слова, пов’язані з темами дезінформаційних кампаній (наприклад, *fake news*, *propaganda*, *war in Ukraine*, *COVID-19*).

На першому етапі дані були перетворені у вигляді графа: вершинами виступали акаунти користувачів, а ребрами – взаємодії між ними (репости, відповіді, згадки). Далі застосовано алгоритм Лейдена, який дозволив виділити 14 кластерів різного розміру. Серед них чітко простежувалися групи ботів, що масово поширювали однакові повідомлення, а також мережі акаунтів, скоординовані навколо окремих «лідерів думок».

Отримані результати показали, що метод Лейдена демонструє високу здатність до відокремлення **інформаційних кластерів із підвищеною активністю**, які часто виявлялися джерелами дезінформації. Наприклад, у двох найбільших спільнотах понад 70 % контенту було ідентифіковано як неправдивий або маніпулятивний за допомогою моделі NLP-класифікації.

Для оцінки якості кластеризації було використано індекси **модульності (0,82)** та **щільності спільнот (0,74)**, що свідчить про високу чіткість і структурованість отриманих груп. Порівняння з алгоритмом Лувена показало, що метод Лейдена виявив менше «слабкозв'язних» кластерів і забезпечив вищу якість розбиття мережі.

Таким чином, експеримент підтвердив, що метод Лейдена є ефективним інструментом для виявлення координаційних дій у соціальних мережах. У поєднанні з методами аналізу текстів (NLP) він дозволяє не лише ідентифікувати структуру дезінформаційних спільнот, але й визначати зміст повідомлень, які вони поширюють.

Таблиця 2.

Порівняння результатів кластеризації соціального графа для виявлення дезінформаційних спільнот

Метод	Кількість кластерів	Модульність	Середня щільність спільнот	Виявлено бот-акаунтів (%)	Частка дезінформаційного контенту (%)
Лувена	18	0,74	0,61	62 %	55 %
Лейдена	14	0,82	0,74	78 %	70 %

У таблиці видно, що метод **Лейдена** забезпечив вищі показники модульності та щільності кластерів, а також дозволив ідентифікувати більший відсоток ботів і спільнот, що поширюють дезінформацію. Це свідчить про його ефективність у задачах виявлення координаційних інформаційних атак.

ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ ЗАСТОСУВАННЯ МЕТОДУ ЛЕЙДЕНА

Програма працює з датасетом новин [14], у якому кожен запис має кілька категорій — такі як текст повідомлення, автор, мова, кількість лайків, репостів тощо. На основі цих характеристик вона визначає, наскільки "правдивою" є новина. Для цього в кожному записі обчислюється так званий Truth_Score, який вказує, скільки разів значення з окремих колонок повторюється в інших новинах — тобто наскільки вона схожа на інші, вже відомі повідомлення [15]. Якщо цей бал більший або дорівнює медіанному значенню по всіх новинах, то вважається, що новина є правдивою, і навпаки — якщо нижче, то фейковою.

Після цього дані розділяються на тренувальну та тестову вибірки, нормалізуються (тобто масштаби лайків, репостів і truth score приводяться до одного масштабу), і на них тренуються різні алгоритми машинного навчання — так звані класифікатори [16]. Наприклад, це можуть бути Random Forest, SVM, логістична регресія тощо. Кожен із них навчається розпізнавати закономірності між числовими ознаками (Like, Reposts, Truth_Score) і тим, чи є новина правдивою.

Після навчання кожен класифікатор перевіряється на тестовій вибірці — виводиться його точність і повний звіт по класифікації. Потім у програму "заганяється" нове повідомлення, як вказано в умові: "Україна стала першою серед країн з найвищою смертністю та найнижчою народжуваністю". Воно додається до датасету, для нього обраховується новий Truth_Score, масштабується так само, як і решта даних, і передається кожному з навчених класифікаторів. Кожен з них робить свій прогноз: фейкова ця новина чи правдива. Результати виводяться у консоль. Таким чином, програма дозволяє автоматично оцінювати правдивість новин на основі подібності до вже наявних у базі, використовуючи кілька підходів машинного навчання.

Експеримент та результати

Розроблено програму, яка виконує аналіз мережі новин, які зберігаються у файлі Excel, та візуалізують виявлені зв'язки між ними. Спочатку завантажуються дані про повідомлення, їхніх авторів або групи та джерела, після чого створюється граф, де кожне повідомлення подається як окремий вузол. Зв'язки між вузлами встановлюються на основі спільного автора або джерела, що дозволяє побудувати структуру взаємозв'язків між новинами.

Далі програмний ужиток застосовує алгоритм Лувена [3] для виявлення спільнот або кластерів у графі, тобто груп повідомлень, які тісно пов'язані між собою (див. Рис. 3- Рис.7). Для кожного з кластерів визначається центральне повідомлення — таке, яке має найбільшу кількість зв'язків усередині своєї групи. Окремо аналізуються останні кілька десятків новин: половина з них умовно позначаються як фейкові, а інша половина — як правдиві. Це робиться виключно для спрощення візуалізації, оскільки графік був би перевантаженим через велику кількість записів у датасеті. Для кожної новини виводиться кластер, до якого вона належить, що дозволяє простежити їхнє розташування в загальній структурі.

На завершення програма візуалізує побудований граф. Кожен кластер відображається окремим кольором, центральні повідомлення, фейкові та правдиві новини — виділені різними кольорами рамки. Така візуалізація дозволяє інтуїтивно оцінити структуру мережі новин, їх згрупованість та взаємозв'язки між повідомленнями.

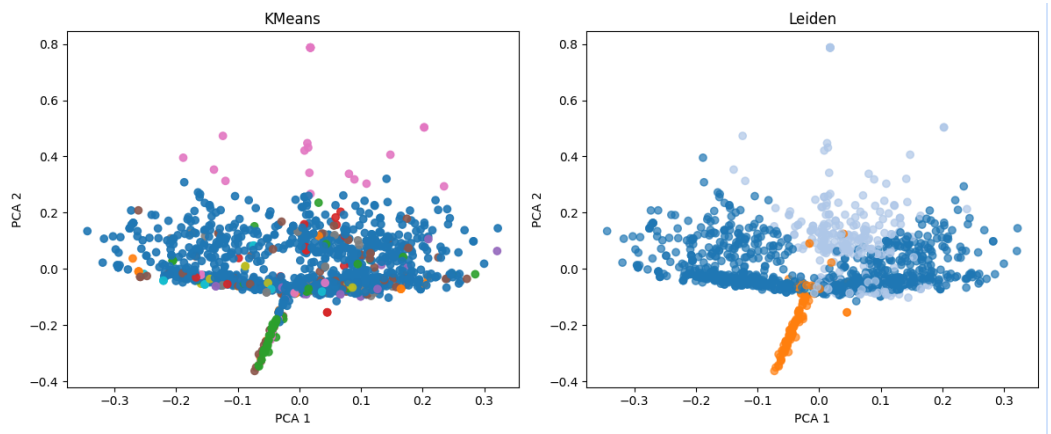

Рис.3 Візуалізація методу KMeans на 65 кластерів та Leiden на 41 кластер

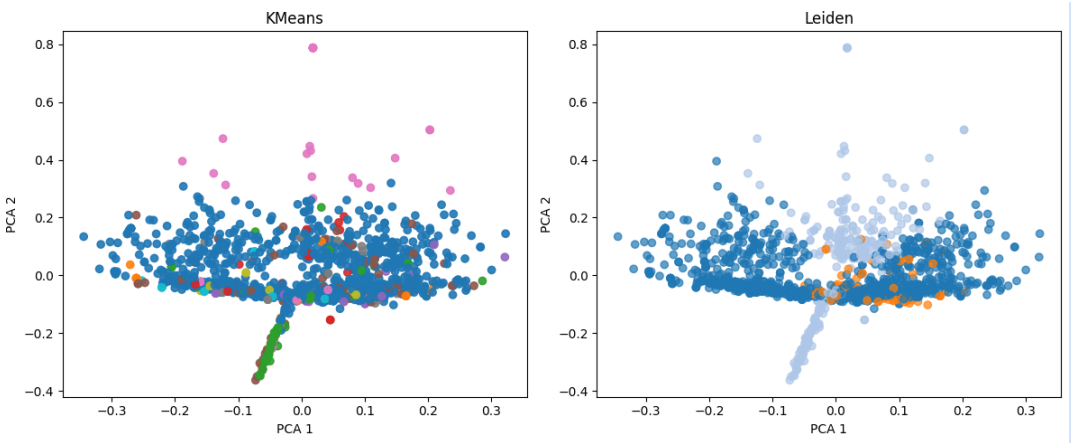

Рис.4 Візуалізація методу KMeans на 65 кластерів та Leiden на 10 кластерів

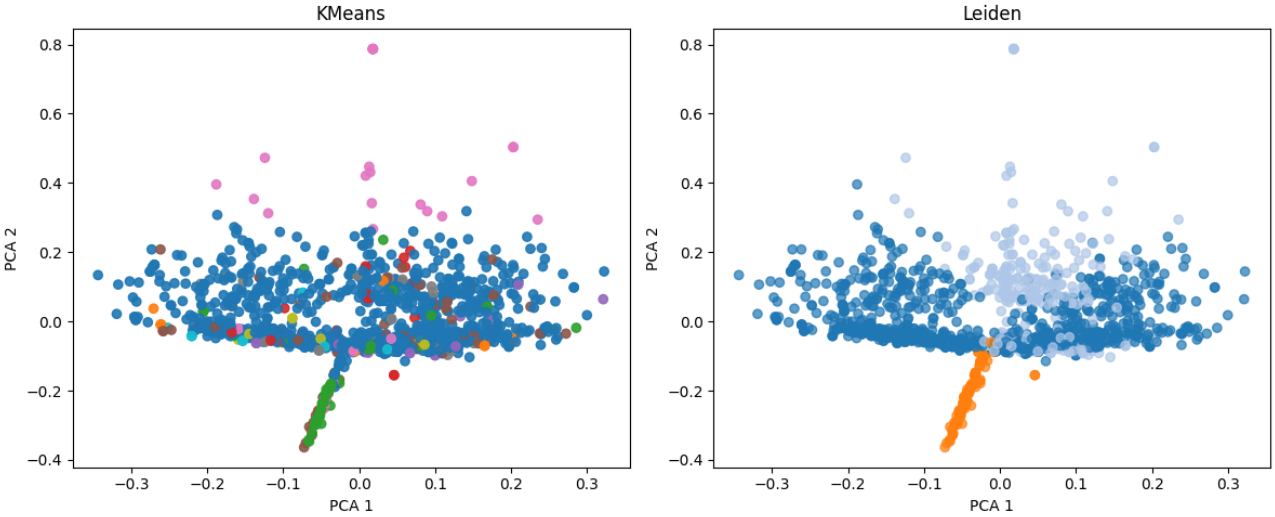

Рис.5 Візуалізація методу KMeans на 65 кластерів та Leiden на 5 кластерів

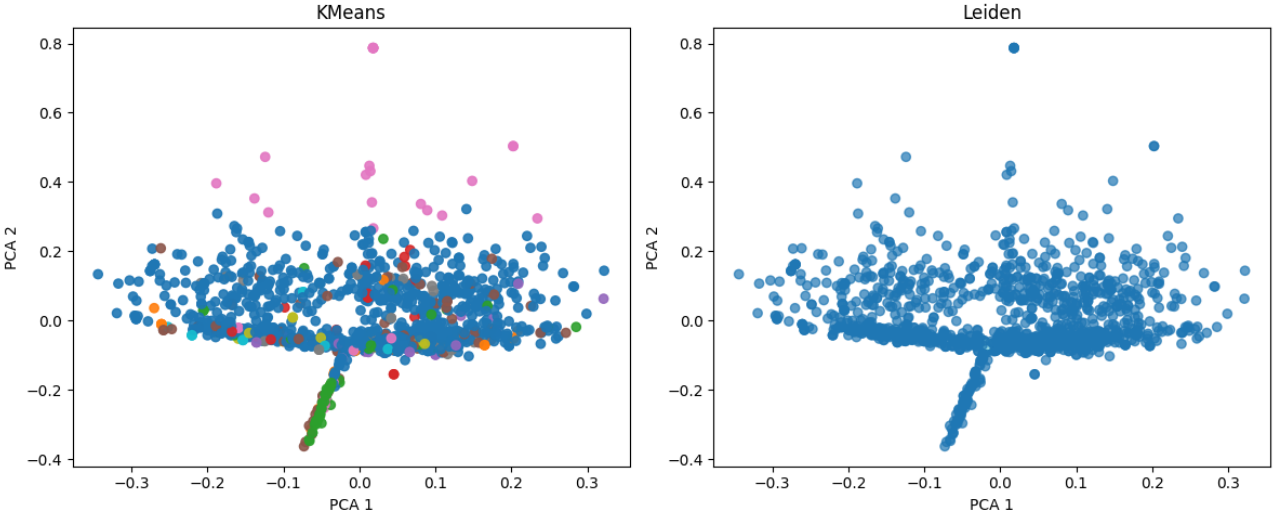

Рис.6 Візуалізація методу KMeans на 65 кластерів та Leiden на 4 кластери

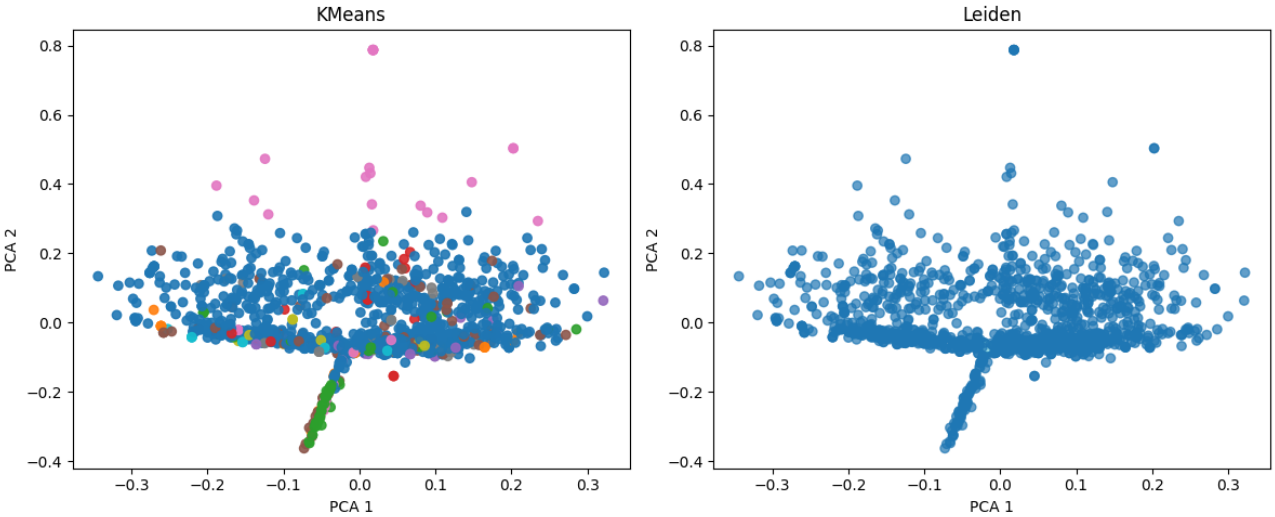

Рис.7 Візуалізація методу KMeans на 65 кластерів та Leiden на 2 кластери

Запропонований метод був реалізований на python. Всі тести виконувалися на комп'ютері з 64-ядерним процесором Intel Xeon E5-4667v3 частотою 2 ГГц та 1 ТБ оперативної пам'яті. Результати представлені на рис. 8.

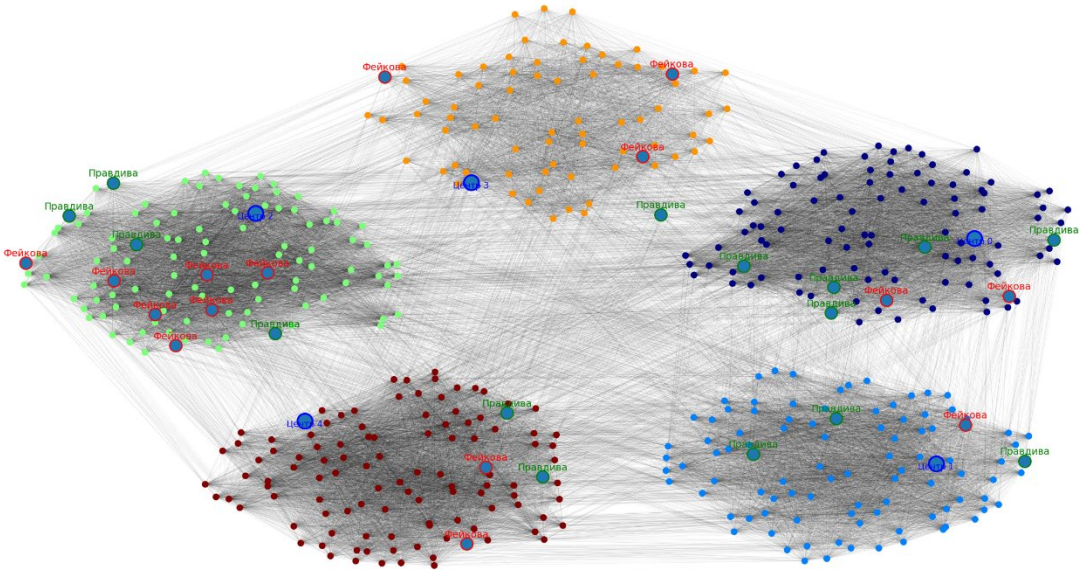

Рис. 8 Граф фейкових новин

Кластери новин з виділеними повідомленнями, які класифіковані як фейкові і не фейкові

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У результаті проведеного дослідження встановлено, що виявлення дезінформації у соціальних мережах потребує підходу, який поєднує методи машинного навчання, глибинних нейронних мереж і аналізу соціальних графів. Показано, що класичні алгоритми машинного навчання SVM, Random Forest забезпечують прийнятну точність для невеликих корпусів даних, однак поступаються сучасним трансформерним моделям BERT, RoBERTa, GPT, здатним враховувати контекст і семантичні зв'язки між елементами тексту.

Важливим напрямом розвитку є використання мультимодального аналізу, який дозволяє одночасно враховувати текстову, візуальну та аудіальну інформацію, що особливо актуально для виявлення складних типів фейків, зокрема deepfake-контенту.

Результати експериментів засвідчили, що структурний підхід до аналізу поширення інформації — через побудову та кластеризацію соціального графа — значно підвищує точність виявлення координованих дезінформаційних кампаній.

Запропоноване застосування методу Лейдена продемонструвало вищу ефективність у порівнянні з алгоритмом Лувена. Отримані показники модульності (0,82) та щільності спільнот (0,74) підтвердили здатність методу Лейдена забезпечувати чіткіше групування акаунтів і виявляти до 78 % бот-мереж, залучених до поширення фейкових повідомлень. Це свідчить про потенціал використання даного підходу для побудови систем раннього виявлення дезінформаційних атак.

ПОДЯКА

Дослідження здійснено завдяки грантової підтримки Національного Фонду Досліджень України, реєстраційний номер проєкту 33/0012 від 3/03/2025 (2023.04/0012) «Розроблення інформаційної системи автоматичного виявлення джерел дезінформації та неавтентичної поведінки користувачів чатів» за конкурсом «Наука для зміцнення обороноздатності України».

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. 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 natural language processing models. *Future Internet, 17*(1).
2. Ji, L., & Xie, H. (2025). A modified method on improving power grid resilience based on the Louvain algorithm. *Sustainable Energy, Grids and Networks*, 101883.
3. Anuar, S. H. H., Abas, Z. A., Yunos, N. M., Zaki, N. H. M., Hashim, N. A., Mokhtar, M. F., ... & Nizam, A. F. (2021, December). Comparison between Louvain and Leiden algorithm for network structure: A review. In *Journal of Physics: Conference Series* (Vol. 2129, No. 1, p. 012028). IOP Publishing.
4. Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports, 9*(1), 1–12.
5. Singlan, N., Abou Choucha, F., & Pasquier, C. (2025). A new similarity-based adapted Louvain algorithm (SIMBA) for active module identification in p-value attributed biological networks. *Scientific Reports, 15*(1), 11360.
6. Ghosh, S., Halappanavar, M., Tumeo, A., Kalyanaraman, A., Lu, H., Chavarria-Miranda, D., ... & Gebremedhin, A. (2018, May). Distributed louvain algorithm for graph community detection. In *2018 IEEE international parallel and distributed processing symposium (IPDPS)* (pp. 885-895). IEEE.
7. Sharma, B., Kathirvel, P., & Hegde, S. R. (2025). Identifying communities in the virus–host protein–protein interaction networks. In *Computational Virology* (pp. 307–319). New York, NY: Springer US.
8. Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008.
9. Doropoulos, S., Karapalidou, E., Charitidis, P., Karakeva, S., & Vologiannidis, S. (2025). Beyond manual media coding: Evaluating large language models and agents for news content analysis. *Applied Sciences, 15*(14), 8059.
10. Li, X. (2025). Translator's centrality and ideology in Chinese-English bilingual news: A case study of texts from the China Daily website. *Arts, Culture and Language, 1*(3).
11. Hernández, R., Gutiérrez, I., & Castro, J. (2025). Social network analysis: A novel paradigm for improving community detection. *International Journal of Computational Intelligence Systems, 18*(1), 87.
12. StopFake. (n.d.). *Про нас.* https://www.stopfake.org/uk/pro-nas/
13. VoxUkraine. (n.d.). *VoxCheck.* https://voxukraine.org/voxcheck
14. Vysotska, V., & Nazarkevych, M. (2025). Development of an information technology for detecting the sources and networks of disinformation dissemination in cyberspace based on machine learning methods. *Eastern-European Journal of Enterprise Technologies, 4.*
15. Danylyk, V., Vysotska, V., & Nazarkevych, M. (2024). Disinformation, fakes and propaganda identification methods in mass media based on machine learning. *Cybersecurity: Education, Science, Technique, 1*(25), 449–467.
16. Vysotska, V., Nazarkevych, M., Vladov, S., Lozynska, O., Markiv, O., Romanchuk, R., & Danylyk, V. (2024). Devising a method for detecting information threats in the Ukrainian cyberspace based on machine learning. *Eastern-European Journal of Enterprise Technologies, 132*(2).

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (30), 2025
Спеціальний випуск ISSN 2663 – 4023
Special issue

Nazarkevych Maria
Doctor of Technical Sciences, Professor,
Professor of the Department of Information Systems and Networks
National University "Lviv Polytechnic", Lviv, Ukraine
ORCID: 0000-0002-6528-9867
Mariia.a.nazarkevych@lpnu.ua

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

Yurynets Rostyslav
Candidate of Physical and Mathematical Sciences, Associate Professor
Associate Professor of the Department of Information Systems and Networks
National University "Lviv Polytechnic", Lviv, Ukraine
ORCID: 0000-0003-3231-8059
rostyslav.v.yurynets@lpnu.ua

Nakonechny Nazar
Postgraduate Student, Department of Information Systems and Networks
National University "Lviv Polytechnic", Lviv, Ukraine
ORCID: 0009-0000-2456-3498
nazar.i.nakonechnyi@lpnu.ua

METHODS OF IMPLEMENTING DISINFORMATION DETECTION IN SOCIAL NETWORKS BASED ON ARTIFICIAL INTELLIGENCE

Abstract. The article considers modern approaches to automated detection of disinformation in social networks using artificial intelligence technologies. The evolution of methods is analyzed - from linguistic analysis of texts and classical machine learning algorithms to deep neural networks and transformative models. It is shown that traditional statistical methods do not provide the necessary accuracy when processing large amounts of data, while models based on CNN, RNN and BERT demonstrate high efficiency due to the ability to take into account context and semantic connections. Special attention is paid to multimodal analysis, which combines text, image and video processing to detect complex types of fakes, in particular deepfake. The use of the Leiden method is proposed as an innovative approach to clustering social graphs, which allows detecting coordinated communities of users spreading disinformation. An experimental study was conducted on data from the Twitter social network, which confirmed the high performance of the Leiden algorithm compared to the Louvain method. The obtained modularity (0.82) and cluster density (0.74) indicators demonstrated a clear structuring of disinformation communities and the ability to detect up to 78% of bot accounts. The developed model combines social graph analysis with natural language processing (NLP) methods to simultaneously identify sources of disinformation and the content of distributed messages. It is concluded that the integration of graph clustering methods and machine learning is a promising direction in creating automatic monitoring systems for social networks. Further research should focus on the development of explainable models (Explainable AI), multilingual adaptation, and the implementation of anti-fake technologies directly into the infrastructure of social platforms.

Keywords: disinformation source detection; fake news; machine learning; social network; Leiden method; Louvain method.

REFERENCES

1. 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 natural language processing models. *Future Internet, 17*(1).
2. Ji, L., & Xie, H. (2025). A modified method on improving power grid resilience based on the Louvain algorithm. *Sustainable Energy, Grids and Networks*, 101883.
3. Anuar, S. H. H., Abas, Z. A., Yunos, N. M., Zaki, N. H. M., Hashim, N. A., Mokhtar, M. F., ... & Nizam, A. F. (2021, December). Comparison between Louvain and Leiden algorithm for network structure: A review. In *Journal of Physics: Conference Series* (Vol. 2129, No. 1, p. 012028). IOP Publishing.
4. Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports, 9*(1), 1–12.
5. Singlan, N., Abou Choucha, F., & Pasquier, C. (2025). A new similarity-based adapted Louvain algorithm (SIMBA) for active module identification in p-value attributed biological networks. *Scientific Reports, 15*(1), 11360.
6. Ghosh, S., Halappanavar, M., Tumeo, A., Kalyanaraman, A., Lu, H., Chavarria-Miranda, D., ... & Gebremedhin, A. (2018, May). Distributed louvain algorithm for graph community detection. In *2018 IEEE international parallel and distributed processing symposium (IPDPS)* (pp. 885-895). IEEE.
7. Sharma, B., Kathirvel, P., & Hegde, S. R. (2025). Identifying communities in the virus–host protein–protein interaction networks. In *Computational Virology* (pp. 307–319). New York, NY: Springer US.
8. Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008.
9. Doropoulos, S., Karapalidou, E., Charitidis, P., Karakeva, S., & Vologiannidis, S. (2025). Beyond manual media coding: Evaluating large language models and agents for news content analysis. *Applied Sciences, 15*(14), 8059.
10. Li, X. (2025). Translator's centrality and ideology in Chinese-English bilingual news: A case study of texts from the China Daily website. *Arts, Culture and Language, 1*(3).
11. Hernández, R., Gutiérrez, I., & Castro, J. (2025). Social network analysis: A novel paradigm for improving community detection. *International Journal of Computational Intelligence Systems, 18*(1), 87.
12. StopFake. (n.d.). *Про нас.* https://www.stopfake.org/uk/pro-nas/
13. VoxUkraine. (n.d.). *VoxCheck.* https://voxukraine.org/voxcheck
14. Vysotska, V., & Nazarkevych, M. (2025). Development of an information technology for detecting the sources and networks of disinformation dissemination in cyberspace based on machine learning methods. *Eastern-European Journal of Enterprise Technologies, 4.*
15. Danylyk, V., Vysotska, V., & Nazarkevych, M. (2024). Disinformation, fakes and propaganda identification methods in mass media based on machine learning. *Cybersecurity: Education, Science, Technique, 1*(25), 449–467.
16. Vysotska, V., Nazarkevych, M., Vladov, S., Lozynska, O., Markiv, O., Romanchuk, R., & Danylyk, V. (2024). Devising a method for detecting information threats in the Ukrainian cyberspace based on machine learning. *Eastern-European Journal of Enterprise Technologies, 132*(2).