КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (33), 2026
ISSN 2663 – 4023
DOI 10.28925/2663-4023.2026.33.1126
УДК 004.89:004.93'1

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

Теплий Ярослав Богданович
аспірант кафедри інформаційних систем та мереж
Національний університет «Львівська політехніка», Львів, Україна
ORCID: 0009-0001-5548-5530
yaroslav.b.teplyi@lpnu.ua

Куриляк Юліан Анатолійович
аспірант кафедри інформаційних систем та мереж
Національний університет «Львівська політехніка», Львів, Україна
ORCID: 0000-0002-6600-2637
yulian.a.kuryliak@lpnu.ua

Торшин Віталій В'ячеславович
аспірант кафедри інформаційних систем та мереж
Національний університет «Львівська політехніка», Львів, Україна
ORCID: 0009-0002-0133-6553
vitalii.v.torshyn@lpnu.ua

ДОСЛІДЖЕННЯ МЕХАНІЗМІВ ВИЯВЛЕННЯ ДЖЕРЕЛ ТА ШЛЯХІВ РОЗПОВСЮДЖЕННЯ ФЕЙКОВИХ НОВИН І ПРОПАГАНДИ КІБЕРПРОСТОРІ СОЦІАЛЬНИХ МЕРЕЖ

Анотація. У статті представлено інформаційну технологію виявлення джерел та маршрутів поширення фейкових новин і пропаганди в соціальних мережах та онлайн-медіа. Запропонований підхід ґрунтується на векторизації текстових публікацій за допомогою мовних моделей, обчисленні косинусної схожості та кластеризації повідомлень для виявлення тематично подібних груп. На основі сформованих кластерів будується хронологічна графова візуалізація, що дозволяє ідентифікувати первинні джерела дезінформації, ключових ретрансляторів і швидкість її поширення. Експериментальне дослідження проведено на зведеному датасеті обсягом понад 2000 публікацій, у якому частки правдивих і фейкових повідомлень є порівняно збалансованими. Аналіз взаємодії показав, що близько 75 % усіх публікацій отримують не більше 20 вподобань, тоді як лише менше 5 % повідомлень формують «довгий хвіст» із сотнями та тисячами реакцій. При цьому дезінформаційні повідомлення частіше або залишаються майже непоміченими, або різко набирають аномально високу популярність за короткий проміжок часу. Аналіз поширень виявив, що приблизно 80 % постів мають не більше 5 репостів, однак серед найбільш поширюваних повідомлень частка фейкових суттєво зростає. Платформний аналіз показав, що на веб-ресурсах і авторитетних медіа переважає достовірний контент, тоді як у соціальних мережах та месенджерах (зокрема Telegram) співвідношення правдивих і фейкових повідомлень є близьким до рівноважного, а в окремих мережах домінує дезінформація. Порівняння моделей ембедінгів засвідчило, що модель OpenAI забезпечує чіткіше розділення повідомлень у просторі ознак та дозволяє виявити до 10-11 значущих кластерів при оптимальному порозі косинусної схожості $\tau \approx 0{,}55\text{-}0{,}60$, тоді як для локальної моделі оптимальний поріг є значно вищим ($\tau \approx 0{,}86\text{-}0{,}88$). Побудовані графи поширення показали, що часовий інтервал між первинною публікацією фейку та його появою на інших

платформах може становити від кількох годин до кількох діб, а трансформація репостів в «оригінальні» публікації є типовим механізмом приховування джерела.

Ключові слова: текстові ембедінги, косинусна схожість, кластеризація, аналіз поширення інформації, графова візуалізація, інформаційна безпека.

ВСТУП

Стрімке зростання обсягів інформації в соціальних мережах та онлайн-медіа суттєво ускладнило контроль за якістю й достовірністю публікацій. Поряд із перевіреним контентом все ширше розповсюджуються фейкові новини та пропагандистські повідомлення, які можуть маніпулювати суспільною думкою, провокувати паніку та використовуватися як інструмент інформаційних атак. Особливу небезпеку становить висока швидкість і масштаб поширення дезінформації, а також можливість її багаторазового копіювання, переформулювання та повторної публікації на різних платформах із втратою посилань на первинне джерело. Існуючі підходи до виявлення фейкових новин переважно зосереджені на класифікації окремих повідомлень за ознакою достовірності, що не дозволяє повною мірою відстежити процес їх розповсюдження та встановити початкові джерела появи дезінформаційних тез. Крім того, значна частина методів ґрунтується на ручній модерації або статичних правилах, які є малоефективними в умовах великих потоків даних, багатомовності контенту та постійної трансформації текстів. Таким чином, актуальною науково-практичною проблемою є розроблення інформаційної технології, здатної автоматично аналізувати великі масиви публікацій, виявляти тематично подібні повідомлення, визначати маршрути їх поширення у часі та просторі, а також ідентифікувати первинні джерела і ключові вузли ретрансляції дезінформації. Розв’язання цієї проблеми є необхідним для підвищення ефективності моніторингу інформаційного простору та раннього виявлення інформаційних загроз.

Постановка проблеми. В умовах цифровізації суспільства та глобалізації інформаційного простору соціальні мережі й онлайн-медіа стали основними каналами формування громадської думки. Платформи на кшталт Facebook, X (Twitter), Telegram та інші забезпечують миттєве поширення повідомлень серед мільйонів користувачів. Разом із позитивними можливостями швидкої комунікації це створює сприятливе середовище для масштабного розповсюдження фейкових новин, маніпулятивного контенту та координованих пропагандистських кампаній.

Особливої актуальності проблема набуває в умовах гібридних конфліктів та інформаційних протистоянь, коли дезінформація використовується як інструмент впливу на суспільні настрої, політичні процеси та міжнародну репутацію держав. Фейкові повідомлення можуть поширюватися лавиноподібно, багаторазово копіюватися, частково змінюватися або маскуватися під нові оригінальні публікації, що ускладнює встановлення їхнього першоджерела. Часто репости трансформуються в нібито «самостійні» дописи без посилання на джерело, що розриває ланцюг поширення та унеможливлює просте трасування інформаційного потоку.

Існуючі підходи до виявлення фейкових новин переважно зосереджені на задачі класифікації окремих повідомлень як «правдивих» або «недостовірних» із використанням методів машинного навчання. Проте такий підхід не дозволяє відтворити повну картину інформаційної динаміки: визначити часову послідовність поширення, виявити ключові вузли ретрансляції, встановити первинні точки входу дезінформаційної тези в інформаційний простір та оцінити швидкість міжплатформної міграції контенту. Крім того, значна частина методів ґрунтується на ручній модерації, статичних правилах або обмежених наборах ознак, що є недостатньо ефективним в умовах великих обсягів багатомовних даних і постійної трансформації текстів. Таким чином, виникає науково-практична проблема розроблення комплексної інформаційної технології, здатної:

- автоматично аналізувати великі масиви текстових публікацій з різних платформ;
- виявляти семантично подібні повідомлення, включно з переформульованими копіями;
- формувати кластери дезінформаційних тез;
- реконструювати хронологічні маршрути їх поширення;
- ідентифікувати первинні джерела та ключових ретрансляторів.

Розв’язання цієї проблеми потребує поєднання сучасних методів векторного подання тексту, обчислення семантичної подібності, кластеризації та графового моделювання інформаційних потоків. Особливої уваги потребує питання вибору оптимальних параметрів подібності та оцінювання ефективності різних embedding-моделей у задачах групування повідомлень і виявлення дезінформаційних каскадів.

Отже, актуальність дослідження зумовлена необхідністю створення інструментарію, який дозволить не лише ідентифікувати окремі фейкові повідомлення, а й відтворювати структуру та логіку їх розповсюдження в цифровому середовищі, що є критично важливим для систем інформаційної безпеки, OSINT-аналітики та моніторингу інформаційних загроз.

Аналіз останніх досліджень і публікацій. Сучасне інформаційне середовище, особливо у воєнний та поствоєнний періоди, дедалі більше піддається впливу дезінформаційних кампаній. Швидке поширення неправдивої інформації через соціальні мережі та онлайн-платформи створює суттєві загрози для безпеки суспільства, впливає на громадську думку та ускладнює процеси прийняття рішень. У зв'язку з цим зростає потреба у створенні автоматизованих інструментів, здатних виявляти джерела дезінформації, аналізувати маршрути її поширення та визначати ключових ретрансляторів фейкових повідомлень. Актуальність дослідження зумовлена необхідністю підвищення інформаційної безпеки в умовах гібридної війни, коли інформація використовується як інструмент впливу. Існуючі підходи здебільшого зосереджені на фактологічній перевірці повідомлень, однак не дають змоги автоматично відстежувати шлях поширення тези від її первинного джерела до аудиторії. Тому дослідження, спрямоване на виявлення хронології та динаміки розповсюдження фейкових постів, має як наукову, так і практичну значущість. В науковому світі основні напрямки подібних досліджень зосередженні на застосуванні відповідних методів та технологій, зокрема:

- Графові методи для виявлення дезінформації;
- Візуалізація та трасування маршрутів поширення;
- Методи векторизації та кластеризації;
- Моделі поширення (епідемічні);
- Аналітика користувацьких зв'язків.

В [1] сформована мета-графова структура, що об'єднує події, семантику та топіки. Застосування GNN дало приріст точності на 3-4 % проти чистої каскадної класифікації. В [2] пропонують геометричне глибоке навчання (Graph-CNN) для аналізу інформаційних каскадів на Twitter – ROC AUC ≈92.7 %, досягаючи раннього виявлення «фейкових» новин. В [3] здійснено огляд графових нейронних мереж для fake-news detection, класифікація методів за контекстом, пропагандою та соціальним контекстом. В [4] описано інструмент для трейсингу дезінформації через графи, боти, ретвіти, URL-мережі тощо. Він використовує Gephi, NodeXL, Hoaxy для аналізу структур і виявлення координованих взаємодій. В [5] здійснено поєднання BERTopic та GNN для візуалізації топікових каскадів та інтерпретованого виявлення фейк-новин. У [6] на прикладі Weibo здіснена семантична векторизація (TF-IDF з dense embeddings) із кластеризацією K-means для виявлення фейкових тем. У [7] описано використання мережевої ембедінга (наприклад, node2vec) для аналізу сумнівів щодо вакцин – приклад комбінованого контенту та структурного аналізу. SIR-based, Independent Cascade та інші є класичними моделями розповсюдження "вірус-подібної" інформації застосовуються для аналізу поширення чуток через соціальні мережі [8]. У [9] здійснено аналіз ключових метрик (центральність, ефективність) в дезінформаційній мережі – виявлено «лідерів розмов» із високою eigenvector-centrality. У [10] здійснено графове векторне моделювання користувачів (спільні фоловери/дружба) якісно додає до content-based фейк-детекції. У [11] описано реальні кейси аналізу і виявлення дезінформаційних кампаній (русських IRA, китайського антивакцинного мистецтва) через мережеве картографування. Застосування графів з GNN є ключем до точного виявлення фейкових каскадів (точність ~90 %+). Візуалізація каскадів (БерТopic, Gephi, Hoaxy) допомагає аналізувати маршрути і типові структура. Векторизація з кластеризацією (TF-IDF, dense embeddings, cosine) є основою для розпізнавання груп схожих повідомлень. Епідемічні та структурно-мережеві моделі допомагають моделювати динаміку поширення. Соціальні зв'язки користувачів є додатковою змінною для кращого контекстного розуміння. Наш підхід поєднує кілька ключових стратегій як векторизація [6], кластеризація за cosine similarity, графова візуалізація поширення [4-5] та мережевий контекст за допомогою GNN-підходів [1-3]. Ця комбінація згідно з літературою є перспективною. Вона синтезує переваги кластерних, графових та візуалізаційних методів для глибокого аналізу інформаційних потоків.

Метою роботи є розробка та реалізація інформаційної системи для автоматизованого аналізу повідомлень у соціальних мережах із можливістю виявлення дезінформаційних кластерів, визначення маршрутів їх розповсюдження та візуалізації отриманих даних. Об'єктом дослідження виступають текстові публікації в соціальних мережах та онлайн-ЗМІ. Предметом дослідження є методи векторного подання тексту, обчислення семантичної подібності та кластеризації повідомлень з метою аналізу інформаційних потоків. Поставлені завдання дослідження:

1. Розробити архітектуру системи для обробки та аналізу масивів соціальних повідомлень;

2. Реалізувати алгоритми векторизації тексту з використанням локальної та хмарної embedding-моделі;
3. Запропонувати метод кластеризації повідомлень за рівнем косинусної подібності;
4. Побудувати графи розповсюдження фейкових повідомлень з візуалізацією хронології;
5. Провести експериментальну перевірку системи на реальних датасетах.

Наукова новизна дослідження полягає у поєднанні векторного групування публікацій з побудовою хронологічних графів поширення інформації, що дозволяє автоматично виявляти не лише схожі повідомлення, а й логіку та швидкість їх розповсюдження. Запропоновано підхід до візуалізації дезінформаційних кластерів, що дозволяє виявляти так звані «нульові» точки входу фейків та способи приховування їх джерел. Практична цінність роботи полягає у створенні дієвого інструменту для фахівців у галузі інформаційної безпеки, журналістських розслідувань та OSINT-аналітики. Результати дослідження можуть бути інтегровані у системи моніторингу інформаційного простору з метою оперативного реагування на спроби маніпуляцій або дезінформаційних атак.

ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

Теоретичною базою дослідження є міждисциплінарне поєднання положень теорії інформаційної безпеки, аналізу соціальних мереж, обробки природної мови (NLP), теорії графів та методів машинного навчання [9-12]. Запропонована інформаційна технологія ґрунтується на сучасних концепціях семантичного моделювання тексту, мережевого аналізу та моделювання інформаційних каскадів.

У межах дослідження під дезінформацією розуміється навмисно створена або модифікована неправдива чи маніпулятивна інформація, спрямована на введення аудиторії в оману з певною метою (політичною, економічною, соціальною).

Фейкова новина – це окремий інформаційний об’єкт (повідомлення, пост, стаття), що містить неправдиві або спотворені відомості та може поширюватися через цифрові платформи. В умовах функціонування соціальних мереж, таких як Facebook, X та Telegram, дезінформація здатна формувати так звані інформаційні каскади – послідовності публікацій, пов’язаних спільною тезою та часовою динамікою поширення. В основі дослідження лежить положення про те, що інформаційна атака має не лише змістову, а й структурну природу, тобто характеризується:

- семантичною подібністю повідомлень;
- часовою послідовністю появи;
- мережевими зв’язками між джерелами;
- повторюваністю та варіативністю формулювань.

Ключовим підходом дослідження є векторна семантична модель тексту, згідно з якою кожне повідомлення перетворюється у числовий вектор фіксованої розмірності (embedding), що відображає його зміст у багатовимірному просторі ознак. Теоретичною основою цього підходу є:

- розподільна гіпотеза (distributional hypothesis), відповідно до якої слова з подібними контекстами мають подібні значення;
- трансформерна архітектура мовних моделей;
- концепція латентного семантичного простору.

У дослідженні використано дві моделі ембедінгів:

- text-embedding-3-small – хмарна модель OpenAI;
- intfloat/multilingual-e5-base – локальна багатомовна модель.

Embedding – це відображення тексту у векторному просторі $\mathbb{R}^d$, де семантично близькі повідомлення розташовуються на меншій відстані одне від одного.

Для кількісної оцінки близькості повідомлень використовується косинусна подібність – міра кута між двома векторами. Косинусна подібність обрана через такі теоретичні переваги:

- інваріантність до масштабу векторів;
- ефективність у високовимірних просторах;
- широку апробацію в задачах інформаційного пошуку та NLP.

Порогове значення τ визначає, чи вважаються два повідомлення семантично подібними. Вибір τ є критичним параметром моделі та впливає на структуру отриманих кластерів. Після визначення семантичної подібності повідомлень формується граф подібності, у якому: вершини – окремі публікації; ребра – зв’язки між повідомленнями з $S \geq \tau$. Теоретично дослідження спирається на базові положення теорії графів, зокрема:

- поняття компоненти зв’язності;
- орієнтовані та неорієнтовані графи;

- часові графи (temporal graphs);
- метрики центральності.

Кластер у межах дослідження визначається як компонента зв'язності графа подібності, що складається щонайменше з двох повідомлень. Подальша побудова хронологічного графа поширення базується на впорядкуванні вершин за часовими мітками та аналізі міжплатформної міграції контенту. Теоретичною основою аналізу маршрутів є модель інформаційного каскаду – процесу, в якому ідея або повідомлення поширюється від одного вузла мережі до інших. Каскад характеризується:

- початковою точкою (первинне джерело);
- ланцюгом ретрансляцій;
- часовими інтервалами між публікаціями;
- глибиною та шириною поширення.

У межах дослідження каскад реконструюється шляхом:

1. виявлення семантично подібних повідомлень;
2. сортування їх за часом;
3. формування орієнтованих зв'язків між послідовними публікаціями.

Кластеризація в дослідженні базується на принципі групування об'єктів за мірою подібності без попередньої розмітки класів (unsupervised learning). Застосований підхід передбачає:

- побудову матриці попарної подібності;
- формування графа при порозі τ;
- виділення компонент зв'язності як кластерів.

Такий підхід дозволяє виявляти:

- прямі копії;
- переформульовані повідомлення;
- модифіковані варіанти однієї тези.

У межах роботи використовуються такі ключові поняття:

- інформаційний простір – сукупність цифрових платформ і каналів комунікації;
- інформаційний потік – послідовність пов'язаних повідомлень;
- первинне джерело – перша за часом публікація тези;
- ретранслятор – вузол мережі, що повторно поширює інформацію;
- кластер повідомлень – група семантично подібних публікацій;
- embedding-модель – модель перетворення тексту у векторний простір;
- порог подібності τ – граничне значення косинусної міри для встановлення зв'язку.

Дослідження спирається на такі принципи:

- Принцип семантичної близькості – змістовно споріднені повідомлення мають близькі векторні представлення.
- Принцип часової причинності – первинне джерело передує ретрансляціям.
- Принцип мережевої структури поширення – дезінформація поширюється не хаотично, а через конкретні вузли мережі.
- Принцип міжплатформної взаємодії – інформаційні каскади можуть переходити між різними соціальними платформами.
- Принцип масштабованості – технологія має працювати з великими масивами даних.

Таким чином, теоретичні основи дослідження формуються на перетині теорії інформаційної безпеки, методів обробки природної мови, векторної семантики, теорії графів, аналізу інформаційних каскадів. Саме поєднання цих концепцій дозволяє перейти від ізольованого визначення фейкових повідомлень до комплексного аналізу механізмів їх виникнення та поширення в цифровому інформаційному середовищі.

МЕТОДИКА ДОСЛІДЖЕННЯ

Ціллю статті є розроблення та експериментальне дослідження інформаційної технології автоматизованого виявлення джерел і маршрутів поширення фейкових новин та пропагандистських повідомлень у соціальних мережах і онлайн-медіа. Запропонована технологія спрямована на аналіз великих масивів текстових публікацій різного походження, їх семантичне порівняння та групування з метою реконструкції хронологічного ланцюга розповсюдження дезінформаційних тез.

Для досягнення поставленої цілі у роботі передбачається: застосування сучасних мовних моделей для перетворення текстових повідомлень у векторні подання; використання метрик семантичної подібності для виявлення схожих за змістом публікацій, зокрема тих, що є копіями або

переформульованими варіантами одного й того ж повідомлення; побудова кластерів тематично близьких постів та подальше формування графових моделей, які відображають часову послідовність і міжплатформні маршрути поширення інформації. Окремою складовою цілі є порівняльний аналіз ефективності локальної багатомовної моделі та хмарної моделі OpenAI для задачі векторизації текстів, а також визначення оптимальних параметрів косинусної схожості, що забезпечують найкраще розділення інформаційних потоків на смислово однорідні групи. Крім того, стаття має на меті дослідити взаємозв'язок між семантичною подібністю повідомлень і показниками взаємодії користувачів (вподобаннями та поширеннями) з метою підвищення точності виявлення потенційних інформаційних атак. Досягнення сформульованої цілі дозволяє створити практичний інструмент для аналітиків та дослідників інформаційної безпеки, який забезпечує не лише ідентифікацію фейкових повідомлень, а й глибше розуміння механізмів їх виникнення та поширення в цифровому інформаційному середовищі.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Система аналізує масив публікацій у соціальних мережах/онлайн-ЗМІ, знаходить схожі за змістом повідомлення, групує їх у кластери та візуалізує хронологічні «маршрути» поширення кожної дезінформаційної тези. Таким чином можна ідентифікувати первинні джерела, ключові ретранслятори та темп розповсюдження фейків.

Таблиця 1

Структура програми

Рівень	Опис	Ключові модулі
Попередня обробка даних	Завантаження таблиці з постами, очищення, перетворення дат та типів	utils.load_data()
Векторизація тексту	Отримання embeddings для кожного повідомлення через локальну модель або OpenAI API	utils.get_embedding()
Обчислення схожості	Косинусна подібність між embedding-ами, формування матриці	sklearn.metrics.pairwise.cosine_ similarity(у ноутбуці)
Кластеризація	Побудова графа за порогом схожості, пошук компонент зв'язності	utils.clusters_from_similarity()
Візуалізація	Граф-таймлайн згрупованих постів за подібністю	visualisation.timeline_graph()

Очікується Excel-файл з мінімальним набором стовпців:

- date, time – дата й час публікації;
- text – повний текст поста;
- language – назва мови;
- post/repost – оригінал (post) чи репост (repost);
- author/group – автор або група-розповсюджувач;
- source – платформа / канал;
- web-address – URL публікації;
- likes, shares – лічильники реакцій;
- flag – мітка достовірності (True - правдивий факт, False - фейк).
- utils.load_data() переформатовує дати (to_timestamp()), нормалізує булеві позначки (transform_bool()), типізує числові колонки та видаляє службові стовпці.

Алгоритм виявлення маршрутів:

1. Векторизація, тобто для кожної публікації створюється масив embeddings розміром d – або локальною моделлю «intfloat/multilingual-e5-base», або через OpenAI («text-embedding-3-small»).
2. Матриця схожості, зокрема, обчислюємо косинусну схожість S між усіма embedding-ами.
3. Поріг τ: зв'язуємо вершини i, j ребром, якщо $S_{ij} > \tau$ (типово $\tau = 0{,}9$).
4. Знаходимо кластери, тобто знаходимо компоненти зв'язності графа (множини самоподібних постів). Вибираємо кластери, де ≥ 2 постів.
5. Знаходимо маршрут, зокрема усередині кожного кластера вершини сортуємо за datetime; послідовно з'єднані ребра показують хронологію поширення.

Якість груп залежить від порогу τ і обраної embedding-моделі. τ = 0.9 для моделі intfloat/multilingual-e5-base і τ = 0.7 для моделі OpenAI text-embedding-3-small.

Після об'єднання трьох вихідних таблиць (1, 2 та 3) ми отримали єдиний датасет із трохи більш як двох тисяч публікацій. Кожен запис зберігає інформацію про текст, автора, дату-час, платформу й кількісні метрики взаємодії (вподобання та поширення), а також прапорець flag, що позначає достовірність. Хоч кількість «правдивих» та «фейкових» дописів виявилась порівняно збалансованою, їхнє «поведінкове» розподілення помітно різниться. На діаграмі зображеній на рис. 1 порівнюється кількість «правдивих» (синій стовпчик) і «фейкових» (червоний стовпчик) публікацій для кожного автора або каналу, які мають щонайменше 10 повідомлень у зведеному наборі. Вісь X перелічує авторів, вісь Y показує абсолютну кількість постів.

- Ліва частина графіка зосереджує авторів, у яких переважають сині стовпчики: це сторінки, де більшість контенту позначена як достовірний.
- У правій частині видно групу авторів, для яких домінують червоні стовпчики, тобто в їхньому потоці публікацій більша частка позначена як недостовірна.

Між цими «полюсами» містяться кілька акаунтів зі змішаною картиною, де сині та червоні стовпчики співіснують у порівняно близьких пропорціях.

Загалом діаграма демонструє, що деякі джерела систематично публікують переважно перевірений контент, інші – переважно сумнівний, а частина має змішаний профіль.

На гістограмі зображеній на рис. 2 добре видно, що переважна більшість повідомлень (приблизно три чверті) збирає не більше ніж двадцять «лайків». Розподіл має довгий, різко правобічний хвіст: лише поодинокі пости досягають кількох сотень реакцій, а в найвіддаленішому «кошику» з акуратним кроком у тисячу ми бачимо штучний «шпиль» – це група яка відповідає за пости з кількостю лайків більше 1 000 (зроблено з метою стиснення графіку, оскільки було багато груп з невеликою кількістю постів у діапазоні 1-100 тис. лайків).

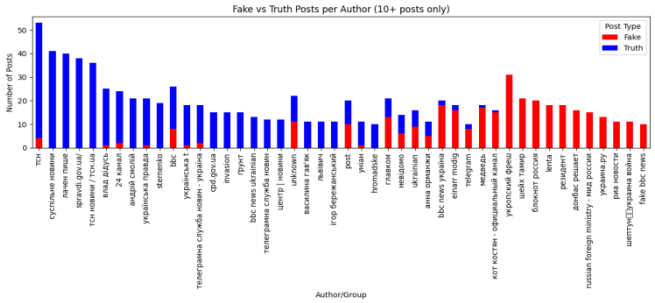


Рис. 1. Стовпчикова діаграма фракції правдивих до неправдивих публікацій за авторами. Відображено авторів які зробили більше 10 публікацій.

Важливо, що синя (truth) і червона (fake) складові у першому «стовпчику» різняться набагато, тоді як далі, у зоні від сорока до трьохсот реакцій, переважає синій колір. Це сигналізує: достовірні новини частіше набирають «середню» популярність, тоді як дезінформаційні здебільшого або лишаються майже непоміченими, або миттєво «вистрілюють» і зникають.

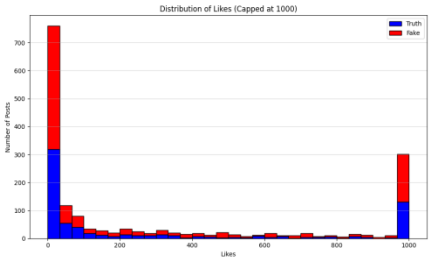


Рис. 2. Гістограма розподілу кількості публікацій за кількістю лайків. Усі пости які мали більше 1000 лайків були віднесені до одної групи(останній стовпчик).

Окрема діаграма стовпчикова діаграма зображена на рис. 3 показує сукупну кількість вподобань за кожним автором чи каналом. Картинка підтверджує інтуїцію: медійні гіганти на кшталт BBC чи офіційного аккаунта facebook акумулюють майже виключно «сині» (правдиві) лайки й на порядок випереджають інших. Натомість серед «червоних» помітні нішеві Telegram-канали й приватні сторінки, які орієнтуються не стільки на регулярний потік контенту, скільки на окремі вірусні пости. Зазвичай такі канали мають свою цільову аудиторію і постять або переважно правду, або переважно фейки.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (33), 2026
ISSN 2663 – 4023

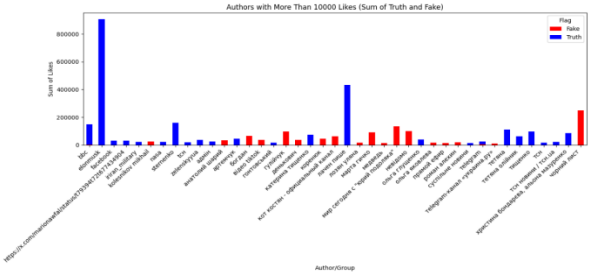


Рис. 3. Стовпчикова діаграма, яка відображає кількість та пропорції лайків для правдивих та фейкових публікацій, згрупованих по авторах. Відображено лише авторів які в сумі мають більше 10000 лайків.

Кілька акаунтів – наприклад, «відео тік-ток» чи «lozan lusher» – демонструють змішану картину: поряд із легітимним матеріалом вони публікують і потенційно маніпулятивний. Це саме ті «сірі зони», де алгоритм виявлення маршрутів фейків виявляється особливо корисним, допомагаючи відстежити, з яких джерел автор запозичує спірні тези. Узагальнюючи, популярність окремих публікацій розподіляється надзвичайно нерівномірно, а дезінформація або губиться в інформаційному шумі, або ж отримує коротке, але різке «підживлення» лайками. Водночас автори з усталеною репутацією здобувають більшу й стабільнішу аудиторію саме на достовірних матеріалах. Це спостереження підтверджує, що механізм кластеризації за текстовою схожістю доцільно доповнювати метриками взаємодії: висока аномальна популярність разом з «червоним» прапорцем може слугувати раннім індикатором інформаційної атаки. Коли ми переходимо від «лайків» до показника поширень, картина виглядає ще контрастнішою. Гістограма зображена на рис. 4 підтверджує, що приблизно 80 % усіх публікацій отримують не більше ніж п’ять перепостів, причому серед таких «малопомітних» повідомлень частка фейкових навіть трохи більша, аніж достовірних. У правому хвості розподілу знову проявляється кластер зі згрупованих публікацій які отримали більше 1000 поширень.

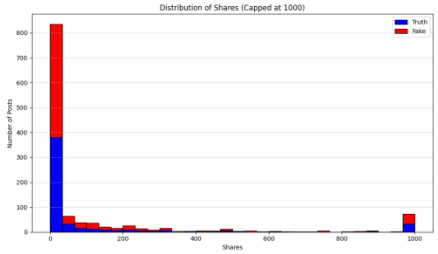


Рис. 4. Гістограма розподілу кількості публікацій за кількістю поширень. Усі пости які мали більше 1000 поширень були віднесені до одної групи(останній стовпчик).

Стовпчикова діаграма зображена на рис. 5 показує загальну кількість публікацій, які зробив кожен з авторів. Лідерами, як і у випадку з лайками, стають міжнародні та наукові медіа (BBC, Science Alert, National Geographic, Nature), а також обліковий запис Ілона Маска – у всіх цих випадках бачимо суцільно синій графік, тобто вміст із високим рівнем довіри.

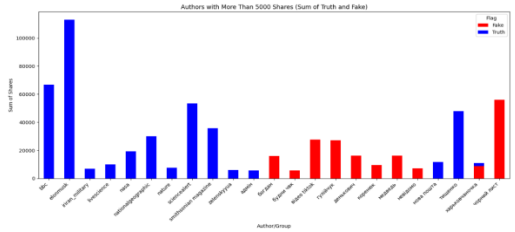


Рис. 5. Стовпчикова діаграма, яка відображає кількість та пропорції поширень для правдивих та фейкових публікацій, згрупованих по авторах. Відображено лише авторів які в сумі мають більше 5000 поширень.

Дещо інша картина в «червоному таборі». Тут з’являються нішеві чи напів анонімні сторінки – наприклад, «Гулийчук» або «чорний лист», які сумарно отримують десятки тисяч перепостів, майже не змішуючи достовірний контент з недостовірним. Тобто аудиторія таких майданчиків поширює насамперед сумнівну інформацію, і саме вони становлять ключову загрозу для інформаційного простору.

На діаграмі зображеній на рис. 6 стовпчики відображають сумарні вподобання, які зібрали пости з різних платформ або доменів, якщо їх назбиралося понад 10 000 «лайків».

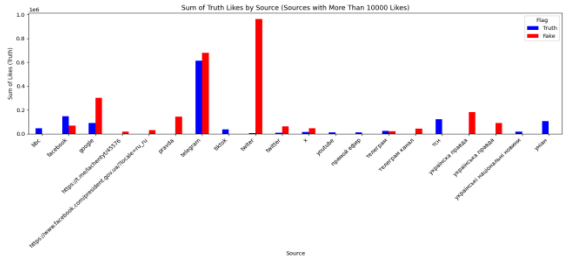


Рис. 6. Стовпчикова діаграма, яка відображає кількість та пропорції лайків для правдивих та фейкових публікацій, згрупованих за платформами. Відображено лише платформи які в сумі мають більше 10000 лайків.

Перевага синього кольору вказує на джерела, де головним чином взаємодіють з неправдивими матеріалами (наприклад, Twitter або окремі новинні сайти). Червоні сегменти показують, що на деяких ресурсах суттєву частку популярності отримують і повідомлення з ознаками дезінформації. Чітко програлаються тенденції, що одні платформи використовують для публікації переважно правдиві пости, а інші – фейкові. Виключенням є телеграм, у якому пропорції фейкових та правдивих постів є приблизно рівними. Тотальна перевага фейкових публікацій у мережі Twitter швидше за все пов'язано незбалансованістю датасету.

На рис. 7 аналогічно підсумовано публікації (тільки для джерел із понад 1000 поширень). Блакитні стовпчики демонструють платформи, на яких активніше поширюють перевірений контент; червоні - ті, де значну кількість репостів отримують сумнівні публікації. Звертає на себе увагу контраст між великою кількістю правдивих перепостів з окремих сайтів/агрегаторів і високими значеннями фейкових публікацій у деяких соціальних мережах.

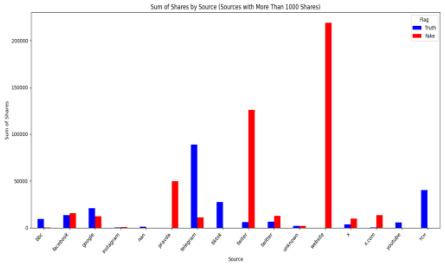


Рис. 7. Стовпчикова діаграма, яка відображає кількість та пропорції поширень для правдивих та фейкових публікацій, згрупованих за платформами. Відображено лише платформи які в сумі мають більше 1000 поширень.

На рис. 8 згруповані стовпчики деталізують, скільки саме оригінальних постів та репостів (поділено за flag) надійшло з популярних джерел. Для кожного джерела показано дві категорії - post та repost. Видно, що на низці платформ (зокрема у Telegram) репостів помітно більше за оригінальних публікацій, причому серед них значна частина позначена як фейкова; натомість на веб-сайтах та низці новинних порталів переважають правдиві оригінальні матеріали.

Далі побудуємо стовпчикові діаграми рис. 9 та рис. 10, на яких кожен стовпчик відповідає окремому кластеру повідомлень, а висота синьої та червоної частини показує, скільки в цьому кластері виявилося підтверджених і недостовірних постів. Уздовж осі X проставлені умовні номери груп, що їх згенерував алгоритм при «оптимальному» порозі схожості, підібраному окремо для кожної моделі. У випадку з OpenAI-ембедінгами є кластери які більш згруповані, проте все ще є зовсім невеликі групи з двох-трьох повідомлень. Здебільшого у всіх кластерах сині й червоні сегменти змішані приблизно порівну. Таке співвідношення натякає, що для більшості правживих постів, зазвичай появиться спотворена копія.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (33), 2026
ISSN 2663 – 4023

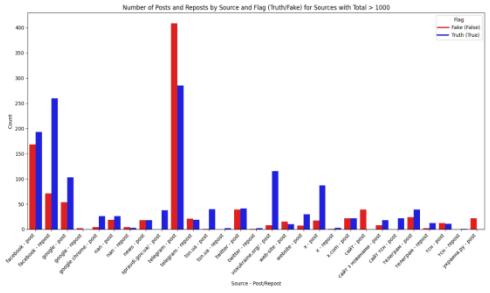


Рис. 8. Стовпчикова діаграма, яка відображає кількість та пропорції поширень для правдивих та фейкових постів та репостів згрупованих за платформами. Відображено лише платформи які в сумі мають більше 1000 публікацій.

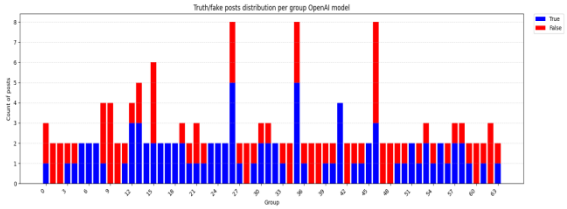


Рис. 9. Стовпчикова діаграма, яка відображає кількість та пропорції для правдивих та фейкових постів, згрупованих за подібністю за OpenAI ембедінгами.

Для локальної моделі кластери загалом менші: майже всі вкладаються у діапазон до шести постів. Розподіл синього та червоного кольору помітно так само рівномірний.

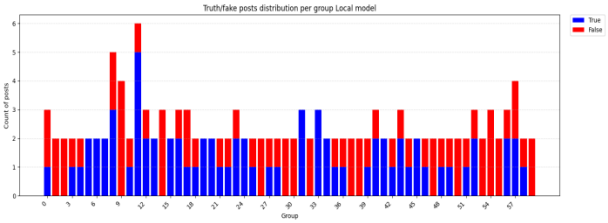
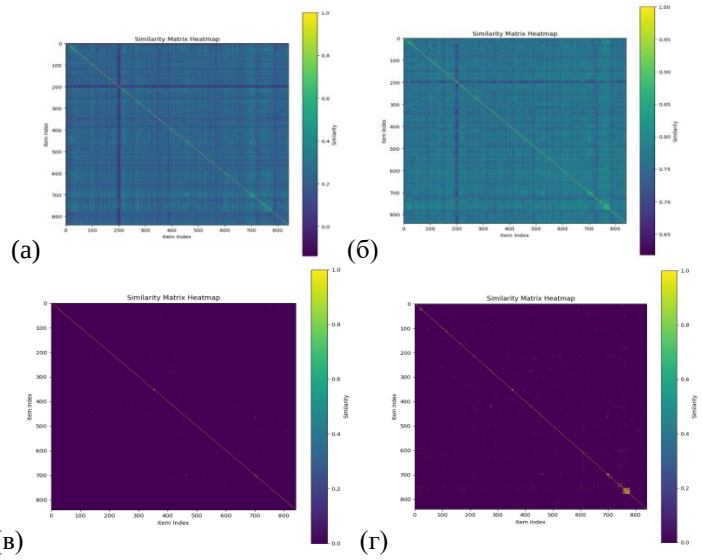


Рис. 10. Стовпчикова діаграма, яка відображає кількість та пропорції для правдивих та фейкових постів, згрупованих за подібністю за Local ембедінгами.

Застосуємо описаний вище алгоритм виявлення подібних постів, використавши, їхні ембедінги та проведемо аналіз результатів.



(а) (б)

(в) (г)

Рис. 11. Теплові мапи косинусної подібності між ембедінгами датасету 1. а - ембедінги отримані за допомогою OpenAI моделі; б - ембедінги отримані за допомогою локальної моделі. в - бінарне

представлення згрупованих OpenAI ембедінгів з косинусною подібністю не менше 0.6, г - бінарне представлення згрупованих локальних ембедінгів з косинусною подібністю не менше 0.85 (Жовтий піксель - пости визначено подібними, фіолетовий - різними).

Датасет 1 (рис. 11) загалом рівномірний, кластерів дубльованих текстів мало. Є помітна вертикально-горизонтальна смужка в околі індексу 200. Це група постів-«аномалій», які істотно відрізняється від решти й дають низьку схожість з усім масивом.В околі діапазонів 10-40, 680-720, 760-810 чітко проглядаються невеликі жовтуваті блоки – це групи схожих дописів (у межах самої групи).

В Датасеті 2 (рис.12) перші ≈ 350 рядків – ряди жовтих квадратів (6-7 компактних блоків). Це групи майже ідентичних тез, що повторювалися пакетами. OpenAI чітко окреслює межі блоків; local теж показує їх, але межі розмиті й частина «внутрішнього» фону зливається з блоками через високі базові значення. Спостерігаються яскраво жовті блоки – це групи репостів, які є ідентичними між собою. Друга половина матриці набагато темніша – більш різнорідні пости, майже без репостів, але часто зустрічаються схожі пости. У цій частині графіку блоків біля діагоналі не видно, оскільки дані не є відсортованими.

Для Датасету 3 OpenAI (рис. 13): графік однорідної зернистості із невеликою, але чіткою «плямою» ≈ індекси 720-760 – серія майже тотожних повідомлень (репости) та деякими поодинокими кластерами подібних постів. Local: така сама структура, але фон значно яскравіший; різницю між типовими й «дубльованими» постами видно гірше.

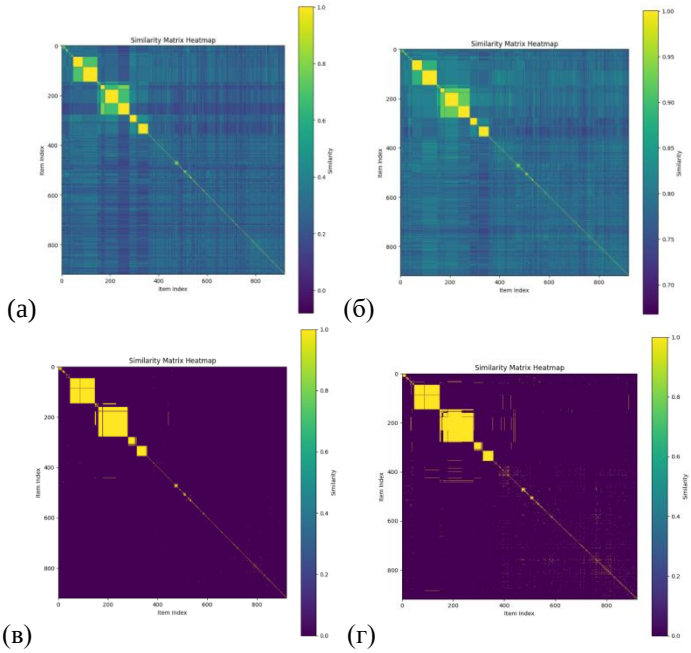


Рис. 12. Теплові мапи косинусної подібності між ембедінгами датасету 2. а - ембедінги отримані за допомогою OpenAI моделі; б - ембедінги отримані за допомогою локальної моделі. в - бінарне представлення згрупованих OpenAI ембедінгів з косинусною подібністю не менше 0.6, г - бінарне представлення згрупованих локальних ембедінгів з косинусною подібністю не менше 0.85 (Жовтий піксель - пости визначено подібними, фіолетовий - різними).

Проаналізуємо ембедінги двох моделей (OpenAI/text-embedding-3-small та intfloat/multilingual-e5-base), побудувавши графік, який відображає зміну кількості кластерів – тобто компонент розміром ≥ 4 пости – коли ми поступово підвищуємо поріг косинусної схожості (τ) для двох моделей. Ліва панель відображає ембедінги OpenAI, права – локальна модель (рис. 14).

В OpenAI-embedding (ліва панель) до $\tau \approx 0.35$ видно лише один великий кластер: майже всі пости злипаються. На $\tau \approx 0.40$-0.45 кластери починають відокремлюватись; кількість зростає до 2-5. Максимум (~11 кластерів) припадає на діапазон 0.55-0.60. Тут модель найкраще розрізняє окремі лінії поширення, не дроблячи їх надміру. Після $\tau > 0.65$ графік швидко «здувається»; при $\tau \geq 0.80$ суттєвих кластерів уже немає. Отже, OpenAI-векторів оптимальний робочий поріг лежить поблизу 0.55-0.60; нижче все зливається, вище кластери розпадаються.

ISSN 2663 – 4023

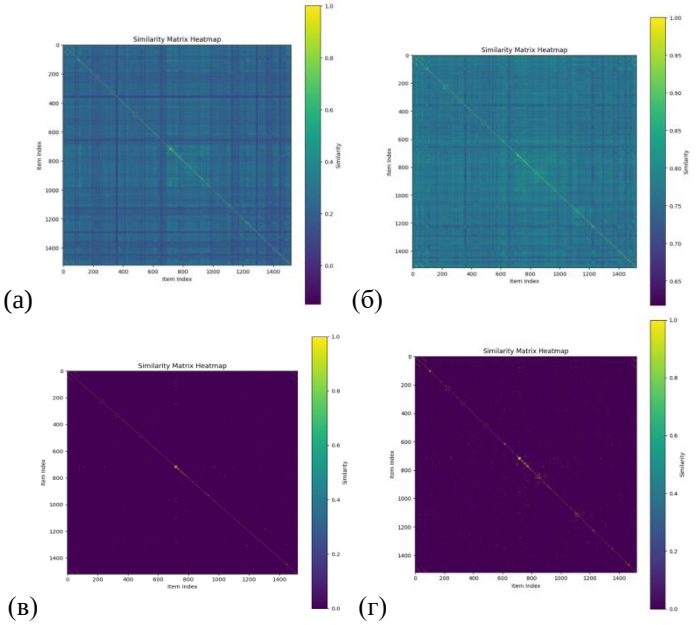


Рис. 13. Теплові мапи косинусної подібності між ембедінгами датасету 3. а - ембедінги отримані за допомогою OpenAI моделі; б - ембедінги отримані за допомогою локальної моделі. в - бінарне представлення згрупованих OpenAI ембедінгів з косинусною подібністю не менше 0.6, г - бінарне представлення згрупованих локальних ембедінгів з косинусною подібністю не менше 0.85 (Жовтий піксель - пости визначено подібними, фіолетовий - різними).

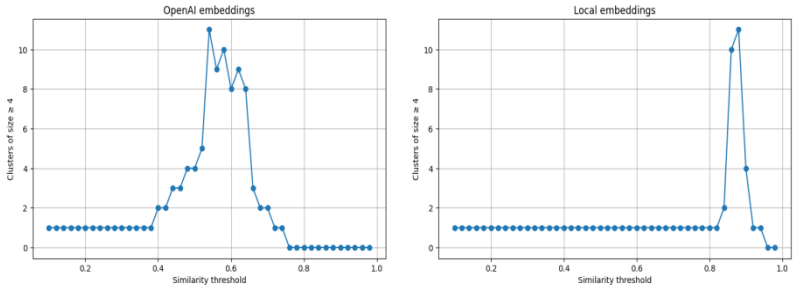


Рис. 14. Графік залежності кількості кластерів від порогу косинусної подібності між ембедінгами постів.

Для локальної моделі (права панель) до $\tau \approx 0.80$ тримається стабільна «плоска» лінія: система бачить лише один великий кластер, бо всі вектори схожі між собою. Кластери різко «вистрілюють» лише після $\tau \approx 0.82$, досягаючи піку 10-11 при 0.87-0.89. Далі, як і в попередньому випадку, кількість швидко паде; уже на $\tau \approx 0.93$ залишається 1 – 2 кластери, а при 0.96 їх немає зовсім. Отже, локальні embeddings мають вищу базову схожість, тому адекватний поріг треба ставити значно вище – приблизно 0.86-0.88. В підсумку OpenAI дає більшу динаміку та контраст, тому легше візуально та алгоритмічно відшукати кластери.

Для подальшої обробки позначимо великі жовті квадрати як «компоненти високої схожості» й подавати на графову візуалізацію. Підозрілі смуги (один пост проти всіх) – корисний індикатор унікальних точок входу дезінформації. Побудуємо граф-таймлайн для датасету 3 та проаналізуємо результат.

- Горизонтальна вісь – це час (ліворуч → праворуч).
- Кожен рядок – окремий кластер, тобто група повідомлень, які виявились дуже схожими за змістом у матриці косинусної подібності ($\tau \geq 0.90$).
- Вершини як коло – первинна публікація (post) та трикутник – репост / ретрансляція (repost)
- Колір: синій – повідомлення позначено як правдиве, червоне – фейк.
- Ребра поєднують публікації всередині кластеру в хронологічному порядку.
- Проміжок ~5 днів між першим TG-джерелом і поширенням на X/Facebook.

- Перетворення репоста у «свіжий» пост (трикутник → коло) з боку одного й того ж автора – типовий спосіб «загубити» оригінальне походження повідомлення.
- У тексті (tooltip) згадано «ракета NASAMS влучила … Охматдит» – дезінформаційна теза РФ.

Рис. 15. Граф поширення інформації

Розглянемо кластер №6 з рисунку 15, який ілюструє типовий маршрут фейку:

Резидент (Telegram, 02 Aug 20:30)
↓ 5 діб
Alexander Kovalenko (X)
↓ 4 год
Andro Newman (X)
↓ 10 хв
Ігор Лієпас (X)
↓ 2 год
Marina (X)
↓ 9 год
Elena Ying ▲ (FB, репост)
↓ кілька секунд
Elena Ying ● (FB, «новий» пост)

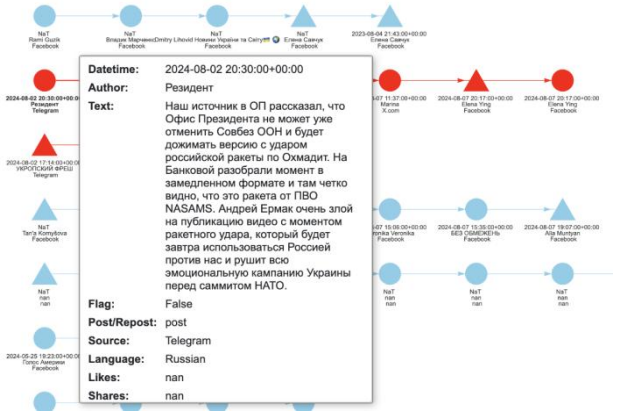


Рис. 16. Візуалізація даних про першу появу інформації

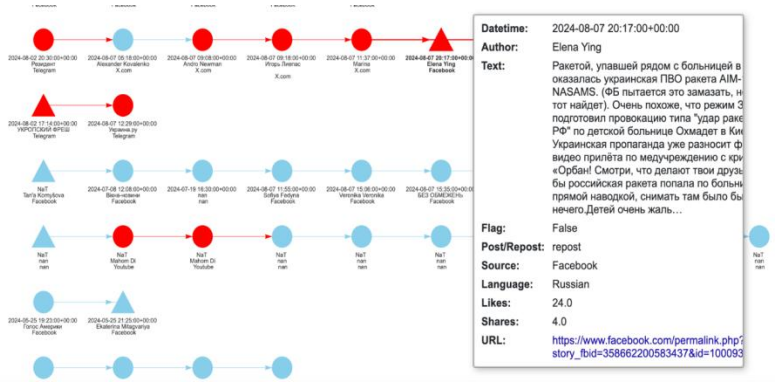


Рис. 17. Візуалізація даних одного з репостів інформації зазначеної на Рис.15.

У ході виконання роботи було створено інформаційну систему, яка автоматично відстежує джерело й подальший шлях дезінформаційних повідомлень. Система аналізує пости з різних платформ, перетворює текст на векторні подання та за косинусною схожістю групує повідомлення у кластери. На основі сформованих кластерів система вибудовує хронологічний граф, де можна побачити, хто і коли вперше опублікував тезу, на яких майданчиках її підхопили, і з якою затримкою фейк перейшов до інших платформ. Практичні експерименти на трьох датасетах показали, що модель OpenAI створює векторні подання які чіткіше виділяються у матриці схожості, завдяки чому легше виділити кластери, тоді як векторні представлення створені локальною моделлю мають вищий поріг подібності. Попри це, обидві моделі впевнено виявляють групи подібних постів та репостів. Робота також виявила слабкі місця даних: пропуски у датах імен авторів впливають на коректність графа, а ручне маркування достовірності залишається вузьким місцем. Водночас запропонований підхід уже сьогодні дозволяє аналітикам миттєво знаходити «нульові» джерела та оцінювати швидкість розповсюдження.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У межах виконаного дослідження розроблено та реалізовано інформаційну систему для автоматизованого аналізу повідомлень у соціальних мережах з метою виявлення дезінформаційних кластерів і відстеження маршрутів їх поширення. Система поєднує сучасні методи обробки природної мови (векторизація за допомогою embedding-моделей), обчислення косинусної подібності, кластеризації схожих постів та графової візуалізації поширення. Успішно реалізовано архітектуру системи, яка дозволяє завантажувати, очищати, векторизувати та аналізувати великі обсяги публікацій з соціальних мереж та онлайн-ЗМІ. Порівняно дві embedding-моделі OpenAI text-embedding-3-small та модель intfloat/multilingual-e5-base. OpenAI text-embedding-3-small забезпечує високу контрастність у матрицях схожості, що спрощує виявлення кластерів при нижчому порозі ($\tau \approx 0.6$-0.7). Локальна модель intfloat/multilingual-e5-base створює вектори з вищою базовою подібністю, що вимагає підвищення порогу ($\tau \approx 0.85$-0.9) для надійного кластерування.

На трьох датасетах проведено експериментальний аналіз поширення інформації. Встановлено, що модель OpenAI дозволяє візуально та алгоритмічно чіткіше виокремити кластери. Виявлено характерні групи майже ідентичних постів та репостів, що підтверджує ефективність обраного підходу. Продемонстровано можливість виявлення маршрутів поширення фейкових тез – наприклад, кластер №6 із повідомленням про «ракети NASAMS» відстежено від Telegram-каналу до Facebook з часовою затримкою близько 5 діб, включно з репостами та навмисними змінами для приховування джерела. Виявлено слабкі місця у вхідних даних, зокрема відсутність уніфікованих імен авторів та неточності у мітках достовірності, що впливають на якість побудованих графів. Практична реалізація системи дає змогу аналітикам швидко визначати першоджерела дезінформації, оцінювати темпи її поширення та виявляти ретрансляторів, що є важливим інструментом у протидії інформаційним атакам. Таким чином, розроблена система може бути використана у сфері інформаційної безпеки, OSINT-розвідки, журналістських розслідувань, а також як основа для подальших досліджень у галузі виявлення фейкових повідомлень і моделювання інформаційних потоків.

Попри отримані результати, запропонована інформаційна технологія виявлення джерел і маршрутів поширення дезінформації відкриває низку напрямів для подальшого розвитку та поглиблення досліджень.

1. Удосконалення семантичного моделювання.

2. Інтеграція графових нейронних мереж.
3. Моделювання динаміки поширення.
4. Автоматизація визначення достовірності.
5. Мультимодальний аналіз.
6. Міжплатформний моніторинг у реальному часі.
7. Поглиблений аналіз поведінкових патернів.
8. Практичне впровадження та оцінювання ефективності.

Перспективним напрямом є розширення експериментів із сучасними мовними моделями та їх адаптація до специфіки дезінформаційного контенту. Доцільним є:

- використання більших та більш контекстно чутливих embedding-моделей;
- тонке налаштування (fine-tuning) моделей на корпусах фейкових новин;
- інтеграція гібридних підходів, що поєднують текстові embeddings з метаданими (автор, час, платформа).

Окремої уваги потребує дослідження стійкості моделей до навмисних семантичних викривлень (перефразування, сарказм, часткова зміна формулювань), які часто застосовуються для маскування першоджерела.

Подальші дослідження можуть бути спрямовані на застосування графових нейронних мереж (GNN) для глибшого аналізу структури інформаційних каскадів. Зокрема, перспективним є:

- використання підходів, подібних до тих, що запропоновані у роботах Monti, Federico та співавторів щодо геометричного глибокого навчання для виявлення фейків;
- розширення мережевої моделі з урахуванням ваг ребер (часові інтервали, інтенсивність взаємодії);
- автоматичне визначення «ключових вузлів» за допомогою навчальних моделей замість фіксованих метрик центральності.

Це дозволить перейти від евристичного аналізу графів до прогнозування майбутніх каскадів поширення.

Перспективним є застосування епідеміологічних моделей поширення інформації (SIR, Independent Cascade) для:

- прогнозування швидкості розповсюдження нової дезінформаційної тези;
- оцінювання потенційного масштабу інформаційної атаки;
- моделювання ефекту втручання (видалення поста, блокування джерела).

Поєднання семантичної кластеризації з динамічним моделюванням дозволить створити систему раннього попередження про інформаційні загрози.

У поточній реалізації використовується ручне або попередньо визначене маркування достовірності (flag). Подальші дослідження можуть бути спрямовані на:

- автоматичну класифікацію правдивості повідомлень;
- інтеграцію зовнішніх фактчекінгових джерел;
- розроблення ансамблевих моделей, що поєднують content-based і network-based ознаки.

Це дозволить зменшити залежність системи від людського маркування та підвищити масштабованість.

Оскільки сучасна дезінформація часто містить зображення, відео та аудіо, перспективним напрямом є:

- інтеграція аналізу зображень (меми, фотоманіпуляції);
- виявлення повторного використання медіафайлів;
- поєднання текстових і візуальних embeddings у єдиному просторі ознак.

Мультимодальний підхід дозволить виявляти інформаційні кампанії, що використовують комбінований контент. Подальший розвиток системи може передбачати: автоматичний збір даних із різних соціальних мереж у режимі реального часу; побудову потокової (streaming) архітектури; інтеграцію API платформ, таких як Telegram та X. Це дозволить оперативно фіксувати «нульові точки входу» дезінформації та скорочувати час реагування. Перспективним є вивчення:

- поведінкових характеристик ретрансляторів;
- координаційних ознак (синхронність публікацій, повторюваність шаблонів);
- виявлення бот-мереж і напівавтоматизованих акаунтів.

Поєднання структурного аналізу з поведінковими індикаторами дозволить точніше ідентифікувати організовані інформаційні кампанії.

Важливим напрямом є апробація системи в реальних умовах:

- інтеграція у системи моніторингу інформаційної безпеки;
- використання в OSINT-аналітиці;
- порівняльне тестування на різних мовах та регіональних сегментах інформаційного простору.

Також доцільним є розроблення формалізованих метрик оцінювання ефективності виявлення маршрутів (точність визначення першоджерела, повнота реконструкції каскаду, час обробки). Подальші дослідження можуть бути спрямовані на підвищення точності семантичного групування, глибшу інтеграцію мережевих методів, автоматизацію визначення достовірності та розширення системи до мультимодального й потокового аналізу. Реалізація зазначених напрямів дозволить створити повноцінну інтелектуальну систему раннього виявлення та прогнозування дезінформаційних атак у цифровому інформаційному середовищі.

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. Paraschiv, M., et al. (2022). A unified graph-based approach to disinformation detection using contextual and semantic relations. *Proceedings of the International AAAI Conference on Web and Social Media, 16*. https://doi.org/10.48550/arXiv.2109.11781
2. Monti, F., et al. (2019). Fake news detection on social media using geometric deep learning. *arXiv*. https://doi.org/10.48550/arXiv.1902.06673
3. Gong, S., et al. (2023). Fake news detection through graph-based neural networks: A survey. *arXiv*. https://doi.org/10.48550/arXiv.2307.12639
4. Papadopoulou, O., et al. (2022). MeVer NetworkX: Network analysis and visualisation for tracing disinformation. *Future Internet, 14*(5), Article 147. https://doi.org/10.3390/fi14050147
5. Soga, K., Yoshida, S., & Muneyasu, M. (2024). Graph-based interpretability for fake news detection through topic- and propagation-aware visualisation. *Computation, 12*(4), Article 82. https://doi.org/10.3390/computation12040082
6. Luo, H., Cai, M., & Cui, Y. (2021). Spread of misinformation in social networks: Analysis based on Weibo tweets. *Security and Communication Networks, 2021*, Article 7999760. https://doi.org/10.1155/2021/7999760
7. Béres, F., et al. (2023). Network embedding aided vaccine skepticism detection. *Applied Network Science, 8*(1), Article 11. https://doi.org/10.1007/s41109-023-00534-x
8. Liu, P., et al. (2025). A thorough comparison between independent cascade and susceptible-infected-recovered models. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(1). https://doi.org/10.1609/aaai.v39i1.32028
9. Muñoz, P., Díez, F., & Bellogín, A. (2024). Modeling disinformation networks on Twitter: Structure, behavior, and impact. *Applied Network Science, 9*(1), Article 4. https://doi.org/10.1007/s41109-024-00610-w
10. Su, T., Macdonald, C., & Ounis, I. (2022). Leveraging users' social network embeddings for fake news detection on Twitter. *arXiv*. https://doi.org/10.48550/arXiv.2211.10672
11. Schiffrin, A., et al. (2022). *AI startups and the fight against mis/disinformation online: An update*. German Marshall Fund of the United States.
12. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter, 19*(1), 22-36. https://doi.org/10.1145/3137597.3137600

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

Yaroslav Teplyi
PhD student of the Information Systems and Networks Department
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0009-0001-5548-5530
yaroslav.b.teplyi@lpnu.ua

Yulian Kuryliak
PhD student of the Information Systems and Networks Department
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0000-0002-6600-2637
yulian.a.kuryliak@lpnu.ua

Vitalii Torshyn
PhD student of the Information Systems and Networks Department
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0009-0002-0133-6553
vitalii.v.torshyn@lpnu.ua

RESEARCH INTO MECHANISMS FOR DETECTING SOURCES AND WAYS OF FAKE NEWS AND PROPAGANDA DISSEMINATION IN SOCIAL NETWORKS CYBERSPACE

Abstract. The article presents an information technology for detecting sources and routes of fake news and propaganda distribution in social networks and online media. The proposed approach is based on vectorization of text publications using language models, calculation of cosine similarity and clustering of messages to identify thematically similar groups. Based on the formed clusters, a chronological graph visualization is built, which allows identifying the primary sources of disinformation, key relays and the speed of its distribution. The experimental study was conducted on a combined dataset of over 2,000 posts, in which the shares of true and fake messages are relatively balanced. Interaction analysis showed that about 75% of all posts receive no more than 20 likes, while only less than 5% of posts form a “long tail” with hundreds and thousands of reactions. At the same time, disinformation posts are more likely to either remain almost unnoticed or to sharply gain abnormally high popularity in a short period of time. Analysis of distributions revealed that approximately 80% of posts have no more than 5 reposts, however, the share of fakes among the most widely shared messages is growing significantly. Platform analysis showed that reliable content prevails on web resources and authoritative media, while on social networks and messengers (in particular Telegram) the ratio of true and fake messages is close to equilibrium, and in some network’s disinformation dominates. A comparison of embedding models showed that the OpenAI model provides a clearer separation of messages in the feature space and allows detecting up to 10-11 significant clusters at an optimal cosine similarity threshold of $\tau \approx 0.55\text{-}0.60$, while for the local model the optimal threshold is much higher ($\tau \approx 0.86\text{-}0.88$). The constructed propagation graphs showed that the time interval between the initial publication of a fake and its appearance on other platforms can be from several hours to several days, and the transformation of reposts into “original” publications is a typical mechanism for hiding the source.

Keywords: text embeddings, cosine similarity, clustering, information dissemination analysis, graph visualization, information security.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Paraschiv, M., et al. (2022). A unified graph-based approach to disinformation detection using contextual and semantic relations. *Proceedings of the International AAAI Conference on Web and Social Media, 16*. https://doi.org/10.48550/arXiv.2109.11781
2. Monti, F., et al. (2019). Fake news detection on social media using geometric deep learning. *arXiv*. https://doi.org/10.48550/arXiv.1902.06673
3. Gong, S., et al. (2023). Fake news detection through graph-based neural networks: A survey. *arXiv*. https://doi.org/10.48550/arXiv.2307.12639
4. Papadopoulou, O., et al. (2022). MeVer NetworkX: Network analysis and visualisation for tracing disinformation. *Future Internet, 14*(5), Article 147. https://doi.org/10.3390/fi14050147
5. Soga, K., Yoshida, S., & Muneyasu, M. (2024). Graph-based interpretability for fake news detection through topic- and propagation-aware visualisation. *Computation, 12*(4), Article 82. https://doi.org/10.3390/computation12040082
6. Luo, H., Cai, M., & Cui, Y. (2021). Spread of misinformation in social networks: Analysis based on Weibo tweets. *Security and Communication Networks, 2021*, Article 7999760. https://doi.org/10.1155/2021/7999760
7. Béres, F., et al. (2023). Network embedding aided vaccine skepticism detection. *Applied Network Science, 8*(1), Article 11. https://doi.org/10.1007/s41109-023-00534-x
8. Liu, P., et al. (2025). A thorough comparison between independent cascade and susceptible-infected-recovered models. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(1). https://doi.org/10.1609/aaai.v39i1.32028
9. Muñoz, P., Díez, F., & Bellogín, A. (2024). Modeling disinformation networks on Twitter: Structure, behavior, and impact. *Applied Network Science, 9*(1), Article 4. https://doi.org/10.1007/s41109-024-00610-w
10. Su, T., Macdonald, C., & Ounis, I. (2022). Leveraging users' social network embeddings for fake news detection on Twitter. *arXiv*. https://doi.org/10.48550/arXiv.2211.10672
11. Schiffrin, A., et al. (2022). *AI startups and the fight against mis/disinformation online: An update*. German Marshall Fund of the United States.
12. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter, 19*(1), 22-36. https://doi.org/10.1145/3137597.3137600

Отримано редакцією журналу / Received: 08.02.26
Прорецензовано / Revised: 21.02.26
Схвалено до друку / Accepted: 25.06.26