КОМП’ЮТЕРНІ НАУКИ

DOI: 10.31319/2519-2884.46.2025.16
УДК 004.8:007.5:351.74

Шаповалова Н.Н., старший викладач, ORCID: 0000-0001-9146-1205,
e-mail: shapovalova@knu.edu.ua
Доценко І.О., старший викладач, ORCID: 0000-0001-7912-2497, e-mail: dotsenko@knu.edu.ua
Саітгареєв Н.Х., к.т.н., доцент, ORCID: 0009-0002-2399-9120, e-mail: saithareiev@knu.edu.ua
Трачук А.А., к.т.н., доцент, ORCID: 0000-0001-6241-1575, e-mail: trachuk@knu.edu.ua
Криворізький національний університет, м. Кривий Ріг

Shapovalova Nonna, Senior lecturer of the Department of Software Modeling
Dotsenko Iryna, Senior lecturer of the Department of Software Modeling
Saithareiev Nail, Candidate of Technical Sciences, Docent at the Department of Software Modeling
Trachuk Annait, Candidate of Technical Sciences, Docent at the Department of Software Modeling
Kryvyi Rih National University, Kryvyi Rih

ІНТЕГРАЦІЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ В СЕМАНТИЧНИЙ АНАЛІЗ OSINT-ДАНИХ: АЛГОРИТМИ ТА РЕЗУЛЬТАТИ

У роботі описується процес розробки універсального програмного модуля для семантичного аналізу даних розвідки на основі відкритих джерел із використанням методів машинного навчання. Запропоноване рішення автоматизує процеси збору, обробки та кластеризації великих масивів інформації, підвищуючи точність аналізу та спрощуючи інтерпретацію результатів для спеціалістів у сферах кібербезпеки, бізнес-аналітики та журналістики.

***Ключові слова**: семантичний аналіз; розвідка на основі відкритих джерел; машинне навчання; алгоритм кластеризації; моделі глибокого навчання; обробка природної мови.*

The paper describes the process of developing a universal software module for semantic analysis of open source intelligence data using machine learning methods. The proposed solution automates the processes of gathering, processing, and clustering large volumes of information, increasing the accuracy of analysis and simplifying the interpretation of results for professionals in the fields of cyber security, business analytics, and journalism.

***Keywords:** semantic analysis; open-source intelligence; machine learning; clustering algorithm; deep learning models; natural language processing.*

Постановка проблеми

У сучасному інформаційному середовищі, де обсяги даних щодня збільшуються на ексабайти, виникає нагальна потреба у створенні інструментів, здатних автоматичноопрацьовувати та аналізувати ці дані. Серед основних джерел інформації виділяють вебсайти, соціальні мережі, наукові публікації, блоги та державні реєстри [1—4]. У цьому контексті OSINT (OpenSourceIntelligence) відіграє важливу роль, оскільки дозволяє збирати та аналізувати інформацію, доступну у відкритих джерелах, без порушення конфіденційності. Серед основних напрямків застосування OSINT виділяють наступні: кібербезпека (виявлення потенційних загроз, аналіз атак, пошук вразливостей у системах); бізнес-аналітика (моніторинг ринкових тенденцій, аналіз конкурентів, управління репутацією); журналістика (дослідження джерел інформації, перевірка фактів, розслідування корупційних справ); національна безпека та розвідка (аналіз геополітичних ризиків, відстеження соціальних настроїв) [5].

Однією з основних проблем, що виникає при роботі з відкритими даними, є їхній великий обсяг і неструктурованість. Відсутність стандартних форматів і значна частка «інформаційного шуму» значно ускладнюють процес обробки таких даних. Наприклад, аналіз інформації з соціальних мереж вимагає вилучення ключових даних, таких як імена, дати, геолокація, з

великої кількості нерелевантних повідомлень.

Традиційні методи обробки не можуть забезпечити необхідної швидкості й точності аналізу, особливо в умовах постійного збільшення обсягів даних. Розробка автоматизованих рішень, зокрема програмних модулів для семантичної обробки, дозволяє оптимізувати процес аналізу даних, забезпечуючи швидке та точне виявлення корисної інформації [6, 7].

Таким чином, створення інструментів, здатних виконувати семантичний аналіз, категоризацію й кластеризацію даних у реальному часі, є критично важливим для ефективного вирішення завдань у галузях кібербезпеки, бізнес-аналітики та національної безпеки [8—10].

Аналіз останніх досліджень та публікацій

У сфері сучасних систем OSINT працює багато інструментів, кожен із яких має свої сильні й слабкі сторони. Наведемо детальніший аналіз деяких із них. Maltego — потужний інструмент, який забезпечує візуалізацію зв’язків між різними об’єктами, такими як IP-адреси, домени, електронні адреси та інше. Однією з переваг є можливість інтеграції з базами даних та іншими сервісами для розширення функціональності. Однак висока вартість робить його недоступним для невеликих організацій. PulsarPlatform — орієнтована на моніторинг соціальних медіа, дозволяючи аналізувати настрої та визначати ключові тренди. Ця платформа особливо популярна серед маркетологів, але її вузька спрямованість обмежує можливості застосування в інших галузях. RavenPack — інструмент, що спеціалізується на аналізі великих масивів текстових даних, включаючи новини та соціальні мережі. Його перевагою є адаптація до фінансових даних, що робить платформу корисною для інвесторів. BrandwatchAnalytics — платформа для моніторингу соціальних мереж та аналізу відкритих даних, яка широко застосовується для оцінки ефективності маркетингових кампаній. Її слабкою стороною є недостатній акцент на аналіз даних поза соціальними медіа. Lexalytics — одна з найпотужніших систем для аналізу текстових даних, яка використовує сучасні методи обробки природної мови. Однак відсутність інтеграції з автоматизованим збором даних обмежує її застосування.

Задля реалізації поставлених завдань буде використовуватись семантичний аналіз у контексті OSINT [11, 12]. Цей метод дозволяє не лише виявляти ключові терміни, а й аналізувати емоційний тон текстів, структуру взаємозв’язків між поняттями та відношення між сутностями.

Формальне представлення алгоритму семантичного аналізу (виділення сутностей) представлений нижче.

Вхід: Текстовий запит Q.

Виконати попередню обробку тексту:

видалити зайві символи;

виконати токенізацію.

Побудувати векторизацію запиту $V(Q)$ за допомогою методу TF-IDF.

Застосувати модель розпізнавання іменованих сутностей (NER), щоб виділити:

імена (людей, компаній, геолокацій);

посилання, електронні адреси.

*Вихід***:** Повернути набір сутностей $E=\{e_1, e_2, ..., e_n\}$.

Алгоритми, такі як TF-IDF, DBSCAN та новітні методи на основі трансформерів, наприклад, BERT(BidirectionalEncoderRepresentationsfromTransformers), допомагають значно підвищити точність і релевантність аналізу [16]. Наведемо формальний алгоритм кластеризації DBSCAN.

Вхід: Набір даних $D=\{d_1, d_2, ..., d_m\}$.

Визначити відстані між об'єктами за допомогою косинусної подібності.

Встановити радіус сусідства ε і мінімальну кількість сусідів *MinPts*.

Позначити всі об'єкти як невизначені.

Для кожного об'єкта:

якщо кількість сусідів >= *MinPts*, утворити кластер;

додати всі сусідні об'єкти до кластера.

Вихід: Повернути набір кластерів $C=\{c_1, c_2, ..., c_k\}$.

Удосконалення семантичного аналізу можливе за рахунок використаннясучасних моделей машинного навчання, таких як BERT [17, 18]. Порівняно з класичними підходами (наприклад, TF-IDF), BERT дозволяє краще враховувати контекст запиту, виділяти взаємозв’язки між поняттями й аналізувати дані різними мовами, що важливо для глобального OSINT. Наведемо формальний опис алгоритму роботиз BERT:

Вхід: Текстовий запит Q.

Токенізувати запит, розбиваючи текст на окремі слова або субслова.

Визначити векторне представлення для кожного токена за допомогою моделі BERT.

Об’єднати токенізовані вектори в один вектор $V(Q)$, що враховує контекст усіх слів у запиті.

Вихід: Векторне представлення запиту V(Q).

Необхідно зазначити, що крім високої вартості більшості платформ, сучасні системи OSINT стикаються із такими труднощами: нестача інтегрованих рішень для збору, обробки й аналізу даних; високі вимоги до обчислювальних ресурсів; складність адаптації до потреб різних користувачів[19, 20].

Формулювання мети дослідження

Головною метою роботи є розробка універсального програмного модуля для семантичної обробки OSINT-даних, який дозволить автоматизувати процеси збору, аналізу й кластеризації інформації. Задля поставленої мети необхідно реалізувати наступні завдання: провести порівняльний аналіз сучасних методів збору та обробки OSINT-даних; розробити нові алгоритми для автоматичної класифікації та фільтрації інформаційного шуму; інтегрувати модуль у клієнт-серверну архітектуру з можливістю масштабування; провести тестування розробленого рішення на реальних даних, включаючи оцінку продуктивності. Серед очікуваних результатів автори виділяють прискорення процесу аналізу даних, зменшення обсягу ручної праці, покращення точності аналізу, завдяки використанню семантичних алгоритмів.

Виклад основного матеріалу

Програмний модуль для семантичної обробки даних OSINT побудований за клієнт-серверною архітектурою, яку представлено у вигляді багаторівневої моделі, що включає наступні компоненти: клієнтський рівень (Frontend) — забезпечує інтерфейс для користувача, включає текстове поле для введення запиту та відображення результатів; серверний рівень (Backend) — реалізує основні обчислення, включаючи семантичний аналіз, інтегрується з джерелами даних через API або веб-скрапери; рівень збору даних (DataCollection) — забезпечує доступ до зовнішніх джерел, таких як соціальні мережі, новини, наукові публікації; база даних — зберігає результати попередніх обробок для підвищення швидкості повторних запитів.

Така архітектура забезпечує масштабованість модуля, гнучкість інтеграції з іншими системами та можливість використання як незалежного мікросервісу. Основні компоненти модуля включають:

Семантичний аналізатор. Використовує методи обробки природної мови (NLP) для виділення ключових сутностей із тексту, таких як імена, назви компаній, географічні місця та посилання. Основні етапи роботи: попередня обробка тексту (очищення від зайвих символів, токенізація); векторизація тексту за допомогою методів, таких як TF-IDF або сучасних моделей BERT; виділення сутностей за допомогою алгоритмів розпізнавання іменованих сутностей (NER).

Модуль збору даних. Інтегрується з різними джерелами інформації: соціальні мережі (збір даних через API); новинні портали **(**використання веб-скрапінгу та спеціалізованих інструментів); наукові публікації (автоматизований пошук через сервіси, такі як GoogleScholar).

Інтерфейс користувача. Реалізовано з використанням бібліотеки Jinja2 для динамічного формування HTML. Основні функції: поле для введення запиту; результати у структурованому вигляді: кластеризовані за типами даних.

Система кластеризації. Використовує DBSCAN для автоматичного групування схожих об'єктів. Це забезпечує виділення основних тематичних блоків у великих обсягах даних.

Модуль обробки даних базується на сучасних алгоритмах машинного навчання. Основні етапи роботи алгоритму включають виділення ключових слів(текст аналізується для визначення найбільш релевантних термінів на основі частоти їх використання), кластеризацію (алгоритм DBSCAN автоматично групує текстові блоки за семантичним змістом, виділяючи основні теми) і фільтраціюінформаційного шуму (видаляються дублікати та нерелевантні дані, що зменшує навантаження на систему й покращує якість аналізу).

Приклади роботи системи. Припустимо, користувач вводить запит: «Аналіз кіберзагроз у Східній Європі». Система виділяє ключові терміни: «кіберзагрози», «Східна Європа» і виконує пошук відповідноїінформації у соціальних мережах, новинах та спеціалізованих базах даних, потім кластеризує результати, виділяючи блоки, наприклад: атаки на фінансові установи, державні ініціативи з кіберзахисту, вразливості інфраструктури тощо. Користувач отримує ці результати у зручному вигляді, що дозволяє швидко оцінити загальну картину.

Також система дозволяє виявляти інформаційні кампаній у соціальних мережах. Так, наприклад, за запитом «Дезінформація у соціальних мережах щодо економічної кризи», було виділено основні сутності: «економічна криза», «соціальні мережі», «дезінформація», зібрано дані з Twitter, Facebook, і новинних порталів. В результаті виявлено три основні кластери: кампанії у Facebook із фейковими новинами, трендові хештеги у Twitter, новини зі спростуванням фейків. Отже, система дозволила виявити джерела дезінформації та відстежити основні теми обговорення.

Розроблений модуль був протестований на реальних даних із різних відкритих джерел, таких як соціальнімережі, новинні портали та наукові публікації. Було оброблено понад 1000 публікацій на теми кібербезпеки в соціальних мережах, з яких система змогла виділити основні тренди з точністю 92 %. Аналіз 500 новинних статей продемонстрував ефективність кластеризації з точністю понад 90 %. Система успішно знаходила ключові публікації, пов'язані із заданою темою, з мінімальною часткою нерелевантних даних. Результати тестування свідчать про високу ефективність модуля навіть за умов роботи з великими обсягами даних.

Для оцінки ефективності роботи модуля було проведено серію експериментів, спрямованих на визначення точності семантичного аналізу, оцінку швидкості обробки запитів і тестування продуктивності на великих обсягах даних.

Задля визначення точності семантичного аналізу, розроблений модуль було протестовано на наборі текстів із різних джерел: новинні портали, соціальні мережі, наукові статті тощо. Для оцінки точності використовувалася метрика F1-міра (середнє гармонійне метрик Precision і Recall). Precisionвизначає, яка частка виділених системою сутностей є правильними, Recallвизначає, яка частка реальних сутностей була виявлена системою. В результаті отримали Precision 91 %, Recall 88 %, F1-міра 89.5 %.

Час обробки запиту залежить від складності тексту та кількості джерел. У середньому прості запити (1—2 ключових слова) обробляються 2.5 с, складні запити (багатокомпонентні) — 5.7 с.

Система була протестована на наборі із 10000 текстових документів. Алгоритм кластеризації DBSCAN зміг ефективно розподілити документи на 25 кластерів за 3 хвилини.

Для оцінки переваг системи було проведено порівняння з кількома популярними платформами. Результати аналізу представлені в табл. 1.

Таблиця 1. Порівняння продуктивності платформ OSINT

Параметр	Maltego	PulsarPlatform	Розроблений модуль
Семантичний аналіз	Обмежений	Високий	Високий
Робота з джерелами	Середня	Висока	Висока
Вартість	Висока	Висока	Низька
Гнучкість інтеграції	Низька	Середня	Висока

Треба відзначити, що перевагою системи Maltego насамперед є потужна візуалізація даних, але серед недоліків можна зазначити високу вартість і обмежений семантичний аналіз. Сервіс PulsarPlatformаналізує соціальні мереж, чим, власне, обмежена його спеціалізація. Розробленому ж модулю притаманні інтеграція багатьох джерел, точний семантичний аналіз і, що досить важливо, доступність.

Висновки та напрямок подальших досліджень

У ході роботи було розроблено програмний модуль для семантичної обробки даних OSINT, який забезпечує автоматичну класифікацію й кластеризацію даних, зручний інтерфейс для користувачів, підтримку інтеграції з багатьма джерелами інформації. Система успішно працює з великими обсягами даних та різноманітними джерелами, а інтеграція семантичного аналізу та кластеризації значно підвищила якість результатів. Ефективність модуля підтверджена тестуванням на реальних даних. Досягнуто точності кластеризації понад 90 %, що є значним покращенням порівняно з існуючими рішеннями.

Попри досягнуті результати, розробка модуля зіштовхнулась із кількома викликами, а саме: обчислювальні ресурси — аналіз великих обсягів даних потребує значної кількості ресурсів; масштабованість — необхідно адаптувати модуль для роботи з багатомовними даними; адаптація до нових форматів — постійно з’являються нові джерела даних, що вимагає оновлення алгоритмів.

Серед пріоритетних напрямків подальших досліджень автори виділяють інтеграцію мультимедійних даних: розширення функціоналу для аналізу зображень, відео та аудіо; аналіз соціальних настроїв: використання алгоритмів для прогнозування трендів на основі даних із соціальних мереж; роботу в реальному часі: оптимізація модуля для швидкого аналізу потокових даних.Задля удосконалення сервісу розглядаються можливості розширення джерел отримання інформації (інтеграція з DarkWeb для виявлення ризиків у прихованих мережах і робота з API популярних платформ, таких як Reddit, Telegram тощо), оптимізація алгоритмів (використання графових нейронних мереж для кластеризації та зниження часу обробки запитів) і реалізація нових функцій, таких як прогнозування трендів і автоматичне створення звітів на основі отриманих даних.

Розроблена система в перспективі буде доповнена мультимодальним аналізом даних, що дозволить аналізувати не лише текстові дані, але й мультимедійні джерела, такі як відео з YouTube або супутникові знімки. Також планується розширення функціональності для аналізу текстів різними мовами, що дозволить охопити глобальні джерела даних, і доповнити функціонал можливістю робити прогнози на основі аналізу історичних даних, наприклад, прогнозування кібератак на основі попередніх інцидентів.

Список використаної літератури

1. Використання технологій OSINT для отримання розвідувальної інформації / О. В. Минько та ін. *Системи управління, навігації та зв’язку*. 2016. 4. С. 81–84.
2. Hwang Y. W., Lee I. Y., Kim H., Lee H., Kim D. Current status and security trend of OSINT. *Wireless Communications and Mobile Computing*. Vol. 2022 (1), ID 1290129.
3. Evangelista J. R. G., Sassi R. J., Romero M., Napolitano D. Systematic literature review to investigate the application of open source intelligence (OSINT) with artificial intelligence. *Journal of Applied Security Research.* 2021. Vol. 16(3). P. 345–369.
4. Литвиненко О. Є, Бурко Д. А. Моделі семантичного аналізу текстів. *Наукоємні технології*. 2009. Т. 4 : 4. С. 55–58.
5. Анісімов А. В., Марченко О. О., Никоненко А. О. Алгоритмічна модель асоціативно-семантичного контекстного аналізу текстів природною мовою. *Проблеми програмування*. 2008. 2-3. С. 379–384.
6. Maulud D. H., Zeebaree S. R., Jacksi K., Sadeeq M. A. M., Sharif K. H. State of art for semantic analysis of natural language processing. *Qubahan academic journal*. 2021. Vol. 1(2). P. 21–28.

7. Velupillai S., Mowery D., South B. R., Kvist M., Dalianis H. Recent advances in clinical natural language processing in support of semantic analysis. *Yearbook of medical informatics*. 2015. Vol. 24(1). P. 183–193.
8. Molenaar A., Lukose D., Brennan L., Jenkins E. L., McCaffrey T. A. Using Natural Language Processing to explore social media opinions on Food Security: sentiment analysis and topic modeling study. *Journal of Medical Internet Research.* 2024. Vol. 26. e47826.
9. Liu J., Li K., Zhu A., Hong B., Zhao P., Dai S., Su H. Application of deep learning-based natural language processing in multilingual sentiment analysis. *Mediterranean Journal of Basic and Applied Sciences (MJBAS).* 2024. Vol. 8(2). P. 243–260.
10. Block L. The long history of OSINT. *Journal of Intelligence History*. 2024. Vol. 23(2). P. 95–109.
11. Analytics of Big Data and social networks. URL: http://www.osp.ru/os/2013/08/13037856 (дата звернення: 16.03.2025).
12. Kontostathis A., Edwards L., Leatherman A. Text mining and cybercrime. Text Mining. Applications and Theory / ed. by Berry M. W., Kogan J. Chichester: Wiley, 2010. P. 149–164.
13. Dua S., Du X. Data Mining and Machine Learning in Cybersecurity. New York, 2011.
14. Text Mining: Applications and Theory / ed. by Berry M. W., Kogan J. Wiley, 2010. 224 p.
15. Thorsten J. Text Categorization with Support Vector Machines: Learning with Many Relevant Features URL: https://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf (дата звернення: 16.03.2025).
16. Sebastiani F. Machine learning in automated text categorization. *ACM Computing Surveys*. 2010. P. 1–47.
17. Esuli A., Sebastiani F. Determining the Semantic Orientation of Terms through Gloss Classification. Conference of Information and Knowledge Management (Bremen). ACM, New York, NY, 2005. P. 617–624.
18. Gupta M., Mishra A., Kumar V. A survey on fake news detection using machine learning techniques. *Journal of Big Data*. 2020. Vol. 7(1). P. 1–21.
19. Zhang Z., Chen Y., Chen L. Methods for detecting fake news: A survey. *ACM Computing Surveys*. 2020. Vol. 53(6). P. 1–37.
20. Browne T. O., Abedin M., Chowdhury M. J. M. A systematic review on research utilising artificial intelligence for open source intelligence (OSINT) applications. *International Journal of Information Security.* 2024. Vol. 23(4). P. 2911–2938.

INTEGRATION OF MACHINE LEARNING METHODS INTO SEMANTIC ANALYSIS OF OSINT DATA: ALGORITHMS AND RESULTS

Abstract

The aim of this study is to develop a universal software module for semantic analysis of open-source intelligence data using machine learning methods, enabling the automation of data collection, processing, and clustering of large datasets. Special emphasis is placed on improving analysis accuracy and simplifying result interpretation for users from various fields, such as cybersecurity, business analytics, and journalism. The study employs modern natural language processing methods, including Term Frequency-Inverse Document Frequency, Bidirectional Encoder Representations from Transformers for semantic text analysis, and the Density-Based Spatial Clustering of Applications with Noise algorithm for data grouping. Data collection was carried out using social media APIs, news portals, and automated web scraping tools. The module's performance was evaluated using precision, recall, and F1-score metrics. For the first time, this study proposes the integration of deep learning models, namely Bidirectional Encoder Representations from Transformers, with clustering algorithms to address open-source intelligence tasks. The system provides flexibility in adapting to different data sources and multilingual processing. A significant achievement is the improvement of semantic analysis by considering query context, which ensures higher result relevance. The developed module can be applied in cybersecurity for threat detection, in business for competitor analysis and market monitor-

ing, and in journalism for fact-checking, research, fake news detection, and analysis of information campaigns. The integration of multiple data sources enables a comprehensive approach to information analysis. The tested module demonstrated high efficiency, achieving a clustering accuracy of over 90 % and the capability to process large datasets (up to 10,000 documents in 3 minutes). The system successfully identifies key entities, automatically clusters data, and provides visualization in a user-friendly format.

References

[1] Mynko, O. V., Iokhov, O. Yu., Olenchenko, V. T., & Vlasov, K. V. (2016). Vykorystannia tekhnolohii OSINT dlia otrymannia rozviduvalnoi informatsii [Using OSINT technologies to obtain intelligence information]. Systemy upravlinnia, navihatsii ta zv'iazku, (4), 81–84. [in Ukrainian].

[2] Hwang, Y. W., Lee, I. Y., Kim, H., Lee, H., & Kim, D. (2022). Current status and security trend of OSINT. *Wireless Communications and Mobile Computing, 2022(*1), 1290129. DOI: https://doi.org/10.1155/2022/1290129

[3] Evangelista, J. R. G., Sassi, R. J., Romero, M., & Napolitano, D. (2021). Systematic literature review to investigate the application of open-source intelligence (OSINT) with artificial intelligence. *Journal of Applied Security Research, 16*(3), 345–369. DOI: https://doi.org/10.1080/19361610.2020.1761737

[4] Anisimov, A. V., Marchenko, O. O., & Nikonenko, A. O. (2008). Alhorytmichna model asotsiatyvno-semantychnoho kontekstnoho analizu tekstiv pryrodnoiu movoiu [Algorithmic model of associative-semantic contextual analysis of texts in natural language]. Problemy prohramuvannia, (2–3), 379–384. [in Ukrainian].

[5] Lytvynenko, O. Ye., & Burko, D. A. (2009). Modeli semantychnoho analizu tekstiv [Models of semantic text analysis]. Naukoiemni tekhnolohii, *4*(4), 55–58. [in Ukrainian].

[6] Maulud, D. H., Zeebaree, S. R., Jacksi, K., Sadeeq, M. A. M., & Sharif, K. H. (2021). State of art for semantic analysis of natural language processing. *Qubahan Academic Journal, 1*(2), 21–28.

[7] Velupillai, S., Mowery, D., South, B. R., Kvist, M., & Dalianis, H. (2015). Recent advances in clinical natural language processing in support of semantic analysis. *Yearbook of Medical Informatics, 24*(1), 183–193.

[8] Molenaar, A., Lukose, D., Brennan, L., Jenkins, E. L., & McCaffrey, T. A. (2024). Using natural language processing to explore social media opinions on food security: Sentiment analysis and topic modeling study. *Journal of Medical Internet Research, 26,* e47826.

[9] Liu, J., Li, K., Zhu, A., Hong, B., Zhao, P., Dai, S., ... & Su, H. (2024). Application of deep learning-based natural language processing in multilingual sentiment analysis. *Mediterranean Journal of Basic and Applied Sciences (MJBAS), 8(*2), 243–260.

[10] Block, L. (2024). The long history of OSINT. *Journal of Intelligence History, 23*(2), 95-109.

[11] Analytics of Big Data and social networks. (n.d.). Retrieved from http://www.osp.ru/os/2013/08/13037856.

[12] Kontostathis, A., Edwards, L., & Leatherman, A. (2010). Text mining and cybercrime. In M. W. Berry & J. Kogan (Eds.), *Text mining: Applications and theory* (pp. 149–164). Wiley.

[13] Dua, S., & Du, X. (2011). *Data mining and machine learning in cybersecurity*. New York.

[14] Berry, M. W., & Kogan, J. (Eds.). (2010). *Text mining: Applications and theory*. Wiley. 224 p.

[15] Thorsten, J. (1998). Text categorization with support vector machines: Learning with many relevant features. Retrieved from https://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf.

[16] Sebastiani, F. (2010). Machine learning in automated text categorization. *ACM Computing Surveys*, 1–47.

[17] Esuli, A., & Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification. *Conference of Information and Knowledge Management (Bremen)*, ACM, New York, NY, 617–624.

[18] Gupta, M., Mishra, A., & Kumar, V. (2020). A survey on fake news detection using machine learning techniques. *Journal of Big Data, 7*(1), 1–21.

[19] Zhang, Z., Chen, Y., & Chen, L. (2020). Methods for detecting fake news: A survey. *ACM Computing Surveys, 53*(6), 1–37.
[20] Browne, T. O., Abedin, M., & Chowdhury, M. J. M. (2024). A systematic review on research utilising artificial intelligence for open-source intelligence (OSINT) applications. *International Journal of Information Security, 23*(4), 2911-2938.

Надійшла до редколегії 20.03.2025