DOI 10.15276/imms.v15.no4.536
УДК 004.056.53:004.912

Informatics and Mathematical Methods in Simulation
Vol.15 (2025), No. 4, pp. 536-544

СИСТЕМА АВТОМАТИЗОВАНОГО АНАЛІЗУ ЦИФРОВОГО СЛІДУ КОРИСТУВАЧА В СОЦІАЛЬНИХ МЕДІА З ВИКОРИСТАННЯМ OSINT

А. В. Власова, В. О. Назаров, І. А. Ярова, Н. І. Кушніренко

Національний університет «Одеська Політехніка»
1, Шевченка пр., Одеса, 65044, Україна
Email: kushnirenko@op.edu.ua

У статті представлено систему автоматизованого аналізу цифрового сліду користувача в соціальних медіа з використанням підходів OSINT, адаптованого для української мови. Зростання активності в соціальних мережах створює ризики для приватності через накопичення персональної інформації. Метою роботи є створення спеціалізованої системи для комплексного аналізу цифрової присутності користувачів у Telegram та YouTube із урахуванням морфологічних особливостей української мови та культурного контексту. Проведено аналіз існуючих методів дослідження соціальних платформ, виявлено обмеження при роботі з українським контентом через недостатню точність розпізнавання мови та відсутність спеціалізованих лінгвістичних моделей. Запропонована система складається з трьох послідовних алгоритмів: алгоритму збору даних через API Telegram та YouTube, алгоритму комплексного аналізу текстового контенту з використанням TF-IDF для виділення ключових термінів та langdetect для автоматичного визначення мови, алгоритму побудови інтегрованого профілю з нормалізацією різнотипних даних на основі текстового, соціального, часового та поведінкового компонентів. Експериментальне тестування на вибірці зі ста користувачів різних вікових категорій показало точність системи 73% для українського контенту, що на 18-25% перевищує міжнародні аналоги при середньому часі обробки 85-95 секунд на користувача. Значення F1-score становлять 0.79 для збору даних, 0.66 для алгоритму текстового аналізу та 0.64 для алгоритму побудови профілю. Наукова новизна полягає у створенні спеціалізованої системи для української мови з урахуванням її морфологічних особливостей. Результати можуть використовуватися спеціалістами з кібербезпеки для аудиту цифрового сліду, дослідниками соціальних мереж для аналізу поведінкових патернів та фахівцями OSINT для верифікації інформації з відкритих джерел.
Ключові слова: OSINT, цифровий слід, соціальні медіа, Telegram, YouTube, машинне навчання, аналіз тексту.

Вступ. Зростання активності користувачів у соціальних мережах призводить до формування значних обсягів цифрових слідів, які містять персональну інформацію про їх власників. Open Source Intelligence являє собою дисципліну збору та аналізу інформації з публічно доступних джерел для отримання розвідувальних висновків. Хоча методи роботи з відкритими джерелами використовувались протягом століть, сучасне розуміння OSINT сформувалось з розвитком інтернету та цифрових комунікацій.

Цифровий слід користувача формується через взаємодію з різноманітними онлайн-платформами та сервісами. Цей слід включає як свідомо залишені дані, такі як публікації в соціальних мережах, так і несвідомо створену інформацію, включаючи метадані, часові відмітки активності та патерни поведінки [1]. Дослідження показують, що навіть обмежений набір даних з соціальних мереж може розкрити значну кількість персональної інформації, що створює як можливості для легітимного аналізу, так і серйозні питання щодо приватності [2].

Методи аналізу соціальних мереж характеризуються різноманітністю підходів. Мережевий аналіз зосереджується на структурі зв'язків між користувачами та поширенні інформації. Контентний аналіз досліджує текстову та мультимедійну інформацію через сентимент-аналіз, тематичне моделювання та класифікацію тексту [3]. Поведінковий

аналіз вивчає патерни активності, часові характеристики та геопросторовий контекст. Інтегровані підходи поєднують мультимодальний аналіз та застосування машинного навчання для комплексного дослідження користувацької активності [4].

Сучасні дослідження все частіше використовують техніки машинного навчання для автоматизації процесу аналізу [5]. Алгоритми класифікації дозволяють категоризувати контент за тематикою чи тональністю. Нейронні мережі можуть обробляти складні мультимодальні дані, включаючи текст, зображення та відео одночасно. Особливого розвитку набули методи обробки природної мови для аналізу текстового контенту [6].

Telegram займає унікальну нішу серед месенджерів через поєднання приватного спілкування з публічними каналами та групами. Архітектура включає приватні чати, групи, супергрупи та канали. Особливістю Telegram є система пересилання повідомлень з збереженням інформації про оригінальне джерело, що створює можливості для відстеження поширення інформації між каналами [7]. YouTube представляє іншу модель цифрової присутності, орієнтовану на довготривалий відеоконтент. Платформа надає багатий набір метаданих для кожного відео, включаючи статистику переглядів, коментарі та інформацію про плейлисти [8].

Основними обмеженнями існуючих OSINT інструментів є орієнтація на англомовний контент, складність налаштування та використання, а також фрагментарність функціональності. Більшість рішень вирішують окремі завдання без можливості комплексного аналізу цифрового сліду користувача. Для української мови це особливо актуально через складну морфологію та відсутність великих анотованих корпусів даних [9].

Мета роботи. Метою даної роботи є розробка системи автоматизованого аналізу цифрового сліду користувача в соціальних медіа з використанням підходів OSINT, адаптованої для української мови та культурного контексту.

Для досягнення поставленої мети необхідно розв’язати такі завдання:

- розробити алгоритм збору публічно доступних даних з соціальних медіа;
- розробити алгоритм комплексного аналізу текстового контенту з урахуванням специфіки української мови;
- розробити алгоритм побудови інтегрованого цифрового профілю користувача;
- реалізувати зазначені алгоритми у складі єдиної системи;
- провести тестування розробленої системи та проаналізувати її ефективність.

Основна частина. Вибір Telegram та YouTube обумовлений їх популярністю серед української аудиторії та наявністю стабільних API для збору публічно доступних даних з каналів, груп, відео та коментарів.

Алгоритм збору даних реалізовано у вигляді послідовного процесу з п'яти основних етапів. Перший етап включає ініціалізацію та ідентифікацію користувача. Алгоритм отримує початкові дані користувача, такі як username або ID профілю, та перевіряє доступність профілю на обох платформах. Для Telegram використовується Telegram Bot API через офіційний клієнт, а для YouTube застосовується Data API v3. Другий етап передбачає збір базової інформації профілю користувача, включаючи ім'я користувача, опис профілю, аватар, дату створення акаунту, кількість підписників та підписок. Ця інформація формує базову структуру цифрового профілю користувача та служить основою для подальшого аналізу. Третій етап зосереджується на аналізі контентної активності користувача. Алгоритм збирає всі доступні публічні публікації, коментарі та реакції користувача. Для Telegram аналізуються повідомлення у публічних каналах та групах, включаючи переслані повідомлення, до яких користувач має доступ або є учасником. Для YouTube збираються відео користувача, його коментарі під власними та чужими відео, створені плейлисти та їх наповнення. Четвертий етап включає побудову соціального графу користувача через аналіз підписок на канали та користувачів, взаємодій з іншими учасниками спільнот, частоти спілкування з

конкретними контактами. П'ятий етап передбачає збір метаданих та темпоральної інформації, включаючи часові мітки активності, інформацію про пристрої користувача та інші технічні метадані.

На основі розробленого алгоритму створено програмний застосунок, який реалізує всі описані етапи збору та обробки даних. Для забезпечення стабільної роботи застосунку передбачено систему обробки помилок. При тимчасовій недоступності API платформи алгоритм переходить в режим очікування з поступовим збільшенням інтервалу між спробами підключення згідно з формулою:

$$t_{wait} = t_0 * 2^n \quad (1)$$

де t_0= 1 секунда, n – номер спроби підключення. Також реалізовано локальний кеш для оптимізації продуктивності та зменшення навантаження на зовнішні API. Час життя кешу визначається типом даних: 300 секунд для профільних даних користувачів, 60 секунд для динамічного контенту (публікації, коментарі) та 900 секунд для статичних метаданих (інформація про канали, групи). Така диференціація забезпечує баланс між актуальністю інформації та ефективним використанням обмежених квот API платформ.

Реалізація етапів 2-5 має специфічні особливості для кожної платформи. Адаптація алгоритму для Telegram включає роботу з месенджер-специфічними функціями через Telegram Bot API. Алгоритм використовує метод getChat для отримання інформації про публічні канали та групи, включаючи назву, опис, кількість учасників та іншу базову інформацію. Метод getChatMember дозволяє аналізувати статус користувача в публічних спільнотах, визначати його роль та рівень активності. Для отримання візуальної інформації застосовується метод getUserProfilePhotos, який забезпечує доступ до аватарів користувачів. Окрему увагу приділено обробці пересланих повідомлень, оскільки Telegram зберігає метадані про оригінальне джерело, що дозволяє відстежувати ланцюжки поширення інформації між каналами. Алгоритм також враховує специфіку публічних та приватних груп, адаптуючи методи збору даних відповідно до рівня доступності інформації.

Для YouTube адаптація алгоритму зосереджується на комплексному аналізі публічних каналів користувачів через YouTube Data API v3. Процес збору даних починається з методу channels.list, який надає базову статистику каналу: кількість підписників, загальну кількість переглядів, дату створення та опис каналу. Метод search.list використовується для отримання списку всіх публічних відео користувача з можливістю фільтрації за датою публікації та типом контенту. Для кожного відео алгоритм збирає детальні метадані через метод videos.list, включаючи назву, опис, теги, тривалість, статистику переглядів, лайків та дизлайків. Особливу цінність представляє аналіз коментарів через метод commentThreads.list, який дозволяє вивчати не тільки коментарі користувача під власними відео, але й його активність під відео інших авторів. Алгоритм також аналізує створені користувачем плейлисти через метод playlists.list та їх наповнення через playlistItems.list, що розкриває тематичні інтереси та патерни споживання контенту. Додатково враховуються дані про підписки користувача на інші канали, що формує уявлення про його соціальне оточення та сфери інтересів у відеоконтенті.

Розроблений алгоритм збору даних реалізує багаторівневу систему кешування для оптимізації продуктивності. Система обробки помилок адаптована до особливостей кожної платформи: Telegram Bot API має обмеження на кількість запитів та специфічні коди помилок, які потребують окремої обробки, а YouTube API має систему квот з лімітом 10000 одиниць на день, що вимагає ретельного планування запитів та оптимізації використання API ресурсів.

Зібрані дані з обох платформ обробляються незалежно та передаються до наступного етапу аналізу, де вони об'єднуються на рівні побудови інтегрованого профілю користувача. Для Telegram аналізуються доступні повідомлення користувача у публічних каналах та групах, частота активності, типи контенту який він публікує або

пересилає. Особлива увага приділяється аналізу форвардинга повідомлень, що може розкрити мережу інформаційних джерел користувача та його уподобання. Для YouTube адаптація зосереджується на аналізі публічних каналів користувачів, їх відео контенту, коментарів, плейлистів та підписок на інші канали. YouTube API дозволяє отримувати детальну статистику каналу, включаючи кількість переглядів, підписників, лайків та коментарів, що надає можливість оцінити рівень впливу користувача на платформі. Алгоритм аналізує метадані відео, включаючи заголовки, описи, теги, тривалість та дату публікації для побудови профілю інтересів користувача.

Текстовий аналіз займає центральне місце в системах OSINT, оскільки мовна поведінка користувачів розкриває найбільше інформації про їх особистісні характеристики, світогляд та соціальні зв'язки. Розроблений алгоритм поєднує традиційні методи обробки природної мови з сучасними підходами частотного аналізу, адаптованими для української мови.

Алгоритм текстового аналізу складається з п'яти етапів обробки. Перший етап включає нормалізацію та очищення тексту від технічних артефактів, HTML тегів, спеціальних символів та інших елементів, які не несуть семантичного навантаження. Процес нормалізації включає приведення тексту до стандартного формату UTF-8, корекцію кодування символів, обробку емоджі та спеціальних символів. Другий етап передбачає токенізацію та лематизацію тексту з урахуванням морфологічних особливостей української мови. Використовується спеціалізований токенізатор, який враховує українські конструкції, прийменники з апострофом, складні числівники та назви власні. Третій етап включає автоматичне визначення мови тексту з використанням бібліотеки langdetect, що важливо для багатомовного контенту користувачів. Алгоритм використовує статистичний підхід на основі частотного аналізу символьних n-грам для української та англійської мов. Визначення мови дозволяє коректно застосовувати відповідні лінгвістичні ресурси: для української мови використовуються спеціалізовані словники та токенізатори, адаптовані до морфологічних особливостей української мови, тоді як для англійської застосовуються стандартні бібліотеки обробки природної мови. Четвертий етап алгоритму реалізує сентимент-аналіз текстового контенту для визначення емоційного забарвлення повідомлень користувача. Використовується гібридний підхід, який поєднує словникові методи з елементами частотного аналізу. Алгоритм класифікує тексти за трьома основними категоріями емоційного забарвлення: позитивне, негативне та нейтральне. П'ятий етап включає частотний аналіз ключових слів та виділення сутностей з тексту користувача. Алгоритм використовує метод TF-IDF для автоматичного виявлення найбільш важливих термінів у корпусі текстів користувача. Вага терміну *t* у конкретному тексті *d* розраховується за формулою:

$$TF - IDF(t,d) = (f(t,d) / max\_freq(d)) * log(N / |\{d \in D : t \in d\}|) \quad (2)$$

де *f(t,d)* - частота терміну t у тексті d, *max_freq(d*) - максимальна частота будь-якого терміну в тексті d, *N* - загальна кількість текстових фрагментів користувача, $|\{d \in D : t \in d\}|$ - кількість текстових фрагментів, що містять термін t.

Також реалізовано обробку неструктурованого тексту з урахуванням інтернет-сленгу, скорочень та неологізмів, характерних для української інтернет-культури. Створено спеціалізований словник з понад 800 популярних українських скорочень з їх розшифровками та контекстуальними значеннями. Це суттєво підвищує точність аналізу сучасного українського цифрового контенту та зменшує кількість помилок при обробці нестандартних текстів.

Побудова комплексного цифрового профілю користувача є третім алгоритмом системи аналізу, який інтегрує результати всіх попередніх стадій обробки у єдину структуровану модель. Алгоритм побудови профілю функціонує через два основних етапи. Початковий етап включає нормалізацію та стандартизацію всіх зібраних даних до єдиної системи координат. Процес нормалізації передбачає приведення всіх метрик до стандартизованої шкали [0,1] за допомогою функції мін-макс нормалізації:

$$x\_norm = (x - x\_min) / (x\_max - x\_min) \quad (3)$$

де x - початкове значення, x_min та x_max - мінімальне та максимальне значення в діапазоні.

Соціальний компонент характеризує мережу соціальних зв'язків користувача, рівень його впливу через співвідношення підписників до підписок, та активність взаємодії з іншими користувачами через кількість коментарів, відповідей та реакцій на публікації інших користувачів. Часовий компонент відображає патерни активності користувача в часі, включаючи періоди найвищої активності, регулярність публікацій та зміни в поведінці протягом різних періодів. Поведінковий компонент аналізує типи контенту, які користувач споживає та створює через категоризацію за форматом (текст, відео, зображення) та тематикою, його реакції та взаємодії з різними типами матеріалів через частоту лайків, коментарів, репостів та переглядів. Другий етап передбачає розрахунок базових компонентів цифрового профілю. Текстовий компонент включає аналіз лексичного багатства користувача, визначення домінуючих тематик на основі частотного аналізу ключових термінів, оцінку емоційного профілю через співвідношення позитивних, негативних та нейтральних повідомлень, та аналіз стилістичних особливостей через довжину речень і використання специфічної лексики. Лексичне багатство розраховується за формулою TTR:

$$TTR = |V| / N \quad (4)$$

де $|V|$ - кількість унікальних слів, N - загальна кількість слів у корпусі текстів користувача.

Результати. Оцінка ефективності розробленої системи проводилась на публічно доступних профілях користувачів Telegram та YouTube різних вікових категорій та рівнів активності. Тестування проводилось на 100 користувачах, профілі яких були відібрані з публічних каналів та груп української тематики. Структура тестової вибірки представлена в табл. 1, яка демонструє розподіл учасників дослідження за віковими групами та рівнем активності на досліджуваних платформах.

Таблиця 1

Характеристики тестової вибірки користувачів

Віковий діапазон	Кількість користувачів	Рівень активності	Telegram	YouTube
18-25 років	36 (36%)	Високий	30	24
26-35 років	40 (40%)	Помірний	36	32
36-50 років	24 (24%)	Низький	20	16
Загалом	100 (100%)	-	86	72

Алгоритм побудови цифрового профілю показує варіативну точність залежно від повноти вхідних даних. Серед 100 досліджуваних користувачів 36 мали високу активність переважно у віці 18-25 років, демонструючи точність профілювання 78%. Користувачі з помірною активністю, які становили 40% вибірки переважно у віці 26-35 років, показали точність 66%. Користувачі з низькою активністю, переважно старші за 36 років і складаючи 24% вибірки, демонстрували точність профілювання 59%.

Аналіз продуктивності алгоритму збору даних показав середній час обробки одного користувача приблизно 85 секунд для Telegram та 95 секунд для YouTube. Дослідження точності алгоритмів текстового аналізу показало високі результати для української мови: визначення мови досягло точності 91%, сентимент-аналіз 74%, тематична класифікація 69%, виділення ключових слів 77%. Для англійської мови результати становили відповідно 92%, 71%, 67%, 74%. Змішаний контент демонстрував дещо нижчі показники: 87%, 65%, 64%, 69%. Особливі виклики для системи становлять тексти з високим рівнем іронії, сарказму або використанням сленгу, де точність аналізу

знижується, що пов'язано зі складністю автоматичного розпізнавання контекстуальних та культурних нюансів української інтернет-комунікації.

Порівняльний аналіз з існуючими OSINT інструментами представлено в табл. 2.

Таблиця 2

Аналіз ефективності системи відносно існуючих рішень

Система	Точність для української мови	Швидкість обробки	Платформи	Спеціалізація
Maltego	55%	180-240 сек	Багатоплатформна	Міжнародна
OSINT Framework	48%	200-280 сек	Багатоплатформна	Міжнародна
Social-Analyzer	52%	190-260 сек	Соціальні мережі	Міжнародна
Розроблена система	73%	85-95 сек	Telegram, YouTube	Українська

Результати показують переваги розробленої системи у спеціалізації на українському контенті, що забезпечує приріст точності на 18-25% порівняно з універсальними міжнародними рішеннями. Показники точності для систем-аналогів отримані шляхом тестування на власній вибірці з 50 українських профілів. Різниця у швидкості обробки пояснюється різним функціоналом систем: універсальні інструменти виконують ширше коло завдань, тоді як розроблена система оптимізована для конкретних платформ Telegram та YouTube.

Функція оцінки якості результатів базується на метриці F1-score:

$$F1 = 2*(precision*recall)/(precision+recall) \quad (5)$$

де $precision = TP/(TP+FP)$, $recall = TP/(TP+FN)$, TP - true positive, FP - false positive, FN - false negative. Середнє значення F1-score для трьох алгоритмів системи становить 0.79 для збору даних, 0.66 для текстового аналізу, 0.64 для побудови профілю, що свідчить про задовільну якість роботи алгоритмів з урахуванням складності завдань автоматичного аналізу соціальних медіа.

Аналіз помилок всієї системи в цілому виявив основні джерела неточностей, які представлені на рис. 1.

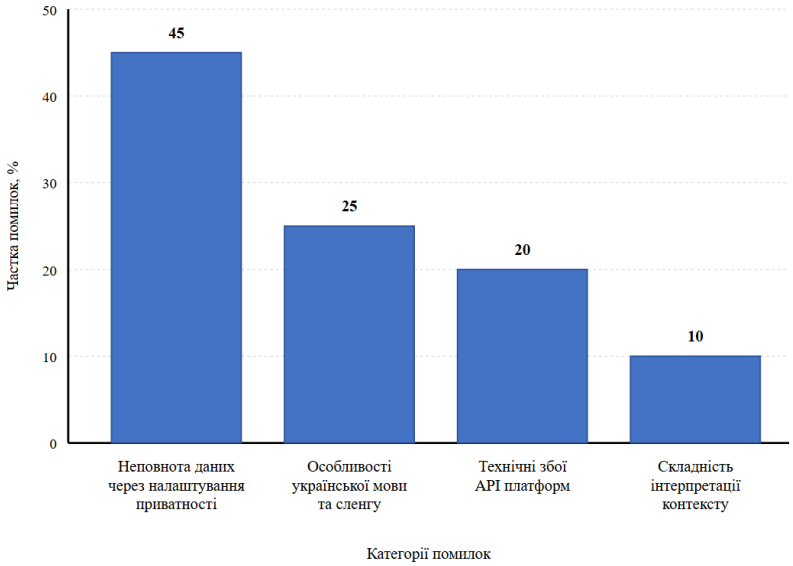

Рис. 1. Гістограма помилок системи

Найбільшу частку становлять проблеми з неповнотою даних через налаштування приватності користувачів (45%), особливості української мови та інтернет-сленгу (25%), технічні збої API платформ (20%) та складність інтерпретації контекстуальних значень (10%). Розподіл причин помилок показує, що найбільший вплив на якість аналізу має доступність даних, тоді як технічні та лінгвістичні виклики становлять менший, але значущий внесок у загальну похибку системи. Оцінка точності побудови комплексного

цифрового профілю проводилась шляхом порівняння результатів роботи системи з експертними оцінками (табл. 3).

Таблиця 3

Точність ідентифікації характеристик цифрового профілю

Характеристика	Точність	Кількість зразків
Основні інтереси користувача	75%	100
Емоційний профіль	70%	100
Рівень соціальної активності	77%	100
Професійна сфера діяльності	65%	87
Часові патерни активності	87%	76

Результати показують найвищу точність у визначенні часових патернів активності (87%), що пояснюється об'єктивним характером темпоральних даних та їх меншою залежністю від інтерпретації. Рівень соціальної активності визначається з точністю 77%, оскільки метрики підписок, коментарів та взаємодій піддаються прямому кількісному аналізу. Основні інтереси користувача ідентифікуються з точністю 75% завдяки застосуванню методу TF-IDF для виділення ключових термінів у текстовому контенті. Емоційний профіль визначається з точністю 70%, що є задовільним результатом з урахуванням складності сентимент-аналізу для української мови. Найнижчу точність демонструє визначення професійної сфери діяльності (65%), оскільки ця характеристика часто не виражена в публічному контенті користувачів.

Загальна точність системи для української мови розраховувалась як середньозважене значення точності визначення п'яти ключових характеристик користувача з таблиці 3: основні інтереси, емоційний профіль, рівень соціальної активності, професійна сфера та часові патерни. Середнє арифметичне цих показників становить 73% для користувачів з помірною активністю, що і використовується як базова метрика порівняння системи з аналогами.

Тестування стабільності роботи алгоритмів при різних умовах експлуатації показало задовільну стійкість до тимчасової недоступності API соціальних платформ. Система автоматичного відновлення з використанням експоненційного backoff забезпечує продовження роботи після усунення технічних проблем з відновленням 89% функціональності протягом 3-5 хвилин. При збільшенні навантаження система демонструє лінійне зростання часу обробки, зберігаючи можливість паралельної обробки до 5 профілів одночасно без суттєвого зниження продуктивності.

Дослідження масштабованості розробленої системи показало, що система ефективно справляється з обробкою користувачів різного рівня активності. Для користувачів з високою активністю (понад 500 публікацій) середній час обробки збільшується до 120-140 секунд, проте точність результатів зростає до 82% завдяки більшому обсягу даних для аналізу. Користувачі з помірною активністю (100-500 публікацій) обробляються за стандартний час 85-95 секунд з точністю 73%. Для користувачів з низькою активністю (менше 100 публікацій) час обробки скорочується до 60-70 секунд, однак точність знижується до 64% через обмежений обсяг вхідних даних.

Висновки. Розроблено систему автоматизованого аналізу цифрового сліду користувача в соціальних медіа з використанням підходів відкритого збору інформації, адаптовану для української мови та культурного контексту. Проведено комплексний аналіз існуючих методів дослідження соціальних платформ та виявлено їх основні обмеження при роботі з україномовним контентом. Розроблено алгоритм збору публічно доступних даних, який включає п'ять основних етапів та забезпечує систематичний підхід до отримання релевантних даних при дотриманні технічних обмежень. Реалізовано спеціалізовані адаптації для роботи з Telegram Bot API та YouTube Data API v3. Запропоновано алгоритм комплексного аналізу текстового контенту з використанням методу TF-IDF та бібліотеки langdetect для автоматичного визначення мови, адаптовану для української мови. Розроблено алгоритм побудови інтегрованого цифрового профілю,

який включає текстовий, соціальний, часовий та поведінковий компоненти. Експериментальне тестування підтвердило ефективність системи з точністю 73% для українського контенту, що на 18-25% перевищує показники міжнародних аналогів при середньому часі обробки 85-95 секунд на користувача.

Список літератури

1. Azucar D., Marengo D., Settanni M. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences.* 2018. Vol. 124. P. 150-159.
2. Feher K. Digital identity and the online self: Footprint strategies – An exploratory and comparative research study. *Journal of Information Science.* 2019. P. 016555151987970.
3. Dang N. C., Moreno-García M. N., De la Prieta F. Sentiment Analysis Based on Deep Learning: A Comparative Study. *Electronics.* 2020. Vol. 9. No. 3. P. 483.
4. Explainable AI for Psychological Profiling from Behavioral Data: An Application to Big Five Personality Predictions from Financial Transaction Records / Y. Ramon et al. *Information.* 2021. Vol. 12. No. 12. P. 518.
5. Nandwani P., Verma R. A review on sentiment analysis and emotion detection from text. Social Network Analysis and Mining. 2021. Vol. 11, no. 1.
6. Medhat W., Hassan A., Korashy H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*. 2014. Vol. 5. No. 4. P. 1093-1113.
7. Predicting Consumers' Decision-Making Styles by Analyzing Digital Footprints on Facebook / Y.-J. Chen et al. *International Journal of Information Technology & Decision Making.* 2019. Vol. 18. No. 02. P. 601-627.
8. Predicting Loneliness through Digital Footprints on Google and YouTube / E. Ahmed et al. *Electronics.* 2023. Vol. 12. No. 23. P. 4821.
9. Deeva I. Computational Personality Prediction Based on Digital Footprint of A Social Media User. *Procedia Computer Science*. 2019. Vol. 156. P. 185-193.

SYSTEM FOR AUTOMATED ANALYSIS OF USER DIGITAL FOOTPRINT IN SOCIAL MEDIA USING OSINT

A. Vlasova, V. Nazarov, I. Yarova, N. Kushnirenko

National Odesa Polytechnic University
1, Shevchenko Ave., Odesa, 65044, Ukraine
Email: kushnirenko@op.edu.ua

The article presents a system for automated analysis of user digital footprint in social media using Open Source Intelligence (OSINT) approaches adapted for the Ukrainian language. Growing user activity on social networks creates privacy risks through accumulation of personal information. The aim of the work is to create a specialized system for comprehensive analysis of user digital presence in Telegram and YouTube considering morphological features of Ukrainian language and cultural context. Analysis of existing social platform research methods revealed limitations when processing Ukrainian content, particularly insufficient recognition accuracy and absence of specialized linguistic models. The proposed system consists of three sequential algorithms: a data collection algorithm through Telegram and YouTube APIs, a comprehensive text content analysis algorithm using TF-IDF method for key term extraction and langdetect library for automatic language detection, and an integrated profile construction algorithm with heterogeneous data normalization based on textual, social, temporal, and behavioral components. Experimental validation on a sample of 100 users across different age categories demonstrated system accuracy of 73% for Ukrainian content with average processing time of 85-95 seconds per user. The highest accuracy was achieved in determining temporal activity patterns (87%) and social activity level (77%). Scientific novelty lies in creating a specialized system for Ukrainian language considering its morphological features, including integration of TF-IDF frequency analysis methods adapted for word forms specificity, development of Ukrainian internet slang dictionary with over 800 entries, and construction of a comprehensive digital profile model based on four components.

Keywords: OSINT, digital footprint, social media, Telegram, YouTube, machine learning, text analysis.