Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)
Кафедра комп'ютерних наук
(повна назва кафедри)

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня

магістр
(назва освітнього ступеня)

на тему: Дослідження взаємодії користувачів із контентом у соціальних мережах із застосуванням методів NLP та поведінкової аналітики

Виконав: студент VI курсу, групи СНнм-61
спеціальності 122 Комп'ютерні науки
(шифр і назва спеціальності)

_____ Боднар Д.В.
(підпис) (прізвище та ініціали)

Керівник _____ Липак Г.І.
(підпис) (прізвище та ініціали)

Нормоконтроль _____ Никитюк В.В.
(підпис) (прізвище та ініціали)

Завідувач кафедри _____ Боднарчук І.О.
(підпис) (прізвище та ініціали)

Рецензент _____ Микитишин А.Г.
(підпис) (прізвище та ініціали)

Тернопіль
2026

Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)

Кафедра комп'ютерних наук
(повна назва кафедри)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Боднарчук І.О.
(підпис) (прізвище та ініціали)
«_13_» _квітня_ 2026 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

на здобуття освітнього ступеня Магістр
(назва освітнього ступеня)

за спеціальністю 122 Комп'ютерні науки
(шифр і назва спеціальності)

Студенту Боднару Денису Володимировичу
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження взаємодії користувачів із контентом у соціальних мережах із застосуванням методів NLP та поведінкової аналітики

Керівник роботи Липак Галина Ігорівна, к. соц. ком., доцент кафедри КН
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджені наказом ректора від «10» березня 2026 року № 4/9-150

2. Термін подання студентом завершеної роботи

3. Вихідні дані до роботи Наукові публікації про методи обробки природної мови (NLP) та поведінкової аналітики для дослідження взаємодії користувачів у соціальних мережах; збору та аналізу текстових даних соціальних платформ Telegram, Facebook, Instagram та X.

4. Зміст роботи (перелік питань, які потрібно розробити)
Вступ. Розділ 1. Теоретичні основи аналізу взаємодії користувачів у соціальних мережах. 1.1 Соціальні мережі як джерело даних. 1.2 Типи користувацьких взаємодій із контентом. 1.3 Методи збору та обробки даних. 1.4 Основи обробки природної мови. 1.5 Поведінковий 1.6 Висновок до розділу. Розділ 2. Методи та інструменти аналізу взаємодій користувачів. 2.1 Аналіз тональності та емоційного забарвлення контенту. 2.2 Кластеризація та тематичний аналіз. 2.3 Графові моделі та аналіз соціальних взаємодій. 2.4 Порівняльний аналіз методів NLP та поведінкової аналітики. 2.5 Інструменти та бібліотеки для аналітики. 2.6 Висновок до розділу. Розділ 3. Дослідження та порівняння методів аналізу взаємодії. 3.1 Опис об'єкта та джерел даних. 3.2 Методика проведення дослідження. 3.3 Програмна реалізація 3.4 Результати аналізу тональності повідомлень. 3.5 Поведінкова кластеризація користувачів. 3.6 Аналіз соціальних графів та структурних патернів. 3.7 Взаємозв'язок між тональністю та залученістю аудиторії. 3.8 Аналіз часових патернів та платформенні відмінності 3.9 Перехресний аналіз результатів 3.10 Висновки до розділу. Розділ 4. Охорона праці та безпека в надзвичайних ситуаціях. 4.1 Вимоги ерногоміки до організації робочого місця оператора ПК. 4.2 Пожежна профілактика на робочому місці. 4.3 Висновок до розділу. Висновки. Перелік джерел. Додатки

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень, слайдів)
1. Титульний слайд. 2. Актуальність та мета. 3. Об'єкт і предмет дослідження. 4. Завдання дослідження. 5. Типи користувацьких взаємодій. 6. Основи NLP. 7. Аналіз тональності. 8. Кластеризація. 9. Графові моделі. 10. Порівняльна таблиця. 11. Програмна реалізація. 12. Результати аналізу тональності. 13. Поведінкова кластеризація. 14. Взаємозв'язок тональності та залученості 15. Висновок.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Охорона праці			
Безпека в надзвичайних ситуаціях	Теслюк В.М., проректор з адміністративно-господарської роботи та будівництва		

7. Дата видачі завдання 13 квітня 2026 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1.	Ознайомлення з завданням до кваліфікаційної роботи	13.04.2026	
2.	Підбір та опрацювання наукових публікацій, збір	13.04.2026-20.04.2026	
	даних по темі роботи		
3.	Виконання дослідження згідно теми кваліфікаційної	21.04.2026-03.05.2026	
	роботи		
4.	Оформлення розділу «Теоретичні основи аналізу	04.05.2026-10.05.2026	
	взаємодії користувачів у соціальних мережах»		
5.	Оформлення розділу «Методи та інструменти аналізу	04.05.2026-10.05.2026	
	взаємодій користувачів»		
6.	Оформлення розділу «Дослідження та порівняння	04.05.2026-10.05.2026	
	методів аналізу взаємодії користувачів у соціальних		
	мережах»		
7.	Виконання завдання до підрозділу «Охорона праці»	27.04.2026-10.05.2026	
8.	Виконання завдання до підрозділу «Безпека в	27.04.2026-10.05.2026	
	надзвичайних ситуаціях»		
9.	Оформлення кваліфікаційної роботи	11.05.2026-13.05.2026	
10.	Нормоконтроль	14.05.2026	
11.	Перевірка на плагіат	15.05.2026	
12.	Попередній захист кваліфікаційної роботи	18.05.2026	
13.	Захист кваліфікаційної роботи	27.05.2026	

Студент _____ Боднар Д.В.
(підпис) (прізвище та ініціали)

Керівник роботи _____ Липак Г.І.
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Дослідження взаємодії користувачів із контентом у соціальних мережах із застосуванням методів NLP та поведінкової аналітики// Кваліфікаційна робота освітнього рівня «Магістр» // Боднара Дениса Володимировича// Тернопільський національний технічний університет імені Івана Пулюя, факультет комп'ютерно-інформаційних систем і програмної інженерії, кафедра комп'ютерних наук, група СНнм-61 // Тернопіль, 2026 // С. 88, рис. – 18, табл. – 16, кресл. – 15, додат. – 2, бібліогр. – 55.

Ключові слова: нейронні мережі, обробка природної мови, аналіз тональності, соціальні мережі, поведінкова аналітика, кластеризація, графовий аналіз, машинне навчання

Кваліфікаційна робота присвячена розробці методики аналізу взаємодії користувачів із контентом у соціальних мережах засобами NLP та поведінкової аналітики.

В першому розділі кваліфікаційної роботи описані теоретичні основи аналізу взаємодії користувачів у соціальних мережах. Висвітлено типи взаємодій із контентом та методи збору даних. Розглянуто основи обробки природної мови. Проаналізовано підходи до поведінкового аналізу користувачів.

В другому розділі кваліфікаційної роботи систематизовано методи аналізу тональності контенту. Досліджено методи кластеризації та тематичного моделювання. Подано порівняльний аналіз методів NLP та поведінкової аналітики.

В третьому розділі кваліфікаційної роботи описано дослідницький конвеєр для аналізу взаємодій користувачів. Проаналізовано результати інтеграції аналізу тональності, кластеризації та графового аналізу. Проведено апробацію на наборі даних чотирьох платформ. Об'єкт дослідження: процеси взаємодії користувачів із контентом у соціальних мережах.

ANNOTATION

Analysis of User Interaction with Social Media Content Using NLP Methods and Behavioral Analytics // The educational level "Master" qualification work // Bodnar Denys Volodymyrovych // Ternopil Ivan Pulyuy National Technical University, Faculty of Computer Information Systems and Software Engineering, Department of Computer Science, SNnm-61 group // Ternopil, 2025 // P. 88, fig. – 18, tables – 16, posters – 15, annexes – 2, ref. – 55.

Key words: neural networks, natural language processing, sentiment analysis, social media, behavioural analytics, clustering, graph analysis, machine learning.

This thesis is devoted to the development of a methodology for analysing user interaction with content on social media using NLP and behavioural analytics.

The first chapter of the thesis describes the theoretical foundations of analysing user interaction on social media. It outlines the types of interaction with content and data collection methods. The fundamentals of natural language processing are examined. Approaches to behavioural analysis of users are analysed.

The second chapter of the thesis systematises methods for analysing the sentiment of content. Methods of clustering and thematic modelling are investigated. A comparative analysis of NLP and behavioural analytics methods is presented.

The third chapter of the thesis describes a research pipeline for analysing user interactions. The results of integrating sentiment analysis, clustering and graph analysis are analysed. Testing was carried out on a dataset from four platforms. Research object: processes of user interaction with content on social networks.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

API (англ. Application Programming Interface) – інтерфейс програмування застосунків.

BERT (англ. Bidirectional Encoder Representations from Transformers) – двонаправлена модель представлення тексту на основі трансформерів.

CSV (англ. Comma-Separated Values) – формат даних з роздільниками-комами.

LSTM (англ. Long Short-Term Memory) – довга короткострокова пам'ять, тип рекурентної нейронної мережі.

ML (англ. Machine Learning) – машинне навчання.

NLP (англ. Natural Language Processing) – обробка природної мови.

RNN (англ. Recurrent Neural Network) – рекурентна нейронна мережа.

SVM (англ. Support Vector Machine) – метод опорних векторів.

TF-IDF (англ. Term Frequency – Inverse Document Frequency) – частотно-зворотна міра важливості терміна.

UBA (англ. User Behavior Analytics) – аналітика поведінки користувачів.

ЗМІСТ

ВСТУП

Соціальні мережі стали невід'ємною частиною сучасного інформаційного простору: щодня платформи Telegram, Facebook, Instagram та X генерують мільярди текстових повідомлень, реакцій і коментарів, у яких відображаються суспільні настрої, комунікаційні патерни та поведінкові тенденції аудиторії. Проте через неструктурованість та різноманітність цих даних їх автоматизована обробка залишається нетривіальним науковим завданням. Методи обробки природної мови і поведінкової аналітики є перспективними інструментами для вирішення цієї задачі, однак питання їх порівняльної ефективності та оптимального поєднання досі залишаються відкритими. Тому дослідження взаємодії користувачів із контентом у соціальних мережах із застосуванням сучасних методів NLP та поведінкової аналітики є актуальним напрямком наукових досліджень.

Метою даної кваліфікаційної роботи освітнього рівня «Магістр» є підвищення рівня повноти подання інформації щодо взаємодії користувачів із контентом у соціальних мережах на основі методів NLP та поведінкової аналітики. Для досягнення поставленої мети необхідно виконати такі завдання:

– проаналізувати стан досліджень у галузі аналізу взаємодії користувачів у соціальних мережах та виявити актуальні підходи до збору й обробки даних;

– дослідити існуючі методи NLP, зокрема аналіз тональності, тематичне моделювання та методи на основі глибокого навчання;

– проаналізувати методи поведінкової аналітики та графового аналізу соціальних мереж;

– виконати порівняння існуючих методів NLP та поведінкової аналітики за критеріями точності, обчислювальної складності та практичної застосовності;

– розробити та визнати комплексну методику дослідження взаємодії користувачів із контентом у соціальних мережах із застосуванням обраних методів.

Об'єкт дослідження: процеси взаємодії користувачів із контентом у соціальних мережах.

Предмет дослідження: методи NLP та поведінкової аналітики для дослідження взаємодії користувачів із контентом у соціальних мережах.

Наукова новизна одержаних результатів кваліфікаційної роботи полягає в тому, що отримано подальший розвинений метод аналізу взаємодії користувачів із контентом у соціальних мережах, де на відміну від відомих підходів, запропонований метод інтегрує з аналізом тональності на основі трансформерних архітектур, кластеризацій поведінкових профілів та графових аналізів соціальних взаємодій в єдиний аналітичний конвеєр, що дозволяє отримати більш повне уявлення про закономірності цифрової комунікації.

Реалізовано дослідницький конвеєр для збору, обробки та аналізу даних взаємодій користувачів у соціальних мережах, схвалений на наборі даних, що відтворює реальну активність користувачів на чотирьох платформах. Отримані результати можуть бути використані як основа при розробці практичних систем моніторингу та аналітики соціальних медіа.

Основні результати проведених досліджень опубліковано у збірнику матеріалів конференції та обговорювались на студентських науково-технічних конференціях Тернопільського національного технічного університету імені Івана Пулюя.

Основні результати кваліфікаційної роботи опубліковано у двох працях конференції (Див. додатки А).

Кваліфікаційна робота складається зі вступу, чотирьох розділів, висновків, списку літератури з 55 найменувань та 2 додатків. Загальний обсяг кваліфікаційної роботи складає 100 сторінок, з них 88 сторінки основного тексту, який містить 18 рисунків та 16 таблиць.

РОЗДІЛ 1. ТЕОРЕТИЧНІ ОСНОВИ АНАЛІЗУ ВЗАЄМОДІЇ КОРИСТУВАЧІВ У СОЦІАЛЬНИХ МЕРЕЖАХ

1.1. Соціальні мережі як джерело даних

У сучасному інформаційному суспільстві соціальні мережі трансформувалися з простих платформ для комунікації у найпотужніше джерело емпіричних даних. Дослідники розглядають їх як глобальну систему фіксації «цифрових слідів» людської діяльності. На відміну від традиційних методів збору інформації, дані із соцмереж дозволяють спостерігати за поведінкою користувачів у їхньому природному середовищі без прямого втручання дослідника.

Використання цих платформ як джерела інформації зумовлене їхньою здатністю акумулювати різнорідні типи даних (текст, медіаконтент, метадані, структури зв’язків), що є цінним ресурсом для методів інтелектуального аналізу (Data Mining) та машинного навчання [1]. Формування консолідованих сховищ соціально-комунікаційних даних є актуальним завданням сучасних інформаційних систем – інтеграція ресурсів різнорідних цифрових платформ дозволяє отримати повніше уявлення про комунікаційні патерни та поведінку аудиторії [2]. Спостерігається стійка тенденція до зростання залежності професійних медіа від цифрових платформ: якщо раніше соцмережі були лише допоміжним інструментом, то зараз вони виступають фундаментом для створення понад 40% новинного потоку (рис. 1.1).

Класифікація потоків даних у соціальних мережах:

1. Текстові дані. Це основний масив інформації, що включає дописи, коментарі та повідомлення. Для інформаційних систем ці дані є базою для семантичного аналізу, виявлення тональності та автоматичної класифікації тем.

2. Реляційні дані. Аналіз цих даних дозволяє будувати графи соціальних комунікацій, виявляти центри впливу та досліджувати архітектуру розповсюдження інформаційних хвиль.

3. Метадані. Додаткова технічна інформація, така як часові мітки, геопозиціонування, ідентифікатори пристроїв та мовні налаштування. Ці дані дозволяють проводити часовий та просторовий аналіз активності аудиторії.

4. Мультимедійні дані. Візуальний контент, який за допомогою алгоритмів комп'ютерного зору може бути перетворений у структуровані теги для подальшого аналізу вподобань або подій.

Рисунок 1.1 – Джерела в українських медіа

Диференціація джерел даних у межах соціальних платформ зумовлена технічними особливостями кожної мережі та характером інформації, що в них циркулює. Зокрема, такі ресурси як Telegram та X виступають джерелами швидкоплинних інформаційних потоків. Вони характеризуються високою швидкістю оновлення контенту та стислістю повідомлень, що робить їх ідеальними для систем моніторингу подій у реальному часі. Використання даних із цих платформ дозволяє алгоритмам оперативно виявляти аномальні сплески

активності та здійснювати детекцію критичних подій (Event Detection) на основі текстових маркерів та хештегів.

Натомість платформи Facebook або LinkedIn є джерелами глибоких структурованих даних. Вони надають розширену інформацію про профілі користувачів, їхні соціальні ієрархії та професійні зв’язки. В інформаційних системах ці дані застосовуються для побудови складних прогнозних моделей поведінки та багатофакторної сегментації аудиторій. Аналіз реляційних зв'язків у таких мережах дозволяє відстежувати довготривалі тренди та стабільні комунікаційні структури, що значно розширює можливості інтелектуального аналізу порівняно з простим моніторингом текстових повідомлень.

Використання соціальних мереж як джерел даних потребує вирішення низки технічних викликів. По-перше, це проблема неструктурованості (понад 80% даних не мають чіткої схеми), що вимагає застосування методів NLP. По-друге, це шум (спам, боти, дублікати), який необхідно фільтрувати на етапі попередньої обробки даних [3].

1.2. Типи користувацьких взаємодій із контентом

Контент – це наповнення вебсайтів, блогів, сторінок у соціальних мережах або електронних листів. Через нього здійснюється взаємодія з користувачами, які звертаються за інформацією, корисними порадами або товарами і послугами. Окрім тексту, таке наповнення охоплює візуал, відео та різноманітні креативи. В основі створення контенту лежать кілька ключових правил: він має бути унікальним, якісним та по-справжньому корисним для людей. Також важливо стежити за грамотністю, триматися однієї стилістики та чітко відповідати темі.

Ефективна стратегія взаємодії з аудиторією в цифровому середовищі базується на гармонійному поєднанні різних типів контенту. Кожен вид виконує специфічні завдання: від формування первинної довіри до безпосереднього

стимулювання збуту. Нижче наведено детальну характеристику чотирьох ключових видів контенту.

1. Інформаційний контент. Фундаментом будь-якого процесу продажу в інформаційному суспільстві є встановлення стійких довірчих відносин із потенційним споживачем. Інформаційний контент виступає базовим інструментом демонстрації професійної компетенції та експертності суб'єкта комунікації (рис.1.2).

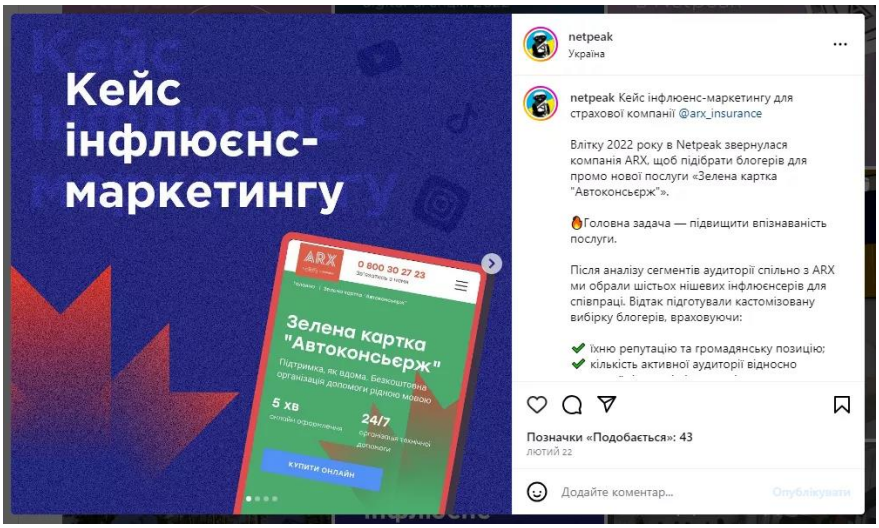

Рисунок 1.2 – Приклад інформаційного контенту

Оптимальна частка таких публікацій у загальній структурі контент-плану має становити близько 50%, оскільки надлишок прямої реклами без підтвердженої цінності призводить до відторгнення аудиторії.

Для досягнення мети інформування та підтвердження професіоналізму використовують такі формати:

- Репутаційні матеріали та експертне підтвердження. Висвітлення професійних здобутків, демонстрація дипломів, сертифікатів та результатів проходження профільних тренінгів. Це формує образ фахівця, якому можна довіряти.
- Аналітика та тренди галузі. Публікація думок лідерів ринку, рейтингів, прогнозів та новітніх методик. Використання інфографіки в таких

дописах дозволяє візуалізувати складні дані, роблячи їх доступними для швидкого засвоєння.

- Огляди та технічні характеристики. Детальний опис функціональних можливостей товарів, порівняння різних моделей, відповіді на часті запитання (FAQ), а також розвінчування поширених міфів про продукт.
- Корпоративна ідентичність. Розкриття місії та історії компанії, представлення ключових фахівців, репортажі з виробництва та робочих процесів. Це висвітлює бренд і створює ефект присутності для клієнта.

2. Розважальний контент. Головною метою даного типу матеріалів є емоційне розвантаження аудиторії та запобігання «втомі» від надмірного потоку професійної інформації. Розважальний контент дозволяє залучити пасивних користувачів до взаємодії, підвищуючи рівень їхньої лояльності через позитивні асоціації з брендом (рис. 1.3).

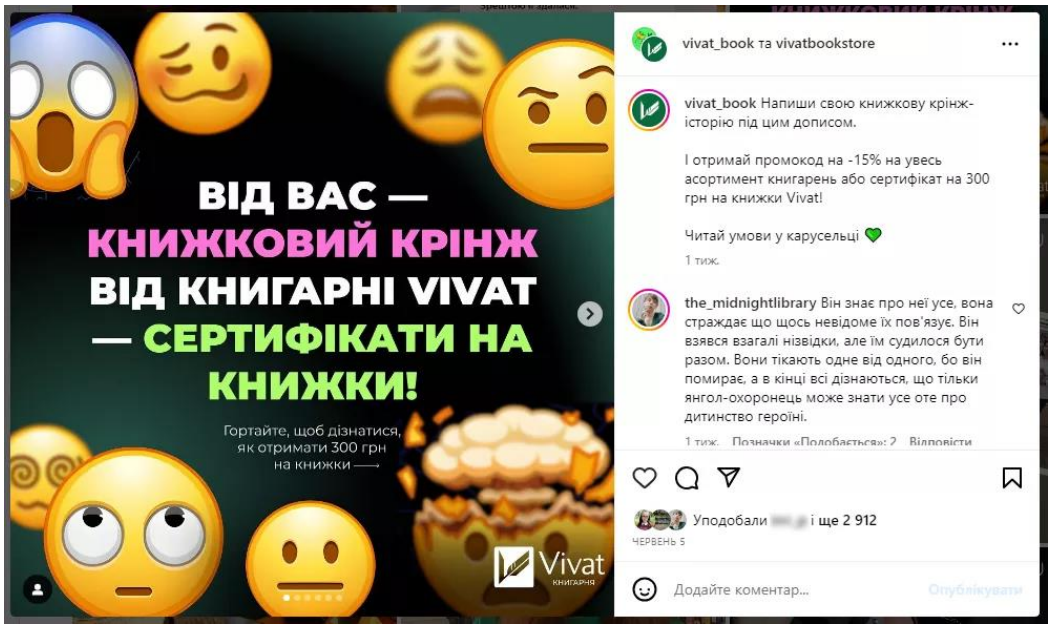

Рисунок 1.3 – Приклад розважального контенту

До основних інструментів розважального впливу на аудиторію належать:

- Інтерактивні форми та гейміфікація. Проведення опитувань, квестів та інтелектуальних ігор. Важливо демонструвати зворотний зв'язок,

підтверджуючи, що думка аудиторії реально впливає на розвиток проєкту чи вибір асортименту.

- Конкурсні механіки. Розіграші та марафони (челенджі), де учасники виконують певні дії для досягнення спільного результату. Це створює відчуття спільноти та стимулює віральне поширення інформації.
- Емоційні та ситуативні публікації. Використання актуальних мемів, тематичного гумору, атмосферних фотографій та цитат, що відповідають цінностям бренду та створюють відповідний настрій у підписників.

3. Контент, що залучає (інтерактивний). Цей тип спрямований на стимулювання активної двосторонньої комунікації та підвищення коефіцієнта залученості (Engagement Rate). Він критично важливий для алгоритмічного просування в соціальних мережах, оскільки стимулює користувачів залишати коментарі, ставити вподобання та зберігати публікації (рис. 1.4).

Рисунок 1.4 – Зразок інтерактивного контенту

Основними видами контенту, що стимулюють активність користувачів, є:

- Дискусійні та дискусійні формати. Публікація статей-думок на гострі теми, провокація конструктивних обговорень або спростування загальноприйнятих думок. Це змушує аудиторію висловлювати власну позицію.
- Технологічні інтерактивні інструменти. Впровадження вікторин, тестів, онлайн-калькуляторів вартості або анімованої графіки, з якою користувач може безпосередньо взаємодіяти.
- Реакція на актуальний порядок денний. Дописи про актуальні події, що дозволяє бренду залишатися в контексті сучасного інформаційного поля.

4. Контент, що продає (комерційний). Після завершення етапів формування довіри та лояльності, впроваджується контент, спрямований на безпосередню конверсію. Саме ці матеріали трансформують читача у покупця та забезпечують фінансову ефективність маркетингової стратегії (рис. 1.5).

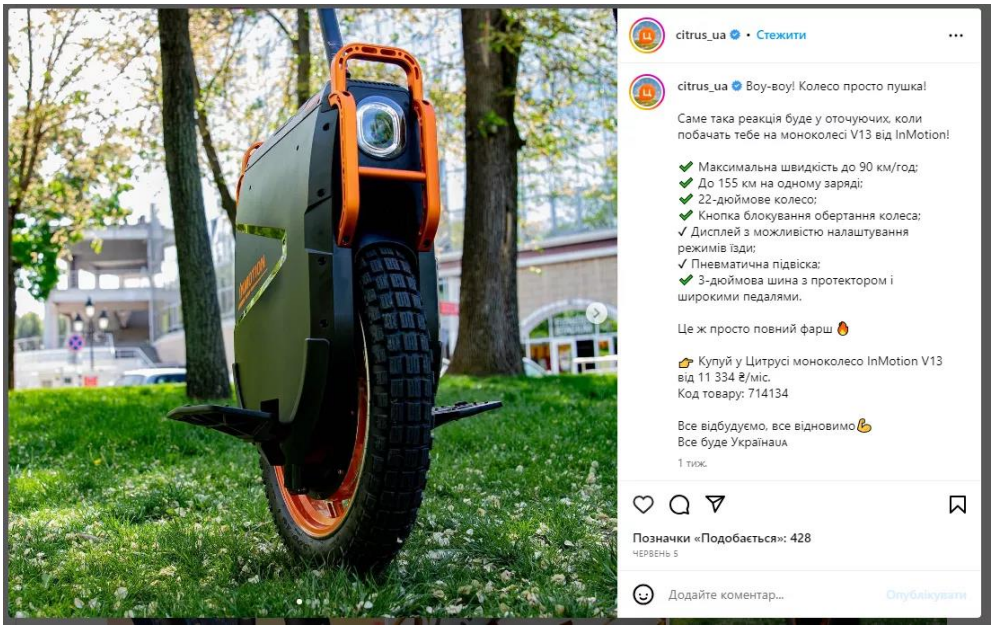

Рисунок 1.5 – Приклад комерційного контенту

Комерційна складова контент-стратегії реалізується через такі елементи:

- Пряма презентація та офери. Демонстрація продукту в дії, формування унікальної торгової пропозиції, акцент на перевагах та вигодах. Використання формату «до та після» наочно підтверджує ефективність пропозиції.

- Соціальні докази. Публікація реальних відгуків, успішних кейсів, історій клієнтів про вирішення їхніх проблем за допомогою продукту. Важливим елементом є використання фотографій із лідерами думок чи відомими особистостями.
- Стимулювання збуту. Анонсування обмежених у часі акцій, знижок, розпродажів або ексклюзивних пропозицій, що спонукають до прийняття швидкого рішення про покупку [4].

1.3. Методи збору та обробки даних

Методи збору та обробки даних – це систематизований процес, що охоплює отримання первинної інформації, її очищення, структурування та перетворення на корисні знання за допомогою програмно-технічних засобів. Це є основою для функціонування інформаційних систем, машинного навчання, Data Science та автоматизації процесів. Сучасні інформаційні системи все частіше використовують технології доповненої та змішаної реальності для представлення даних, що відкриває нові можливості у проєктуванні інтерактивних сервісів збору та відображення інформації [5].

Збір даних сьогодні – це не просто завантаження файлу, а складне архітектурне рішення, де програма виступає активним агентом. У сучасному стеку технологій цей процес перетворився на багаторівневу інженерну екосистему, де ключову роль відіграє здатність до адаптивної фільтрації ще на етапі захоплення сигналу. Наприклад, у хмарних технологіях збір часто реалізується через розгалужену мережу проміжних вузлів, які перехоплюють потоки інформації безпосередньо в момент їх виникнення. Коли мова йде про високонавантажені системи, програма-агент не просто копіює дані, а проводить їх первинну сепарацію, відсікаючи надлишкову інформацію безпосередньо на «краю» мережі (Edge Computing). Це може бути перехоплення пакетів у мережевому трафіку або зчитування подій з мікроконтролерів у реальному часі. Процес збору критично залежить від протоколів передачі, оскільки саме вони

визначають, чи будуть дані отримані в повному обсязі з гарантованою доставкою кожного байта, чи система пожертвує частиною пакетів заради швидкості та безперервності потоку [6].

Коли дані потрапляють у систему, починається етап інтелектуальної трансформації, який виходить далеко за межі простої фільтрації. Обробка в сучасних реаліях часто передбачає застосування алгоритмічних моделей, які на льоту розпізнають структуру в неструктурованому хаосі. Це означає, що система не просто чекає на команду людини, а самостійно класифікує вхідні потоки, визначаючи їх пріоритетність та релевантність для конкретної задачі [7]. Завдяки впровадженню концепції «data-driven» архітектури, алгоритми здатні самостійно підлаштовуватися під динамічні зміни у вхідних структурах без постійного втручання розробника. Наприклад, методи векторизації перетворюють текстову чи візуальну інформацію у багатовимірні числові масиви, з якими комп’ютер може виконувати складні математичні операції, використовуючи тисячі ядер графічних або тензорних процесорів для паралельних обчислень [8].

Кінцевим етапом обробки стає не просто запис у пам'ять, а створення логічних зв’язків між новими даними та тими, що вже існують у системі. У цьому контексті особливого значення набуває семантична індексація – процес, при якому система виявляє приховані концептуальні зв’язки, дозволяючи програмі «розуміти» контекст інформації, а не просто зберігати набір байтів. Це створює ефект «пам'яті контексту», коли система здатна розпізнати аномалію або важливий тренд на основі порівняння поточного моменту з накопиченим історичним досвідом. Таким чином, дані проходять шлях від фізичного імпульсу до «живої» інформаційної моделі, яка постійно оновлюється та виступає цифровим активом, готовим для використання в автоматизованих системах прийняття рішень [9].

1.4. Основи обробки природної мови

Обробка природної мови постає як інноваційна сфера штучного інтелекту, що забезпечує можливість комунікації між людьми та технічними системами за допомогою повсякденного мовлення. Прикладами практичного втілення цієї технології є функціонування голосових помічників, таких як Siri, або робота сервісів Google Translate. В обох випадках програмне забезпечення демонструє здатність інтерпретувати людську мову та генерувати адекватні відповіді.

Технології NLP інтегровані в щоденну взаємодію з мобільними та стаціонарними пристроями. Це проявляється у використанні асистентів Siri, Alexa чи Cortana, в автоматичних алгоритмах фільтрації спаму в Gmail, роботі онлайн-перекладачів та інструментах корекції текстів, таких як Grammarly [10].

Оскільки більшість користувачів не є програмістами, NLP стає критично важливим інструментом. Ця технологія дозволяє взаємодіяти навіть зі складним програмним забезпеченням без потреби в спеціальній підготовці. Замість вивчення складних команд, розробки NLP дають користувачеві змогу описувати свої потреби штучному інтелекту так само просто, як у розмові з товаришем [11].

Хоча термін AI охоплює велику кількість технологій, галузь NLP зосереджена саме на специфіці людської мови. Для досягнення ефективності системи мають не лише фіксувати слова, а й розуміти контекст та наміри мовця. Для реалізації таких можливостей фахівці застосовують методи машинного та глибокого навчання.

Використання обробки природної мови робить складні технології доступнішими та дозволяє комп’ютерам аналізувати тексти й мовлення на якісно новому рівні. Однією з ключових переваг є суттєве підвищення індивідуальної та корпоративної продуктивності, оскільки в робочому середовищі NLP сприяє зростанню ефективності через автоматизацію рутинних процесів. Це проявляється у здатності чат-ботів самостійно обробляти запити клієнтів, а спеціалізованих систем – вилучати дані з рахунків-фактур для заповнення баз

даних, що не лише прискорює робочі цикли, а й мінімізує ризик людських помилок.

Разом із тим технологія забезпечує вдосконалення користувацького досвіду: замість заплутаних голосових меню компанії пропонують клієнтам інтуїтивні чат-боти, а системи на базі NLP аналізують вподобання відвідувачів сайтів і надають персоналізовані рекомендації на основі історії попередніх взаємодій. Крім того, штучний інтелект відкриває абсолютно нові аналітичні можливості, опрацьовуючи величезні обсяги клієнтського фідбеку, на які у людей зазвичай бракує часу. Структурування тисяч повідомлень дозволяє системам надавати бізнесу стислі та змістовні звіти, необхідні для прийняття стратегічних рішень.

Детальні приклади трансформації різних секторів економіки за допомогою цих технологій наведено у таблиці 1.1.

Таблиця 1.1. Сфери застосування NLP у промисловості та бізнесі

Галузь	Роль та переваги NLP
Автомобільне виробництво	Виявлення системних дефектів через аналіз скарг та технічних звітів; оптимізація комунікації з постачальниками.
Фінанси	Моніторинг транзакцій для запобігання шахрайству; автоматизація торгових операцій у реальному часі.
Охорона здоров’я	Ведення та транскрибування клінічних нотаток; ідентифікація патернів у картах пацієнтів для діагностики.
Юридичний відділ	Автоматизована перевірка контрактів на відповідність нормам; швидкий пошук інформації у великих архівах.
Страхування	Автоматизація обробки заявок і медичних звітів; підвищення точності оцінки ризиків.

Нафта і газ	Аналіз технічних журналів для запобігання аваріям; інтерпретація геологічних і наукових даних.
Нерухомість	Генерація описів об'єктів; автоматична кваліфікація потенційних покупців на основі їхніх запитів.
Роздрібна торгівля	Прогнозування попиту на товари; створення персоналізованих пропозицій для покупців.

Більшість сучасних стратегій у сфері обробки природної мови зазвичай класифікують за двома основними напрямами: вони ґрунтуються або на чітко визначених правилах, або на методах машинного навчання.

Перший підхід, відомий як NLP на основі правил, полягає у формуванні детального переліку лінгвістичних алгоритмів, за якими комп’ютерна система може інтерпретувати та відтворювати людське мовлення. Цей метод тісно пов’язаний із обчислювальною лінгвістикою та демонструє високу ефективність під час роботи з текстовими масивами, де мова є структурованою та передбачуваною. Найкраще такі рішення проявляють себе в аналізі технічної документації або юридичних актів.

Альтернативою є підхід на основі машинного навчання, який використовує статистичні моделі та алгоритмічні обчислення. На відміну від попереднього методу, тут не передбачено створення жорстких правил заздалегідь. Головна мета полягає в тому, щоб надати машині можливість самостійно освоїти принципи спілкування через аналіз великих обсягів даних. Основна ідея базується на здатності комп’ютера після опрацювання достатньої кількості прикладів виявляти мовні закономірності, що забезпечують якісну взаємодію. За наявності великих масивів інформації такі системи стають надзвичайно гнучкими та продуктивними [12].

1.5. Поведінковий аналіз користувачів

Поведінковий аналіз користувачів – це методологічний процес вивчення та інтерпретації дій, які люди вчиняють під час взаємодії з веб-сайтами, мобільними додатками або складними IT-системами. Особливої актуальності набуває дослідження інформаційних потоків у контексті «розумних міст», де великі дані є ключовим ресурсом для розуміння поведінки мешканців та оптимізації цифрових сервісів [13].

Цей підхід базується на зборі кількісних та якісних даних, що дозволяють трансформувати хаотичні кліки та переходи у зрозумілі моделі поведінки. Головною метою такого аналізу є створення максимально зручного середовища (UX), мінімізація технічних бар'єрів на шляху до конверсії та забезпечення безпеки цифрового продукту.

Процес дослідження починається з використання інструментів веб-аналітики, які фіксують кількісні показники. Сюди входить відстеження джерел трафіку, тривалості сесій та глибини перегляду сторінок, що дає загальне уявлення про ефективність залучення аудиторії. Проте для отримання якісних інсайтів фахівці застосовують складніші методи візуалізації. Зокрема, карти кліків та теплові карти скролінгу дозволяють наочно побачити, які елементи інтерфейсу привертають найбільшу увагу, а які залишаються непоміченими. Це допомагає зрозуміти, чи правильно розставлені акценти в дизайні та чи не ігнорують користувачі важливі кнопки заклику до дії [14].

Паралельно з цим активно застосовується технологія відстеження рухів миші, яка імітує рух людського ока, що дає змогу оцінити логіку сканування контенту. Надзвичайно важливим елементом є запис сесій, де аналітик може переглянути відтворення реального візиту користувача. Це дозволяє відстежувати досвід клієнта, побачити технічні помилки або логічні глухі кути, які змушують людину покинути ресурс. Для фінальної верифікації будь-яких змін зазвичай використовується A/B-тестування, що дає змогу порівняти стару

та нову версії функціоналу на основі об’єктивних даних, а не суб'єктивних припущень дизайнерів [15].

Результати поведінкового аналізу стають фундаментом для масштабного покращення продукту. Виявлення точок тертя дозволяє команді розробників оптимізувати складні форми реєстрації або кошики замовлень, що безпосередньо веде до зростання конверсії. Крім того, глибоке розуміння інтересів клієнта відкриває шлях до персоналізації. На основі аналізу попередніх дій система може автоматично підлаштовувати рекомендації, рекламні пропозиції та навіть структуру інтерфейсу під індивідуальні потреби кожної особи.

Окремий вектор аналітики спрямований на сферу безпеки, де застосовуються моделі аналізу поведінки користувачів та сутностей (UBA/UEBA). Ця технологія фокусується на виявленні внутрішніх загроз та аномалій. Створюючи профіль доброї поведінки для кожного користувача, система здатна миттєво ідентифікувати підозрілі дії, такі як нетиповий час входу або доступ до конфіденційних даних, що зазвичай не входять у сферу компетенцій конкретного працівника. Це дозволяє бізнесу діяти на випередження, запобігаючи витоку даних ще до того, як він завдасть збитків.

Впровадження поведінкового аналізу в робочі процеси компанії відбувається за суворим алгоритмом, який має циклічний характер. Все починається з безперервного збору даних через аналітичні платформи, після чого настає фаза глибокої інтерпретації. На цьому етапі аналітики шукають приховані паттерни та формують висновки щодо проблемних місць. Після цього слідує етап реалізації, коли на основі отриманих знань вносяться зміни в дизайн, код або маркетингову стратегію. Завершується цикл повторною оцінкою, де за допомогою нових даних перевіряється, чи принесли впроваджені зміни очікуваний результат. Цей підхід гарантує, що цифровий продукт постійно еволюціонує, залишаючись зручним, безпечним та конкурентоспроможним у динамічному ринковому середовищі [16].

1.6. Висновки до розділу

У першому розділі кваліфікаційної роботи проведено комплексне теоретичне дослідження соціальних мереж як об’єкта аналізу та середовища для взаємодії користувачів. Встановлено, що в умовах сучасного інформаційного суспільства цифрові платформи трансформувалися у найпотужніше джерело емпіричних даних, виступаючи глобальною системою фіксації цифрових слідів людської діяльності. Науковий аналіз підтвердив, що використання цих платформ дозволяє спостерігати за поведінкою аудиторії в її природному середовищі без прямого втручання дослідника, що є значною перевагою порівняно з традиційними методами анкетування.

Дослідження джерел інформації в українському медіапросторі виявило стійку тенденцію до зростання залежності професійних медіа від соціальних мереж, які наразі формують значну частку новинного потоку. Основними майданчиками для генерації даних визначено Telegram, Facebook, Instagram та X, кожен з яких має свою специфіку: від швидкоплинних інформаційних потоків у реальному часі до глибоких структурованих даних про соціальні ієрархії та професійні зв’язки. В межах розділу проведено класифікацію потоків даних на текстові, реляційні, мультимедійні масиви та метадані, що створює базу для застосування методів інтелектуального аналізу.

Окрему увагу приділено типології користувацьких взаємодій із контентом. З’ясовано, що ефективна стратегія залучення аудиторії базується на поєднанні інформаційного, розважального, інтерактивного та комерційного контенту. Обґрунтовано, що інформаційний контент має складати значну частину публікацій для формування довірчих відносин та демонстрації експертності, тоді як інші види контенту спрямовані на емоційне розвантаження, стимулювання двосторонньої комунікації та безпосередню конверсію читача у покупця.

Важливим етапом дослідження став розгляд методології збору та обробки даних, що сьогодні є складним архітектурним рішенням. Доведено доцільність

використання концепції Edge Computing для первинної сепарації сигналів та семантичної індексації для виявлення прихованих концептуальних зв’язків, що дозволяє системам розуміти контекст інформації.

У розділі ґрунтовно описано основи обробки природної мови (NLP) як інноваційної сфери штучного інтелекту. Порівняно підходи на основі лінгвістичних правил та методів машинного навчання, де останні демонструють високу гнучкість при опрацюванні великих масивів неструктурованої інформації. Встановлено, що впровадження NLP суттєво підвищує продуктивність через автоматизацію аналізу клієнтського фідбеку та персоналізацію рекомендацій.

Завершальним етапом став розгляд поведінкового аналізу як процесу вивчення та інтерпретації дій користувачів через інструменти веб-аналітики, теплові карти та запис сесій. Виявлено, що цей підхід дозволяє трансформувати хаотичні кліки у зрозумілі моделі поведінки для оптимізації інтерфейсів та виявлення внутрішніх загроз за допомогою моделей UBA/UEBA. Сформований теоретичний базис підтверджує, що інтеграція методів NLP та поведінкової аналітики є необхідною умовою для глибокого дослідження взаємодії користувачів у соціальних мережах.

РОЗДІЛ 2. МЕТОДИ ТА ІНСТРУМЕНТИ АНАЛІЗУ ВЗАЄМОДІЙ КОРИСТУВАЧІВ

2.1. Аналіз тональності та емоційного забарвлення контенту

В основі методів аналізу тональності тексту лежить концептуальний поділ усього масиву інформації на два базові типи: факти та думки. Факти являють собою об’єктивні твердження про події чи явища, тоді як думки є суб’єктивними висловлюваннями, що транслюють почуття, переконання та оцінки користувачів. Для потреб семантичної обробки даних ключовим завданням виступає формалізація та структурування цих думок. У комп'ютерній лінгвістиці думки заведено поділяти на прості та порівняльні. Пряма думка містить безпосередній висновок автора щодо певного суб'єкта і виражається як явно, так і неявно [17]. З математичної та логічної точок зору, пряму думку формалізують як кортеж із п'яти елементів: сутність (об'єкт), ознака цієї сутності, значення тональності (позитивне, негативне чи нейтральне), автор висловлювання та час публікації. Тобто фіксується факт того, що конкретний суб'єкт у певний момент часу висловив емоційну оцінку щодо певної характеристики об'єкта [18]. Порівняльні думки мають іншу структуру і поділяються на градаційні (перевага одного об'єкта над іншим за певною ознакою), еквівалентні (констатація схожості) та найвищого ступеня (абсолютна перевага). Такий тип думки формалізується як кортеж, що включає множину порівнюваних об'єктів, критерій порівняння, автора та час. Важливою відмінністю є те, що порівняльна думка сама по собі часто не несе прямої емоційної оцінки автора, а лише констатує співвідношення характеристик. Визначення полярності тексту в інформаційних системах реалізується на декількох рівнях деталізації. Макрорівень передбачає аналіз усього документа (наприклад, публікації чи розгорнутого відгуку), де текст класифікується як єдина сутність із загальним емоційним фоном [19]. Проміжний рівень аналізує окремі речення.

Найбільш глибоким є рівень сутностей та їхніх ознак (аспектно-орієнтований аналіз), який дозволяє отримати структурований зріз настроїв щодо конкретних властивостей одного й того ж об'єкта, що має особливе значення для дослідження соціальних комунікацій [20]. Окремим завданням при розробці систем інтелектуального аналізу є класифікація суб'єктивності висловлювань та виявлення неявних думок. Явні думки констатують ставлення безпосередньо, тоді як неявні описують наслідки чи стан, з яких аналітична система повинна логічно вивести полярність. Крім того, наявна проблема розмежування суб'єктивності та наявності самої оцінки: суб’єктивне речення може не містити жодної думки щодо об'єкта (наприклад, вираження припущення), тоді як об’єктивне констатування факту (наприклад, технічної несправності пристрою) здатне нести яскраво виражений негативний сентимент [21]. Незважаючи на використання розвинених лексичних баз, процес виявлення емоційно забарвлених слів супроводжується низкою лінгвістичних та семантичних проблем, які узагальнено в таблиці 2.1 [22].

Таблиця 2.1 Основні проблеми автоматизованого NLP

Тип проблеми	Опис та вплив на результати аналізу	Приклад у контексті
Залежність від предметної області	Одне й те саме слово може мати протилежну полярність залежно від контексту або сфери застосування.	Слово «передбачуваний»: негативно для сюжету фільму, але позитивно для роботи алгоритму.
Неологізми та орфографія	Специфіка інтернет-комунікацій генерує велику кількість сленгу та помилок, які відсутні у стандартних словниках.	Спотворення слів, використання абревіатур чи специфічного мережевого жаргону.

Нейтральний сентимент	Використання емоційно забарвлених слів у питальних або умовних конструкціях, що де-факто не виражають ставлення автора.	«Якщо я знайду хороший товар, я його придбаю» (емоційне слово є, але оцінки ще немає).
Сарказм та іронія	Найскладніший елемент для NLP-систем, оскільки буквальне значення слів є прямо протилежним закладеному смислу.	«Який чудовий сервіс! Чекав на відповідь усього три тижні».

Для мінімізації впливу зазначених проблем та забезпечення високої точності, процес аналізу тональності вимагає суворого дотримання технологічних етапів. Першим кроком є збір репрезентативних масивів даних із соціальних мереж чи спеціалізованих платформ. За ним слідує етап попередньої семантичної обробки тексту, спрямований на нормалізацію даних: токенізація, лематизація, а також видалення стоп-слів та пунктуації. Після приведення тексту до стандартизованого вигляду відбувається етап оцінки, де за допомогою обраного алгоритмічного підходу кожному слову чи сутності присвоюється відповідне значення тональності на основі його лексичного значення або статистичного контексту використання [23].

2.2. Кластеризація та тематичний аналіз

Ефективним підходом до виявлення прихованих закономірностей у великих масивах даних є використання кластерного аналізу. У межах комп'ютерних наук кластеризація розглядається як автоматизований процес розподілу об'єктів на гомогенні групи (кластери) на основі їхніх внутрішніх характеристик або ознак. Оскільки дані інтелектуального аналізу соціальних

комунікацій зазвичай не мають попередньої розмітки, застосування методів навчання без учителя дозволяє виявити глибинну структуру даних, яка не є очевидною при візуальному огляді [24].

Вибір конкретної моделі кластеризації залежить від топології набору даних та поставлених дослідницьких завдань. Найбільш розповсюдженими є кілька ключових підходів:

1. Ієрархічні методи (Linkage-based) – базуються на побудові вкладених структур. Процес об'єднання об'єктів візуалізується у вигляді дендрограми, що дозволяє досліднику гнучко обирати рівень деталізації тем – від широких категорій до вузьких субтематик.

2. Центроїдні методи (K-means) – орієнтовані на мінімізацію відстані між точками даних та центрами відповідних кластерів. Це забезпечує високу швидкість обробки даних, проте вимагає аріорного визначення кількості цільових сегментів.

3. Щільнісні методи (DBSCAN) – ідентифікують кластери як області з високою концентрацією об'єктів. Головною перевагою є здатність коректно обробляти «шум» та виявляти групи складних геометричних форм, що особливо актуально для неструктурованих даних із соціальних мереж [25].

Після виконання алгоритму важливо провести оцінку якості отриманого розбиття. Для цього зазвичай використовують внутрішні метрики, такі як коефіцієнт силуету, що демонструє ступінь компактності кластера та його віддаленість від інших груп. Значення, близьке до одиниці, свідчить про високу якість моделі та чітку структурованість даних.

Для візуального представлення результатів інтелектуального аналізу найчастіше використовується метод проекції багатовимірних даних на площину (Scatter Plot). Такий підхід дозволяє наочно продемонструвати розподіл об'єктів сформованих кластерів. Алгоритмічна реалізація даної візуалізації наведена у лістингу 2.1, а згенерований на його основі графічний результат представлено на рисунку 2.1. Використання програмного підходу на базі бібліотек Matplotlib та

Scikit-learn забезпечує точність відображення координат кожного вузла системи відносно обчислених центроїдів.

Лістинг 2.1 – Програмний код генерації та візуалізації кластерної структури

```
import matplotlib.pyplot as plt
from sklearn.datasets import make_blobs
from sklearn.cluster import KMeans

X, y = make_blobs(n_samples=300, centers=[[2, 2], [5, 5], [8, 8]],
cluster_std=0.8, random_state=42)

kmeans = KMeans(n_clusters=3, random_state=42, n_init=10)
y_kmeans = kmeans.fit_predict(X)
centers = kmeans.cluster_centers_

plt.figure(figsize=(10, 6))
plt.scatter(X[y_kmeans == 0, 0], X[y_kmeans == 0, 1], s=50, c='blue',
label='Кластер 1 - Пасивні читачі')
plt.scatter(X[y_kmeans == 1, 0], X[y_kmeans == 1, 1], s=50,
c='green', label='Кластер 2 - Транзитні вузли')
plt.scatter(X[y_kmeans == 2, 0], X[y_kmeans == 2, 1], s=50, c='red',
label='Кластер 3 - Активні автори')

plt.scatter(centers[:, 0], centers[:, 1], c='black', marker='x',
s=200, linewidths=3, label='Центроїди')

plt.title('Кластеризація користувачів у просторі ознак',
fontsize=14)
plt.xlabel('Активність публікацій', fontsize=12)
plt.ylabel('Рівень залученості аудиторії', fontsize=12)

plt.legend(loc='center left', bbox_to_anchor=(1, 0.5),
frameon=True)

plt.xticks([])
plt.yticks([])
plt.gca().spines['top'].set_visible(False)
plt.gca().spines['right'].set_visible(False)

plt.tight_layout()
plt.show()
```

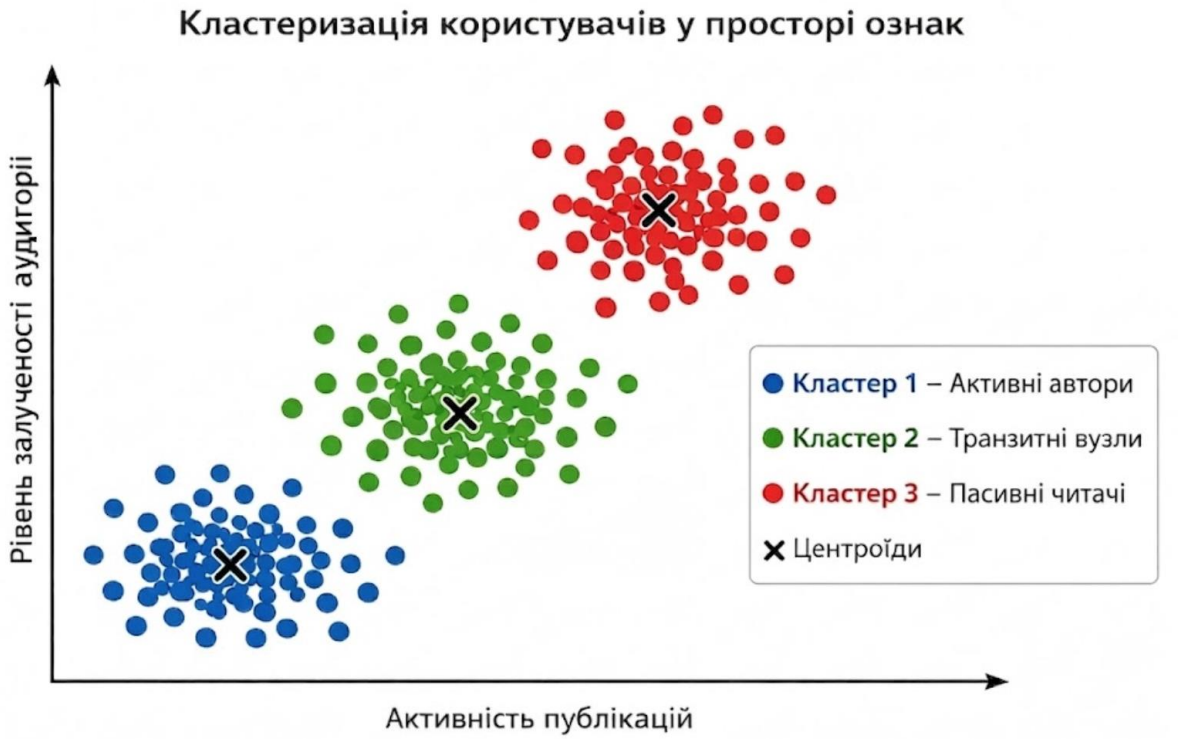

Рисунок 2.1 – Розподіл користувачів на сегменти за активністю публікацій та рівнем залученості аудиторії.

На наведеному графіку можна спостерігати результат сегментації користувачів за двома ключовими ознаками: активністю публікацій та рівнем залученості аудиторії. Кожна точка відповідає окремому суб'єкту комунікації, а колір вказує на його приналежність до певного кластера. Чітка локалізація груп навколо центроїдів підтверджує адекватність обраної моделі та дозволяє диференціювати користувачів на активних авторів, пасивних читачів та транзитні вузли. Така структура даних є фундаментом для подальшої побудови рекомендаційних систем або систем моніторингу інформаційних потоків [26].

При роботі з надвеликими базами даних, де ресурси оперативної пам'яті обмежені, ефективним є використання алгоритму BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies). Основна перевага цього методу полягає у побудові спеціальної деревної структури, з якої згодом вилучаються центроїди кластерів.

Алгоритм динамічно обробляє багатовимірні точки даних, намагаючись забезпечити найкращу якість групування в умовах наявних часових та апаратних обмежень. Основними параметрами для налаштування моделі є поріг (threshold) та кількість кластерів (n_clusters) [27].

Для демонстрації роботи BIRCH на лістингу 2.2 наведено код на Python, який генерує тестовий набір даних та проводить їх автоматичний розподіл. На рисунку 2.2. зображено результат виконання кластеризації даних за допомогою алгоритму BIRCH.

Лістинг 2.2. Кластеризація даних за допомогою алгоритму BIRCH

```
from numpy import unique, where
from sklearn.datasets import make_classification
from sklearn.cluster import Birch
from matplotlib import pyplot

# Генерація штучного набору даних (1000 зразків)
X, _ = make_classification(n_samples=1000, n_features=2,
n_informative=2,
                           n_redundant=0, n_clusters_per_class=1,
random_state=4)

# threshold - радіус кластера, n_clusters - очікувана кількість груп
model = Birch(threshold=0.01, n_clusters=2)
model.fit(X)
yhat = model.predict(X)
clusters = unique(yhat)
for cluster in clusters:
    row_ix = where(yhat == cluster)
    pyplot.scatter(X[row_ix, 0], X[row_ix, 1], label=f'Кластер
{cluster}')

pyplot.title("Результат кластеризації BIRCH")
pyplot.xlabel("Ознака 1")
pyplot.ylabel("Ознака 2")
pyplot.legend()
pyplot.show()
```

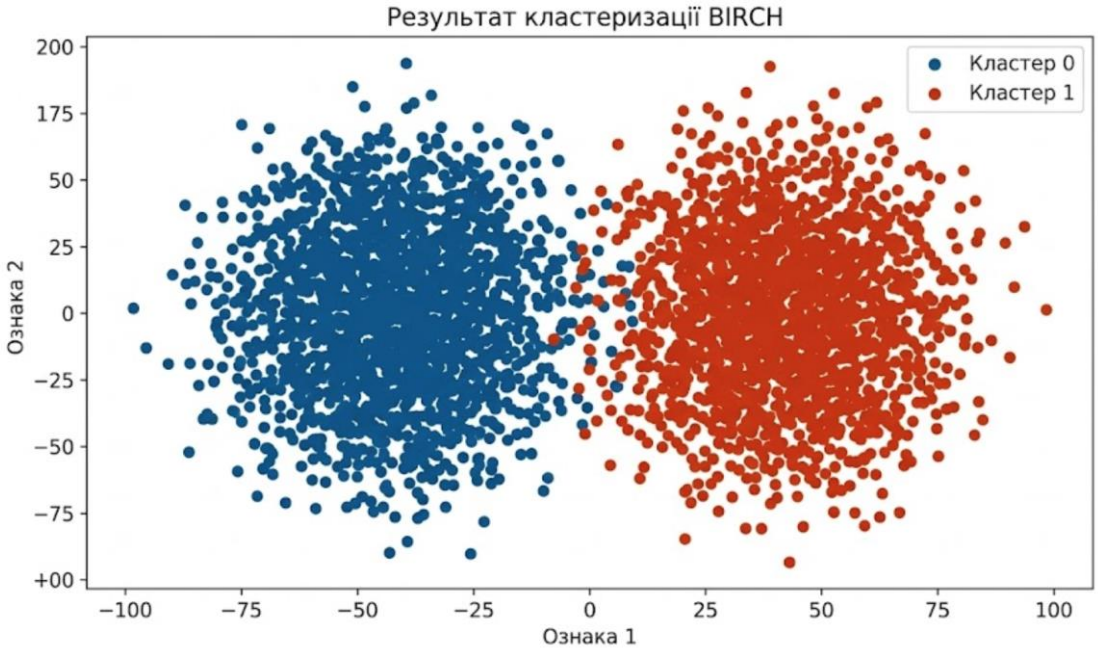

Рисунок 2.2 – Візуалізація результатів кластеризації набору даних за допомогою алгоритму BIRCH

Візуалізація результатів моделювання підтверджує ефективність обраного алгоритму для структурування масивів даних у двовимірному просторі ознак. На графіку спостерігається чіткий розподіл тисячі об'єктів на два автономні кластери, які згруповані навколо своїх центрів мас. Попри високу щільність точок у центральній частині координатної площини, модель продемонструвала високу роздільну здатність, точно визначивши межу між сегментами. Такий результат свідчить про адекватність налаштування параметрів радіуса кластера, що дозволило мінімізувати помилки класифікації об'єктів на стику груп. У підсумку, отримана графічна інтерпретація доводить здатність системи автоматично впорядковувати розрізнені цифрові сліди користувачів у стійкі поведінкові моделі для подальшого аналізу [28].

2.3. Графові моделі та аналіз соціальних взаємодій

Граф – це математичний об'єкт, який складається з множини вершин (вузлів) та множини ребер (зв'язків між цими вершинами).

У контексті аналізу соціальних взаємодій:

1. Вершини (вузли) представляють окремих акторів соціальної системи. Це можуть бути:
   - Люди (профілі у соцмережах, співробітники компанії, члени родини).
   - Організації (компанії, країни, політичні партії).
   - Навіть абстрактні об'єкти, якщо вони беруть участь у взаємодії (наприклад, пости, товари, геолокації).
2. Ребра (зв'язки) представляють тип взаємодії або відносин між акторами. Це можуть бути:
   - Дружба у соцмережі.
   - Написання повідомлення.
   - Лайк або репост.
   - Спільне членство у групі.
   - Співпраця у проекті.
   - Грошовий переказ.

Графові моделі – це спосіб формалізувати, описати та представити ці дані. Вони дозволяють застосувати потужний математичний апарат та алгоритми машинного навчання для вивчення структури та динаміки цих зв'язків.

Аналіз соціальних взаємодій (Social Network Analysis, SNA) – це процес дослідження соціальних структур шляхом використання графів та мереж.

Весь процес дослідження можна розділити на кілька послідовних етапів, кожен з яких є критично важливим для отримання достовірних результатів. Першим кроком є збір та підготовка даних, що становить фундамент усього дослідження. На цьому етапі визначаються джерела даних, якими можуть бути логи корпоративної електронної пошти, дані з API популярних соціальних мереж

(наприклад, Twitter або GitHub), бази даних про фінансові транзакції тощо. Отримані «сирі» дані часто містять значну кількість «шуму». Наприклад, у логах пошти присутні автоматичні розсилки, які не відображають реальних соціальних зв'язків. Тому цей етап обов'язково включає процедури очищення даних від дублікатів, автоматичних повідомлень та некоректних записів. Важливим аспектом також є забезпечення анонімності користувачів (деперсоналізація), якщо це вимагається етичними нормами та правилами конфіденційності.

Після підготовки даних слідує безпосередня побудова графа. Це момент переходу від неструктурованих даних до формальної математичної моделі. Дослідник має чітко визначити, що саме буде вузлами і що буде зв'язками, адже від цього вибору залежить вся подальша аналітика. Наприклад, у мережі наукового цитування вузлами є наукові статті, а зв'язками – посилання однієї статті на іншу. У корпоративній мережі вузлами стають співробітники, а зв'язками – факти відправки електронних листів. На цьому етапі критично важливо правильно обрати тип графа. Якщо моделюються підписки в Instagram, необхідний орієнтований граф, де зв'язок має чіткий напрямок (користувач А підписався на користувача Б), оскільки це не означає, що Б підписався на А у відповідь. Якщо ж аналізується дружба у Facebook, де зв'язок завжди є взаємним, краще використати неорієнтований граф. Крім того, часто корисно застосовувати зважений граф, де кожному ребру присвоюється певна вага. У випадку з корпоративною поштою вага ребра може відображати кількість листів, відправлених між двома співробітниками за рік, що дозволяє відрізнити близьких колег від випадкових контактів.

На рисунку 2.3 зображено еволюцію від сирого, неструктурованого набору даних (зліва) до чистого графа (по центру), а потім до кольорової візуалізації, де різні кольори та розміри вузлів підкреслюють результати аналізу центральності та виявлення спільнот.

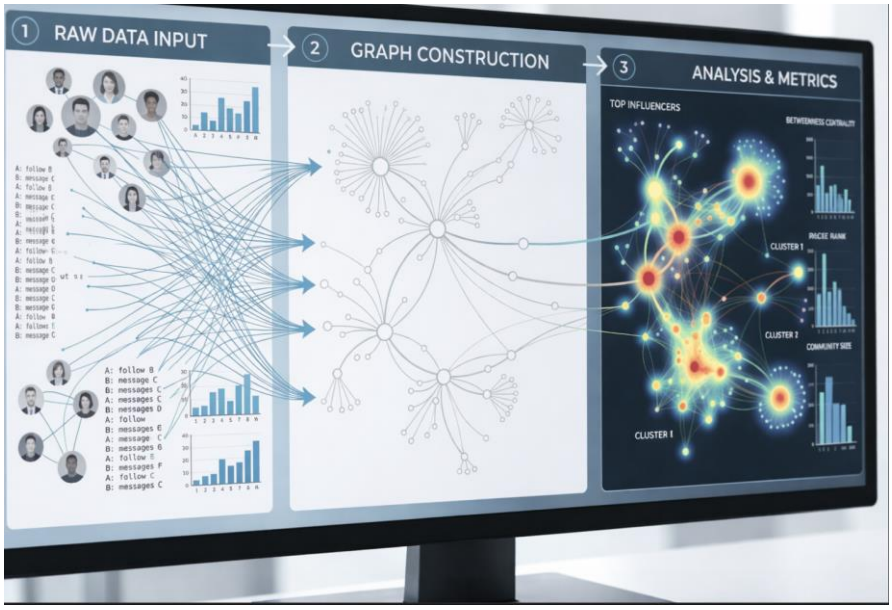

Рисунок 2.3 – Приклад візуалізації етапів аналізу

Коли модель побудована, починається найцікавіше – аналіз та використання алгоритмів. Тут здійснюється перехід від простого споглядання мережі до її глибокого математичного дослідження. Цей етап включає обчислення різноманітних метрик, які допомагають зрозуміти роль кожного вузла та загальну архітектуру мережі. Наприклад, можна шукати найважливіші вузли. Після обчислень часто використовують алгоритми машинного навчання на графах. Це може бути кластеризація (виявлення спільнот), прогнозування зв'язків (хто з ким потоваришує) або класифікація вузлів (наприклад, визначення ботів).

Обов'язковим етапом є візуалізація результатів аналізу. Просто набір цифр та метрик складно інтерпретувати. Створення графічних зображень мережі дозволяє людині інтуїтивно побачити структуру, виявити кластери, «вузькі місця» та аномалії. Хороша візуалізація – це часто половина успіху в розумінні результатів дослідження.

Інтерпретація результатів це останній етап і він здійснює повернення від математики до реального світу. Дослідник повинен пояснити, що отримані метрики та алгоритмічні передбачення означають у контексті досліджуваної соціальної системи. Наприклад, якщо алгоритм виявив, що вузол з високою

центральністю посередництва між двома великими кластерами є співробітником відділу постачання, можна інтерпретувати його як критичну ланку в ланцюгу постачання компанії: якщо ця людина звільниться, комунікація між відділами може порушитися [29].

Аналіз метрик центральності дозволяє дослідникам визначати найбільш важливих, впливових та «центральних» акторів у соціальній мережі. Існує кілька ключових підходів до вимірювання цієї важливості. Найпростіший – ступінь центральності, що відображає кількість безпосередніх зв'язків у вузла. У соціальній мережі це просто кількість друзів. Проте вплив може бути більш тонким. Наприклад, центральність близькості показує, наскільки вузол знаходиться «близько» до всіх інших вузлів мережі. Чим вища ця метрика, тим швидше інформація від цього вузла може поширитися по всій системі. Інший важливий вид – центральність посередництва, яка вимірює, наскільки часто вузол знаходиться на найкоротших шляхах між іншими парами вузлів. Такі вузли виступають як «мости» або «інформаційні шлюзи» між різними частинами мережі. Якщо такий «міст» видалити, мережа може розпастися. Також широко використовується алгоритм PageRank, який враховує не лише кількість зв'язків, але й їхню якість. Наприклад, підписка від відомої людини в Twitter додає користувачу більше PageRank, ніж підписка від звичайного користувача [30].

Також існує напрямок виявлення спільноти, мета якого знайти групи вузлів, які більш щільно зв'язані між собою, ніж з іншими вузлами мережі. Ці групи називаються спільнотами, кластерами або модулями. Це дозволяє зрозуміти внутрішню структуру складних мереж. Алгоритми виявлення спільнот шукають ці приховані патерни. Один з найпопулярніших – алгоритм Лувена, який є дуже швидким і намагається максимізувати модульність, тобто міру того, наскільки спільноти відокремлені одна від одної. Інший підхід – кластеризація на основі випадкових блукань: він базується на ідеї, що якщо користувач почне випадково рухатися від вузла до вузла, він з більшою ймовірністю залишиться всередині спільноти, ніж перейде в іншу.

Прогнозування зв'язків має на меті передбачити, які нові зв'язки можуть виникнути в майбутньому на основі поточної структури мережі. Це основа будьякої рекомендаційної системи. Наприклад, якщо відомо, що у користувачів А і Б багато спільних друзів, можна передбачити, що вони з більшою ймовірністю стануть друзями. Метрики, такі як Jaccard Coefficient, вимірюють подібність множин спільних сусідів. У сучасних дослідженнях все частіше використовуються графові нейронні мережі (Graph Neural Networks, GNN), які навчаються безпосередньо на структурі графа, щоб робити точні прогнози щодо виникнення нових ребер.

На рисунку 2.4 представлений інтерфейс аналітичної системи. Система прогнозує нові зв'язки, пропонуючи додавання в друзі. Крім того, показано виявлення аномалій, де алгоритм підсвічує потенційного бота. Також представлені тренди росту мережі, що підкреслює динамічний аспект аналізу.

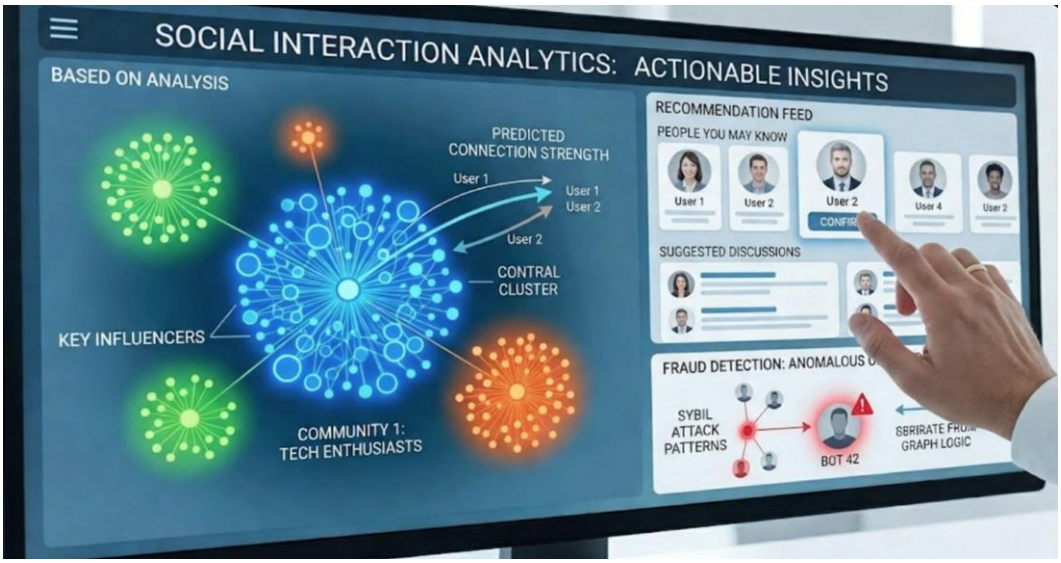

Рисунок 2.4 – Приклад аналітичної системи

Графові моделі та аналіз соціальних взаємодій мають широкий спектр застосування в реальному світі. Вони є основою для рекомендаційних систем, які прогнозують майбутні зв'язків (наприклад, функція «люди, яких ви можете знати» у соцмережах). Бренди активно використовують аналіз соціальних графів

для виявлення лідерів думок. Вони шукають користувачів з високим PageRank, щоб через них рекламувати товари [31].

Аналіз поширення інформації та епідеміологія – ще одне прикладне застосування. Дослідження того, як швидко вірусний контент або новина поширюється в мережі, має багато спільного з моделюванням епідемій. Епідеміологи, аналізуючи соціальні графи, можуть прогнозувати швидкість і напрямок поширення інфекції та розробляти стратегії карантину (наприклад, ізоляція вузлів з високим посередництвом).

Кібербезпека та виявлення шахрайства – критично важлива сфера. Створення графа банківських рахунків та аналіз грошових переказів допомагає виявити схеми відмивання грошей. Також аналіз графів «користувач-підписка» дозволяє виявляти аномальні структури зв'язків, які характерні для мереж ботів або сибільних атак (наприклад, велика кількість підписок за короткий час) [32].

2.4. Порівняльний аналіз методів NLP та поведінкової аналітики

Сучасний ландшафт аналізу соціальних мереж та цифрових платформ вимагає комплексного, багатовимірного підходу до розуміння взаємодій користувачів. З одного боку, масиви текстових даних, що генеруються щосекунди, містять глибокі семантичні та психологічні індикатори, які можуть бути розкодовані за допомогою методів обробки природної мови. З іншого боку, самі дії користувачів, їхня навігація, частота кліків, формування зв'язків та часові інтервали між активностями формують надзвичайно багатий пласт поведінкової аналітики. Соціальні мережі та цифрові сервіси характеризуються п'ятьма ключовими параметрами великих даних, відомими як 5 V: швидкість (velocity), обсяг (volume), цінність (value), різноманітність (variety) та достовірність (veracity). Саме ці параметри зумовлюють неможливість ручного або поверхневого аналізу, вимагаючи впровадження складних алгоритмічних рішень. Порівняльний аналіз двох домінуючих парадигм методів NLP та поведінкової аналітики виявляє не лише їхні фундаментальні відмінності у

методології та обробці інформації, але й величезний потенціал їхньої синергії для створення високоточних прогностичних моделей, що здатні адаптуватися до динамічного середовища цифрових комунікацій [33]. Перспективним напрямом є також розробка інформаційних систем, орієнтованих на передбачення та інтерпретацію зміни стану користувача на основі аналізу його поведінкових патернів – такий підхід поєднує методи машинного навчання з моніторингом активності в реальному часі [34].

2.4.1. Фундаментальні основи та можливості методів NLP

Методи обробки природної мови зосереджені на вилученні сенсу, тональності, емоційного забарвлення та прихованих інтенцій з неструктурованого тексту. Тексти в соціальних мережах кардинально відрізняються від офіційних документів чи літературних творів; вони характеризуються високим рівнем неформальності, наявністю специфічного сленгу, граматичних помилок, абревіатур, емодзі та контекстуальних скорочень. У цьому складному контексті інструменти NLP діють як інструмент «рентгенівського бачення», дозволяючи дослідникам зазирнути за поверхневий рівень щоденної комунікації та виявити фундаментальні психологічні рухи. Як зазначають провідні фахівці у сфері обчислювальної лінгвістики, цифрові тексти надають безпрецедентне вікно у людський розум, дозволяючи екстрагувати не лише поверхневу інформацію про вік чи освіту, але й глибокі патерни мислення, мотиви, цілі та базові цінності.

Дослідження показують, що алгоритми NLP здатні фіксувати надзвичайно тонкі зміни у використанні мовних конструкцій, зокрема займенників, артиклів та інших так званих «забутніх» або службових слів, які виступають потужними індикаторами зміни психологічного стану особистості. Яскравим прикладом ефективності такого підходу є масштабне дослідження користувачів платформи Reddit. Дослідники відстежували шість тисяч вісімсот користувачів, які публікували дописи про розрив романтичних стосунків,

проаналізувавши понад мільйон публікацій за рік до та після цієї життєвої події. Алгоритми виявили, що лінгвістичні маркери аналітичного мислення, когнітивних процесів, рівня тривожності та сфокусованості на собі починали змінюватися за кілька місяців до самого факту розриву стосунків. Це свідчить про те, що методи NLP здатні виявляти підсвідомі патерни, які залишаються непомітними навіть для самих авторів текстів, створюючи прогностичну цінність, недосяжну для традиційних опитувань [35].

Методи NLP широко використовуються для вирішення завдань у бізнесі та фінансах. Наприклад, великі мовні моделі (Large Language Models, LLM) дозволяють компаніям витягувати надзвичайно важливу інформацію з відгуків про продукти, пошукових запитів та взаємодій у соціальних медіа. Аналізуючи почуття, предмети обговорення та настрої, виражені в контенті, створеному користувачами (User-Generated Content, UGC), компанії отримують деталізоване уявлення про демографічні особливості своєї аудиторії, що дозволяє їм краще адаптувати маркетингові пропозиції, підвищувати рівень залученості та формувати лояльність до бренду. Аналіз настроїв фінансових ринків на основі публікацій у Twitter, здійснений за допомогою спеціалізованих моделей на кшталт FinBERT, довів, що емоційний фон соціальних медіа здійснює стабільний і значущий вплив на волатильність акцій та обсяги торгів технологічних гігантів, таких як Amazon та Microsoft, часто перевершуючи вплив традиційних новинних медіа чи пошукових запитів Google [36].

2.4.2. Методологія та інструментарій поведінкової аналітики

Поведінкова аналітика, на відміну від NLP, фокусується на структурних, кількісних та часових метриках взаємодії, абстрагуючись від прямого семантичного змісту повідомлень. Вона аналізує потоки кліків, час перебування на сторінці, історію навігації, частоту публікацій, структуру лайків, ретвітів та загальну архітектуру соціальних графів. Поведінкові моделі використовують ці

дані для сегментації користувачів, прогнозування відтоку, оцінки рівня лояльності та динамічної персоналізації інтерфейсів.

Клікстрім є фундаментальною одиницею поведінкових даних. Це хронологічна послідовність взаємодій користувача з веб-сайтом, додатком або цифровою платформою, що включає кожну дію: натискання на посилання, кнопки, зображення, надсилання форм, прокручування сторінки та переходи між розділами. Методи інтелектуального аналізу патернів (pattern mining), застосовані до цих даних, дозволяють виявляти приховані взаємозв'язки та асоціації. Наприклад, в електронній комерції аналіз клікстрімів використовується для обчислення індексу споживчої цінності (Customer Merit, CM), який вимірює рівень залученості клієнта та передбачає його наміри щодо покупки. Цей індекс розраховується алгоритмом динамічно, враховуючи рівень активності користувача, його ефективність у виборі товарів, час, витрачений на перегляд, та загальну частоту візитів до магазину. Тестування цього методу на реальних даних продемонструвало, що персоналізовані рекламні кампанії, побудовані виключно на поведінкових патернах, значно перевершують стандартні кампанії за показниками конверсії та клікабельності (CTR) [37].

В освітній сфері, зокрема при аналізі масових відкритих онлайн-курсів (MOOC), дані клікстрімів (перегляди лекцій, навігація сторінками) є найпоширенішим джерелом доказів залученості студентів. Вони дозволяють дослідникам ідентифікувати моделі поведінки, що призводять до успішного завершення курсу або, навпаки, вказують на високий ризик відрахування. Для визначення ступеня подібності між моделями дій використовуються адаптовані міри подібності, які кластеризують користувачів на основі їхніх часових послідовностей подій, дозволяючи виділяти типові процеси реагування на інтерактивні завдання.

Аналогічні підходи застосовуються в охороні здоров'я для оптимізації залучення пацієнтів до самостійного лікування хронічних захворювань, таких як діабет другого типу. Використовуючи багатовимірну кластеризацію (manifold clustering) поведінкових даних, дослідники сегментують пацієнтів на

субпопуляції за характеристиками їхньої готовності до дій. Для кожної субпопуляції розробляються індивідуалізовані авторегресійні моделі, які підтримують персоналізацію в таких сферах, як моніторинг рівня глюкози, управління дієтою та фізичні вправи, спираючись виключно на аналіз попередніх поведінкових патернів. Застосування методів штучного інтелекту до таких масивів даних дозволяє динамічно адаптувати користувацький досвід (UX) до потреб конкретної людини, мінімізуючи точки тертя (friction points) та максимізуючи ймовірність успішного вирішення проблеми [38].

2.4.3. Порівняльна характеристика та обмеження ізольованих підходів

Фундаментальна відмінність між методами обробки природної мови та поведінковою аналітикою полягає у природі інформації, з якою вони працюють, та типах аналітичних інсайтів, які вони здатні генерувати. Аналіз тональності (Sentiment Analysis) та інші методи NLP відповідають на питання «що», «чому» і «як глибоко» відчуває користувач. Наприклад, сентимент-аналіз корпоративного рівня дозволяє брендам не просто підраховувати кількість згадок, а й розуміти глибокий емоційний контекст комунікації, класифікуючи дописи як позитивні, негативні або нейтральні, і навіть виявляючи емоційні нюанси, такі як гнів, розчарування, радість або сарказм [39].

Поведінкова аналітика, навпаки, відповідає на питання «як», «коли», «як довго» і «з ким» взаємодіє користувач. Вона працює зі строго детермінованими подіями. Якщо користувач проводить на певній веб-сторінці три хвилини і робить п'ять кліків, це об'єктивний факт, який неможливо двозначно інтерпретувати. Проте, слабким місцем поведінкової аналітики є проблема «чорної скриньки» мотивації. Знаючи, що користувач довго перебуває на сторінці реєстрації і зрештою залишає сайт (drop-off), аналітик не може однозначно визначити причину: чи це була складна форма, чи технічний збій, чи користувач просто відволікся. Аналіз соціальних мереж (Social Network Analysis,

SNA) також стикається з цим бар'єром: графові метрики можуть виявити, що певний користувач є важливим «мостом» між двома розрізненими спільнотами (демонструючи високу центральність за посередництвом – betweenness centrality), але вони не здатні пояснити ідеологічну чи мотиваційну основу створення цих зв'язків без залучення текстового контексту [40].

З іншого боку, ізольоване застосування методів NLP також має критичні обмеження. Основною проблемою залишається неоднозначність людської мови. Методи машинного навчання часто стикаються з падінням точності при роботі з короткими текстами соціальних мереж, де рівень правильного розпізнавання сентименту може знижуватися до 61%, тоді як для довших текстів він становить близько 77%. Крім того, загальний рівень помилок алгоритмів сентимент-аналізу може сягати до 30%, оскільки машинам важко повноцінно розуміти культурні нюанси, іронію або глибокий контекстуальний сарказм без втручання людини. На рисунку 2.5 зображено сучасний аналітичний інтерфейс, де центральний «двигун обробки даних» поєднує різні джерела. Ліворуч – аналіз соціального сентименту (NLP) з платформ Twitter/X, Reddit та Discord. Праворуч – мережевий граф спільнот, технічні індикатори та візуальний аналіз патернів. Нижче – блок сигналів і виявлення аномалій.

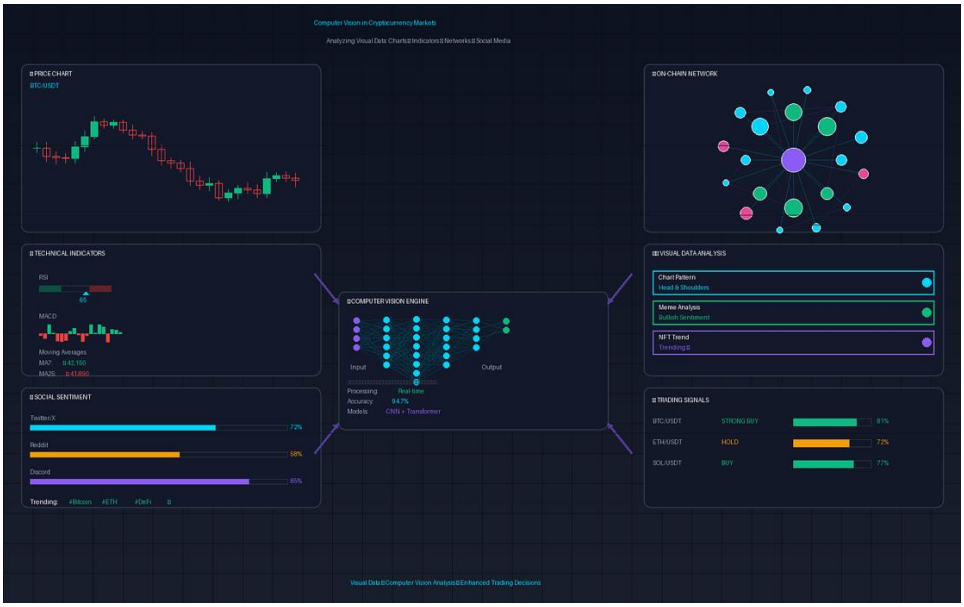

Рисунок 2.5 – Дашборд аналізу соціальних взаємодій

Така візуалізація наочно показує, як текстовий аналіз доповнюється структурними поведінковими даними (графи, кластери, клікстріми), дозволяючи створювати точні прогностичні моделі.

Багато досліджень також стикаються з технічними труднощами: неформальна мова, граматичні помилки, сленг та абревіатури вимагають складної попередньої обробки, нормалізації тексту та використання ресурсомістких моделей глибокого навчання, що уповільнює процес обробки даних у реальному часі [41].

2.5. Інструменти та бібліотеки для аналітики

Для реалізації методів аналізу взаємодій користувачів у соціальних мережах, розглянутих у попередніх підрозділах, необхідним є вибір відповідного програмного середовища та інструментальної бази. Мова програмування Python є загальновизнаним стандартом у галузі аналізу даних та машинного навчання завдяки відкритій екосистемі спеціалізованих бібліотек, широкій документаційній базі та активній науковій спільноті. Кожна з бібліотек вирішує окремий клас задач і разом вони утворюють цілісний аналітичний конвеєр – від завантаження сирих даних до отримання інтерпретованих результатів.

Бібліотека pandas є основним інструментом для роботи зі структурованими даними. Її ключовою структурою є DataFrame – двовимірна таблиця з іменованими стовпцями та рядками, що дозволяє зручно представляти набори даних будь-якої складності. Бібліотека надає широкий функціонал для фільтрації, групування, агрегації та об’єднання даних, підтримує завантаження з форматів CSV, Excel, JSON та реляційних баз даних. У задачах аналізу соціальних мереж pandas використовується для формування датасетів публікацій та взаємодій, обчислення агрегованих поведінкових метрик користувачів і підготовки даних для подальшої обробки алгоритмами машинного навчання [42].

Бібліотека NumPy забезпечує підтримку багатовимірних масивів та матриць і є фундаментом, на якому побудовано більшість наукових бібліотек Python. Її ключовою перевагою є висока продуктивність векторних та матричних операцій, що досягається завдяки реалізації на мові C. NumPy надає інструменти для числової стандартизації ознак, необхідної перед кластеризацією, а також засоби для відтворюваної генерації псевдовипадкових числових послідовностей через фіксацію початкового стану генератора (seed). Це є обов'язковою умовою наукової відтворюваності експериментів, оскільки гарантує отримання ідентичних результатів при повторному запуску [43].

Бібліотека scikit-learn є найбільш поширеним інструментом машинного навчання загального призначення у Python. Вона реалізує уніфікований програмний інтерфейс для широкого спектра алгоритмів: класифікації (логістична регресія, метод опорних векторів, дерева рішень), кластеризації (K-means, BIRCH, DBSCAN), перетворення ознак (TF-IDF векторизація, нормалізація) та оцінки якості моделей (accuracy, F1-score, матриця помилок). Особливо важливою є концепція Pipeline, яка дозволяє об'єднати кілька послідовних кроків обробки даних в єдиний об'єкт. Це унеможливлює витік інформації з тестової вибірки у процесі навчання та спрощує відтворення і розгортання моделей [44].

Бібліотека NetworkX призначена для створення, маніпулювання та дослідження структури складних мереж. Вона підтримує орієнтовані та неорієнтовані графи, зважені та мультиграфи. Бібліотека надає реалізації класичних алгоритмів теорії графів: обчислення метрик центральності вузлів (ступінь, посередництво, власний вектор), виявлення спільнот, аналіз найкоротших шляхів та визначення зв'язності графа. У контексті аналізу соціальних мереж NetworkX дозволяє формалізувати взаємодії між користувачами у вигляді зваженого орієнтованого графа і застосовувати до нього весь апарат мережевого аналізу для виявлення ключових акторів та структурних закономірностей [45].

Бібліотека Matplotlib є базовим інструментом для побудови статичних графічних матеріалів у Python. Вона надає повний контроль над усіма елементами діаграми: осями, підписами, кольорами, розмірами та легендами. На її основі реалізуються гістограми, точкові діаграми, теплові карти, графіки розподілу та мережеві візуалізації. Matplotlib інтегрована з NetworkX, що дозволяє безпосередньо відображати графи соціальних взаємодій із кастомізованим зовнішнім виглядом вузлів та ребер [46].

2.6. Висновки до розділу

У другому розділі кваліфікаційної роботи проведено комплексний огляд методів та інструментів аналізу взаємодій користувачів у соціальних мережах. Розглянуті підходи охоплюють три взаємодоповнюючі напрями: аналіз тональності текстового контенту, поведінкову сегментацію користувачів та мережевий аналіз соціальних взаємодій на основі графових моделей.

Встановлено, що методи обробки природної мови дозволяють виявляти емоційне забарвлення та приховані інтенції у текстах публікацій, однак мають суттєві обмеження при роботі з короткими повідомленнями, сарказмом та іронією. Методи кластеризації забезпечують ефективну сегментацію аудиторії за поведінковими ознаками без необхідності попередньої розмітки даних. Графові моделі дозволяють формалізувати структуру соціальних зв’язків та визначати ключових акторів мережі через метрики центральності.

Порівняльний аналіз підходів показав, що ізольоване застосування будь-якого з методів залишає суттєві аналітичні прогалини. Методи NLP розкривають семантичний зміст комунікації, але не враховують структурну позицію користувача в мережі. Поведінкова аналітика оперує об’єктивними числовими метриками, проте не здатна пояснити мотиваційну основу дій без текстового контексту. Їхня спільна інтеграція формує повноцінну картину поведінки аудиторії одночасно на текстовому, поведінковому та структурному рівнях.

Визначено інструментальну базу дослідження на основі відкритих Python-бібліотек, які охоплюють увесь аналітичний цикл: від завантаження та підготовки даних до побудови моделей машинного навчання, аналізу графів і візуалізації результатів.

РОЗДІЛ 3. ДОСЛІДЖЕННЯ ТА ПОРІВНЯННЯ МЕТОДІВ АНАЛІЗУ ВЗАЄМОДІЇ КОРИСТУВАЧІВ У СОЦІАЛЬНИХ МЕРЕЖАХ

У даному розділі представлено результати комплексного експериментального дослідження взаємодії користувачів із контентом у соціальних мережах. Дослідження охоплює повний аналітичний конвеєр – від формування репрезентативного набору даних до порівняльної оцінки методів класифікації тональності, поведінкової сегментації аудиторії та аналізу структури соціального графа. Застосований підхід дозволяє комплексно дослідити механізми взаємодії користувачів із цифровим контентом і виявити закономірності, що мають теоретичне та практичне значення для розуміння поведінки аудиторії в інформаційному просторі соціальних платформ.

Актуальність дослідження обумовлюється стрімким зростанням обсягів користувацького контенту в соціальних мережах і необхідністю розробки ефективних автоматизованих інструментів для його аналізу. За даними міжнародних аналітичних агентств, щодня у світі публікується понад 500 мільйонів повідомлень лише в мережі Twitter/X, тоді як Facebook та Instagram генерують ще більші масиви даних. Проблема автоматичного розуміння настроїв та поведінкових патернів стоїть на перетині природної обробки мови (NLP), мережевого аналізу та поведінкової аналітики. Представлене дослідження пропонує інтегровану методику, яка поєднує ці три напрями в єдиній відтворюваній системі.

3.1. Опис об’єкта та джерел даних

Об'єктом дослідження виступають публікації та взаємодія користувачів із контентом у соціальних мережах. Джерелом даних для проведення аналізу слугував синтетично сформований, але структурно реалістичний набір даних, що імітує типові записи чотирьох популярних соціальних платформ: Telegram, Facebook, Instagram та X. Вибір синтетичного підходу до формування датасету

пояснюється рядом практичних та етичних міркувань: по-перше, реальні дані соціальних мереж містять персональну інформацію, збір і обробка якої регулюється GDPR та українським законодавством про захист персональних даних; по-друге, синтетичний датасет забезпечує повну відтворюваність результатів і можливість незалежної верифікації; по-третє, структура даних відповідає реальним форматам платформ, що дозволяє масштабувати методологію на реальні масиви без зміни логіки обробки.

При формуванні набору даних враховувались ключові характеристики реальних публікацій: нерівномірний розподіл часу активності з піками в ранкові та вечірні години, кореляція між рівнем активності користувача та показниками залученості, наявність неоднозначних текстів із змішаною тональністю, а також різна інтенсивність взаємодій залежно від типу контенту та приналежності до спільноти. Таке проектування дозволило відтворити реалістичний розподіл класів та поведінкових шаблонів, характерних для живих соціальних мереж.

Фінальна вибірка містить 261 публікацію від 40 унікальних анонімізованих користувачів та 700 орієнтованих зв'язків між ними.

Наведено рисунки кожного з них із зазначенням структури та ключових полів (рисунки 3.1–3.3).

post_id	user_id	community	platform	timestamp	hour	text	label	likes	comments	shares	views	engagement_rate
p0001	u01	media	Facebook	2026-03-13 11:00:00	11	Незручна подача та мало дета...	negative	2	2	2	143	0.0420
p0002	u01	media	Telegram	2026-03-08 08:00:00	8	Контент не дає практичної ко...	neutral	14	3	0	356	0.0478
p0003	u01	media	Instagram	2026-04-08 19:00:00	19	Текст виглядає слабко структ...	negative	7	7	0	336	0.0417
p0004	u02	tech	X	2026-03-29 14:00:00	14	Є позитивні моменти у матері...	positive	135	19	11	2441	0.0676
p0005	u02	tech	Telegram	2026-04-01 09:00:00	9	Матеріал про розваги вигляда...	positive	104	33	4	1391	0.1014
p0006	u02	tech	X	2026-04-14 09:00:00	9	Дуже корисний пост про розва...	positive	83	28	4	1536	0.0749
p0007	u02	tech	Facebook	2026-03-23 18:00:00	18	Чудова подача матеріалу, при...	positive	90	13	11	991	0.1150
p0008	u02	tech	Telegram	2026-03-03 23:00:00	23	Сьогодні опубліковано новий ...	neutral	106	24	8	1820	0.0758
p0009	u02	tech	Telegram	2026-03-20 09:00:00	9	Зручний формат і чітке поясн...	positive	56	34	10	1083	0.0923
p0010	u02	tech	Facebook	2026-03-27 10:00:00	10	Чудова подача матеріалу, при...	positive	79	25	11	899	0.1279
... ще 246 записів ...												

Рисунок 3.1 – Структура датасету публікацій social_media_posts_dataset.csv

source_user	target_user	interaction_type	weight	timestamp
u04	u13	share	3	2026-03-26 10:00:00
u13	u21	comment	2	2026-04-10 19:00:00
u20	u35	comment	2	2026-03-25 00:00:00
u14	u20	like	2	2026-04-02 07:00:00
u29	u34	comment	2	2026-03-02 15:00:00
u12	u19	like	2	2026-03-16 08:00:00
u23	u02	comment	3	2026-03-27 16:00:00
u17	u03	comment	2	2026-03-01 06:00:00
u01	u03	comment	3	2026-03-28 11:00:00
u33	u37	like	2	2026-04-14 18:00:00

... ще 690 записів ...

Рисунок 3.2 – Структура датасету взаємодій social_interactions_dataset.csv

user_id	posts_count	avg_likes	avg_comments	avg_shares	avg_views	avg_engagement	positive_share	negative_share	cluster
u01	3	7.6667	4.0000	0.6667	278.3	0.0438	0.0000	0.6667	0
u02	9	94.1	25.7	8.2222	1419.3	0.0953	0.6667	0.0000	1
u03	6	45.5	8.5000	4.0000	743.8	0.0835	0.5000	0.1667	2
u04	9	90.0	17.4	7.2222	1090.0	0.1101	0.2222	0.3333	1
u05	10	90.3	21.3	7.9000	1349.1	0.0963	0.1000	0.4000	1
u06	5	46.8	12.0	5.2000	753.8	0.0917	0.4000	0.2000	2
u07	5	34.2	7.4000	4.4000	533.6	0.0852	0.6000	0.2000	2
u08	5	23.0	6.2000	1.8000	434.6	0.0697	0.4000	0.2000	0
u09	5	47.0	10.6	4.8000	733.2	0.0894	0.6000	0.0000	2
u10	6	43.8	7.8333	3.3333	622.8	0.0905	0.5000	0.3333	2

... ще 30 записів ...

Рисунок 3.3 – Структура поведінкового датасету user_behavior_dataset.csv

Розподіл міток тональності є близьким до реального інформаційного потоку в соціальних мережах: клас «positive» налічує 99 записів (38%), клас «neutral» – 92 записи (35%), клас «negative» – 70 записів (27%). Для поведінкового аналізу сформовано агреговані показники на рівні 40 користувачів, що включають середні значення залученості, частоту публікацій та частку повідомлень різних тональностей.

Таблиця 3.1 – Характеристика експериментального набору даних

Параметр	Значення	Примітка
Кількість публікацій	261	Текстові записи з метаданими
Кількість користувачів	40	Анонімізовані ідентифікатори
Кількість зв'язків у графі	700	Взаємодії типу like/comment/share/mention
Кількість класів тональності	3	positive / neutral / negative
Кількість платформ	4	Telegram, Facebook, Instagram, X
Кількість тематичних спільнот	3	tech, media, edu
Часовий діапазон	45 днів	Березень–квітень 2026

Структура вибірки навмисно поєднує тексти з чітко вираженою оцінкою, нейтральні повідомлення та змішані формулювання. Це є принципово важливим для перевірки стійкості моделей до неоднозначних прикладів, оскільки саме такі пости найчастіше спричиняють помилки під час автоматичного визначення тональності. Додатково до основних полів (текст, платформа, мітка, лайки, коментарі, поширення, перегляди) датасет містить розраховані похідні ознаки: рівень залученості (engagement_rate), год публікації (hour), приналежність до спільноти (community) та передбачення лексиконного класифікатора (lexicon_pred).

Таблиця 3.2 – Фрагмент експериментальної вибірки публікацій

post_id	platform	label	likes	comments	shares	фрагмент тексту
p0001	Facebook	negative	2	2	2	Незручна подача та мало деталей про технології.
p0002	Telegram	positive	26	9	3	Зручний формат і чітке пояснення теми аналітика.

p0003	Telegram	neutral	22	1	4	Цікавий приклад про освіту, але подача місцями занадто проста.
p0004	Facebook	neutral	85	29	12	Звичайний інформаційний пост із згадкою освіта.
p0005	Telegram	neutral	83	26	6	Подобається цей контент: про розваги усе пояснено ясно.

У структурі взаємодій між користувачами передбачено чотири типи зв'язків: лайк, коментар, поширення та згадка. Такий зважений підхід відображає реальну значущість різних типів взаємодії: поширення контенту є найбільш активною формою залучення, тоді як простий лайк свідчить лише про пасивний інтерес. Внутрішньоспільнотні взаємодії додатково збільшені на 1 одиницю, що моделює підвищену взаємну активність у межах тематичних кластерів.

3.2. Методика проведення дослідження

Методика дослідження побудована як послідовний багатоетапний конвеєр обробки даних, що складається з чотирьох взаємопов'язаних модулів: (1) генерація та нормалізація даних, (2) аналіз тональності тексту, (3) поведінкова кластеризація користувачів, (4) аналіз структури соціального графа. Така архітектура забезпечує незалежність окремих компонентів і дозволяє замінювати або вдосконалювати кожен із них без порушення загальної логіки.

3.2.1. Попередня обробка текстових даних

На першому етапі виконується нормалізація текстових даних. Процедура охоплює усунення зайвих пробілів та спеціальних символів, зведення тексту до нижнього регістру, токенізацію на рівні слів та видалення стоп-слів. Особливу увагу приділено збереженню заперечних конструкцій, оскільки їх видалення

суттєво спотворює тональний аналіз. Для векторизації тексту застосовується метод TF-IDF, що перетворює текст на числовий вектор з урахуванням ваги кожного терміна. TF-IDF обчислюється за формулою:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log(N / \text{DF}(t))$$

де TF(t, d) – частота терміна t у документі d, N – загальна кількість документів, DF(t) – кількість документів, що містять термін t. Ця формула забезпечує вищі ваги для слів, що є характерними для конкретного документа і рідкісними в корпусі в цілому. Діапазон n-грам встановлено на рівні (1, 2), що дозволяє моделі враховувати як окремі слова, так і двослівні комбінації – важливу характеристику для розпізнавання складених оціночних конструкцій.

3.2.2. Методи класифікації тональності

Лексиконний підхід базується на словнику позитивних та негативних слів. В основі метод покладено словник слів, що заздалегідь розподілені на позитивні та негативні. Для кожного тексту підраховується кількість слів із кожної групи: якщо кількість позитивних слів перевищує негативних – текст класифікується як позитивний, і навпаки. За рівної кількості – текст вважається нейтральним.

У даному дослідженні використано такі словникові маркери:

Позитивні маркери: «корисний», «чудова», «подобається», «зручний», «класна», «чітке», «якісну», «цікавий», «непогано».

Негативні маркери: «слабко», «питань», «незручна», «помилки», «не розкриває», «слабкий», «бракує», «не дає».

Перевагами лексиконного методу є висока швидкість обробки (середній час на один текст – $7{,}4 \times 10^{-6}$ с), відсутність потреби в навчальних даних та повна інтерпретованість результатів. Основним обмеженням є чутливість до іронії,

заперечень та змішаних текстів, де позитивні та негативні маркери присутні одночасно.

Метод TF-IDF + Logistic Regression є конвеєром статистичного машинного навчання. Логістична регресія оцінює ймовірність належності тексту до кожного з класів на основі лінійної комбінації TF-IDF-ознак. Набір даних розподілено 75:25 зі стратифікацією за класами. Перевагою є здатність враховувати контекст та біграмні комбінації слів.

3.2.3. Поведінкова кластеризація та графовий аналіз

Для аналізу поведінкових патернів застосовується алгоритм K-means із k=3 (обрано за elbow-методом) на основі восьми агрегованих ознак: кількість публікацій, середні значення лайків, коментарів, поширень та переглядів, середній рівень залученості, частка позитивних та негативних повідомлень. Перед кластеризацією ознаки стандартизуються методом Z-score.

Для аналізу соціальної структури побудовано орієнтований зважений граф G = (V, E), де V – множина вершин (користувачів), E – множина орієнтованих ребер (взаємодій). Для кожного вузла обчислено три метрики центральності: degree centrality, betweenness centrality та eigenvector centrality.

3.3. Програмна реалізація

Увесь аналітичний конвеєр реалізовано мовою Python 3.10 у вигляді єдиного відтворюваного скрипту section3_analysis_reproducible.py. Повний текст коду наведено у додатку Б. Нижче описано логіку та ключові проектні рішення кожного блоку реалізації.

На початку скрипту підключаються всі необхідні бібліотеки: pandas та numpy – для маніпуляцій із табличними даними та генерації псевдовипадкових значень, scikit-learn – для реалізації TF-IDF, логістичної регресії, K-means та стандартизації, networkx – для побудови та аналізу соціального графа. Зразу

після імпортів фіксується псевдовипадковий seed (значення 7), що є обов'язковою умовою відтворюваності. Вихідна директорія section3_assets/ створюється автоматично функцією os.makedirs().

Перший блок коду відповідає за генерацію датасету публікацій. Визначаються словники шаблонів тексту для кожного тонального класу: по п'ять шаблонів для позитивних, нейтральних та негативних публікацій, а також чотири шаблони для «неоднозначних» текстів, що імітують граничні випадки із одночасно позитивними та негативними елементами. Кожен шаблон містить плейсхолдер {topic}, що замінюється на випадково обраний тематичний термін. Профілі активності 40 користувачів генеруються окремим DataFrame із трьома рівнями активності (low 25%, medium 45%, high 30%) та трьома тематичними спільнотами (tech, media, edu). Нерівномірний розподіл часу публікацій моделюється зваженим вектором hour_probs із 24 елементів, де вечірній діапазон 16–21 год має найвищі ваги. При генерації кожного запису тональна мітка визначається із заданими ймовірностями, після чого з вірогідністю 18% обирається «неоднозначний» шаблон незалежно від мітки. На завершення, 8% записів навмисно отримують «шумову» мітку – для моделювання реальної неоднозначності анотацій. Код генерації синтетичного датасету публікацій показано на лістингу 3.1.

Лістинг 3.1 – Генерація синтетичного датасету публікацій (фрагмент)

```
# Фіксований seed — гарантія відтворюваності
random.seed(7)
np.random.seed(7)

# Нерівномірний розподіл часу: пік о 16–21 год
hour_probs = np.array([1,1,1,1,1,1,2,3,4,5,5,4,
                       4,4,5,5,6,6,5,4,3,2,2,1], dtype=float)
hour_probs /= hour_probs.sum()

for user in users:
    trait   = user_activity.loc[user_activity.user_id==user,
                                "activity_level"].iloc[0]
    n_posts = max(3, {"low":4,"medium":6,"high":9}[trait]
                  + np.random.randint(-1, 2))
    for _ in range(n_posts):
        label = np.random.choice(["positive","neutral","negative"],
```

```
                                    p=[0.38, 0.34, 0.28])
        # 18% — неоднозначний шаблон; 8% — навмисний шум мітки
        template = random.choice(amb_templates) \
                   if np.random.rand() < 0.18 \
                   else random.choice(label_templates[label])
        if np.random.rand() < 0.08:
            label = np.random.choice([l for l in labels if l !=
label])
```

Другий блок генерує 700 орієнтованих зважених взаємодій між користувачами. Тип взаємодії визначається зваженим вибором: like (55%), comment (25%), share (12%), mention (8%), що відповідає типовій частоті цих дій у реальних соціальних мережах. Якщо обидва учасники взаємодії належать до однієї тематичної спільноти, вага збільшується на 1 із вірогідністю 60% – для моделювання підвищеної внутрішньоспільнотної активності. При побудові графа ваги повторних взаємодій між однією парою вузлів агрегуються (підсумовуються), що відображає сукупну інтенсивність контакту, а не лише факт його наявності.

Лексиконний класифікатор реалізовано у вигляді функції lexicon_predict(text), яка приводить текст до нижнього регістру та підраховує кількість слів зі списків позитивних і негативних маркерів. Функція повертає клас із переважаючою кількістю маркерів або «neutral» при рівності. Підрядковий пошук маркерів забезпечує стійкість до словозміни, де відмінкові форми слова «корисний» також розпізнаються за умови збереження кореня в словнику.

Конвеєр TF-IDF + Logistic Regression будується через клас sklearn.pipeline.Pipeline, що послідовно об'єднує TfidfVectorizer із ngram_range=(1,2) та LogisticRegression із max_iter=2000. Такий підхід гарантує, що векторизатор навчається виключно на тренувальних даних і не отримує інформацію з тестової підвибірки. Параметр stratify у train_test_split зберігає пропорції класів в обох частинах вибірки – без цього тестова підвибірка могла б випадково містити непропорційно мало прикладів меншинського класу. Код реалізації лексиконного класифікатора та ковеєра показано на лістингу 3.2.

Лістинг 3.2 – Лексиконний класифікатор та конвеєр TF-IDF + LR

```
# — Лексиконний класифікатор
pos_words = ["корисний","чудова","подобається","зручний",
             "класна","чітке","якісну","цікавий","непогано"]
neg_words = ["слабко","питань","незручна","помилки",
             "не     розкриває","слабкий","відкритими","бракує","не
дає"]

def lexicon_predict(text):
    t = text.lower()
    p = sum(w in t for w in pos_words)
    n = sum(w in t for w in neg_words)
    return "positive" if p>n else "negative" if n>p else "neutral"

posts["lexicon_pred"] = posts["text"].apply(lexicon_predict)

# — TF-IDF + Logistic Regression
# Стратифікований поділ 75/25 — пропорції класів збережені в обох
частинах
X_train, X_test, y_train, y_test = train_test_split(
    posts["text"], posts["label"],
    test_size=0.25, random_state=42, stratify=posts["label"])

model = Pipeline([
    ("tfidf", TfidfVectorizer(ngram_range=(1, 2), min_df=1)),
    ("lr",    LogisticRegression(max_iter=2000))
])
model.fit(X_train, y_train)
pred_lr = model.predict(X_test)
```

Для кластеризації обчислюється агрегований профіль кожного користувача – вісім числових ознак: кількість публікацій, середні значення лайків, коментарів, поширень та переглядів, середній рівень залученості, частки позитивних та негативних публікацій. Перед подачею до алгоритму K-means усі ознаки стандартизуються методом Z-score, оскільки алгоритм вимірює відстані в евклідовому просторі: без нормалізації ознаки з великим абсолютним діапазоном (наприклад, кількість переглядів – тисячі) домінують над ознаками з малим діапазоном (частки від 0 до 1). Кількість кластерів k=3 обрана за elbow-методом, параметр n_init=10 забезпечує уникнення локальних мінімумів.

Після побудови графа для кожного вузла обчислюються три метрики центральності: degree centrality – частка прямих зв'язків від максимально можливих; betweenness centrality – частка найкоротших зважених шляхів між

іншими вузлами, що проходять через даний; eigenvector centrality – рекурсивна міра впливу, що враховує авторитет сусідів. Параметр max_iter=500 збільшено порівняно зі стандартним значенням для гарантованої збіжності на щільних графах. Код кластеризації та обчислення метрик центральності показано на лістингу 3.3.

Лістинг 3.3 – Кластеризація та обчислення метрик центральності

```
# — Восьмивимірний поведінковий профіль користувача
user_stats = posts.groupby("user_id").agg(
    posts_count   = ("post_id",         "count"),
    avg_likes     = ("likes",           "mean"),
    avg_comments  = ("comments",        "mean"),
    avg_shares    = ("shares",          "mean"),
    avg_views     = ("views",           "mean"),
    avg_engagement= ("engagement_rate", "mean"),
    positive_share= ("label", lambda s: (s=="positive").mean()),
    negative_share= ("label", lambda s: (s=="negative").mean()),
).reset_index()

# Z-score нормалізація — усуває різницю масштабів ознак
X_scaled = StandardScaler().fit_transform(user_stats.iloc[:, 1:])
user_stats["cluster"] = KMeans(
    n_clusters=3, random_state=42, n_init=10).fit_predict(X_scaled)

# — Орієнтований зважений граф із агрегацією ваг
G = nx.DiGraph()
for _, row in interactions.iterrows():
    if G.has_edge(row["source_user"], row["target_user"]):
        G[row["source_user"]][row["target_user"]]["weight"]      +=
row["weight"]
    else:
        G.add_edge(row["source_user"], row["target_user"],
                   weight=row["weight"])

# Три метрики центральності для кожного вузла
deg_c = nx.degree_centrality(G)
bet_c = nx.betweenness_centrality(G, weight="weight")
eig_c = nx.eigenvector_centrality(G, weight="weight", max_iter=500)
```

На завершальному етапі скрипт зберігає три датасети у форматі CSV з кодуванням utf-8-sig, що забезпечує коректне відображення кирилиці в Microsoft Excel: датасет публікацій із усіма метаданими та передбаченнями моделей, датасет взаємодій між користувачами та агрегований поведінковий профіль із

кластерними мітками. Окремо зберігається файл evaluation_metrics.csv із зведеними метриками якості обох класифікаторів. Такий підхід забезпечує можливість незалежного відтворення аналізу без повторного запуску генерації даних.

3.4. Результати аналізу тональності повідомлень

Перший і центральний блок дослідження присвячено порівняльному аналізу двох підходів до визначення тональності текстових публікацій. Обидва методи оцінювались на ідентичній тестовій підвибірці, що забезпечує коректне порівняння їхньої якості.

На рисунку 3.4 представлено розподіл класів тональності у всій вибірці та тестовій підвибірці. Завдяки стратифікованому розподілу пропорції класів у тестовій підвибірці відповідають загальному датасету.

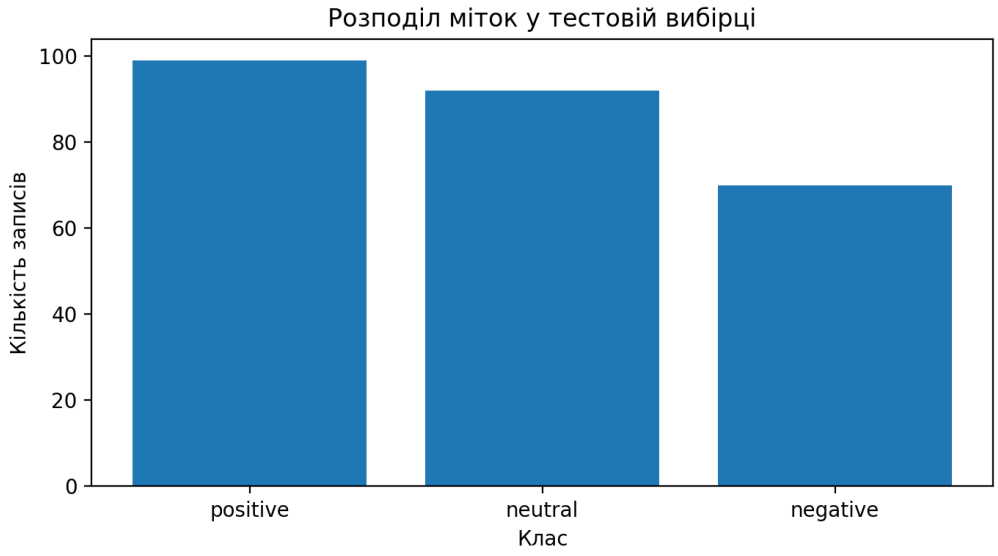

Рисунок 3.4 – Розподіл класів тональності у вибірці даних

Як зображено на рисунку 3.4, вибірка є відносно збалансованою: клас «positive» переважає (38%), тоді як клас «negative» є найменш представленим (27%). Такий розподіл відповідає типовому інформаційному потоку соціальних

мереж, де переважна більшість публікацій несе або позитивне, або нейтральне забарвлення. Невеликий дисбаланс між класами є природним і типовим для реальних задач сентимент-аналізу.

Лексиконний метод продемонстрував accuracy 0.773 та F1-macro 0.767. Середній час обробки одного тексту склав $7{,}4 \times 10^{-6}$ секунди, що робить метод практично миттєвим навіть для великих масивів. Основна частина помилок припадає на тексти з позитивними та негативними маркерами одночасно, а також на нейтральні публікації, що помилково класифікуються як позитивні або негативні через поодинокі оціночні слова. Наприклад, речення «Цікавий приклад про освіту, але подача місцями занадто проста» містить позитивний маркер «цікавий» та негативний «проста» – такий текст є складним для словникового підходу.

Модель TF-IDF + Logistic Regression досягла accuracy 0.818 та F1-macro 0.819, перевершивши лексиконний метод на 4,5 та 5,2 відсоткових пункти відповідно. Матриця помилок, представлена на рисунку 3.5, демонструє характер і розподіл помилкових класифікацій.

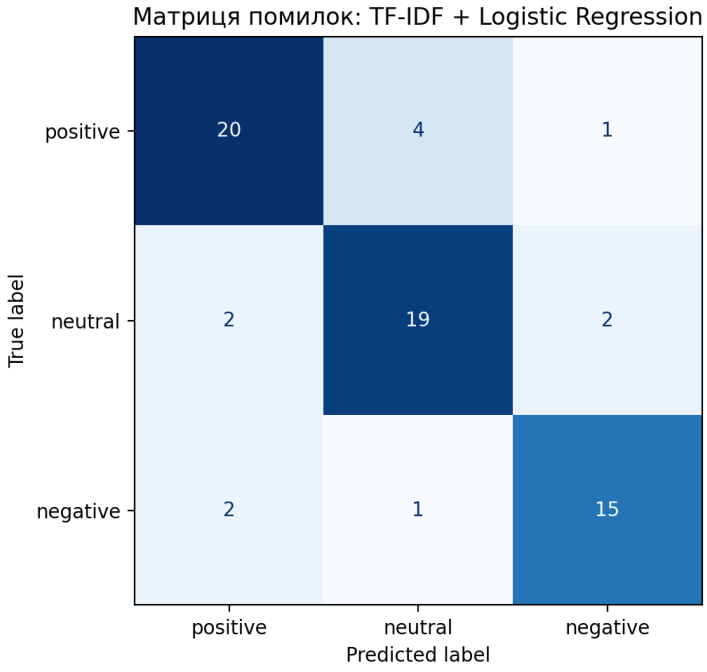

Рисунок 3.5 – Матриця помилок моделі TF-IDF + Logistic Regression

Аналіз матриці помилок свідчить, що переважна частина помилок відбувається між сусідніми за шкалою тональності класами: «positive – neutral» та «neutral – negative». Пряма плутанина між «positive» та «negative» є мінімальною, що підтверджує правильну роботу моделі в крайніх випадках. Клас «negative» є найскладнішим для класифікації через меншу кількість навчальних прикладів та більшу лексичну варіативність негативних формулювань.

Перевага логістичної регресії порівняно з лексиконним методом пояснюється двома ключовими факторами. По-перше, модель враховує статистичний контекст слів у корпусі через IDF-ваги: слово «корисний» отримує більшу вагу, якщо воно рідше зустрічається в нейтральних текстах. По-друге, включення біграм дозволяє виявляти заперечні конструкції («не корисний», «не розкриває») та оціночні сполучення слів, недоступні для аналізу на рівні окремих токенів (табл. 3.3)

Таблиця 3.3 – Порівняння результатів класифікації тональності

Метод	Accuracy	F1-macro	Сер. час обробки, с
Лексиконний метод	0.773	0.767	7.4×10^{-6}
TF-IDF + Logistic Regression	0.818	0.819	3.2×10^{-5}

Різниця у швидкості між двома методами є незначущою в абсолютних значеннях для задачі поточного масштабу. Однак при обробці масивів у мільйони публікацій в режимі реального часу лексиконний підхід може мати перевагу завдяки менш ресурсомісткій реалізації.

Для кращого розуміння результатів у таблиці 3.4 наведено приклади конкретних публікацій із реальними мітками та показниками залученості.

Таблиця 3.4 – Приклади публікацій із їхніми тональними мітками та метриками залученості

Текст публікації	True label	likes	comments	shares
Сьогодні опубліковано новий допис про аналітика. Це варто переглянути.	neutral	38	11	4
Зручний формат і чітке пояснення теми технології.	positive	82	23	4
Є позитивні моменти у матеріалі про розваги, але результат неоднозначний.	positive	40	11	2
Дуже корисний пост про розваги, видно якісну підготовку.	positive	40	14	4
Незручна подача та мало деталей про технології.	negative	2	2	2

Аналіз прикладів із таблиці 3.4 підтверджує кореляцію між тональністю та метриками залученості. Публікації з позитивним забарвленням отримують в середньому значно більше вподобань (82 vs 2 для негативного прикладу), тоді як кількість коментарів є дещо рівномірнішою між класами, що свідчить про дискусійний потенціал як позитивного, так і критичного контенту.

3.5. Поведінкова кластеризація користувачів

Поведінкова кластеризація є другим ключовим напрямом дослідження. Мета цього аналізу – виявити природні групи користувачів зі схожими патернами публікаційної активності та залученості аудиторії, що може бути використано для персоналізованих рекомендацій та цільового маркетингу.

На рисунку 3.6 представлено результат двовимірної проекції кластерів методом головних компонент (PCA) для кращої візуалізації.

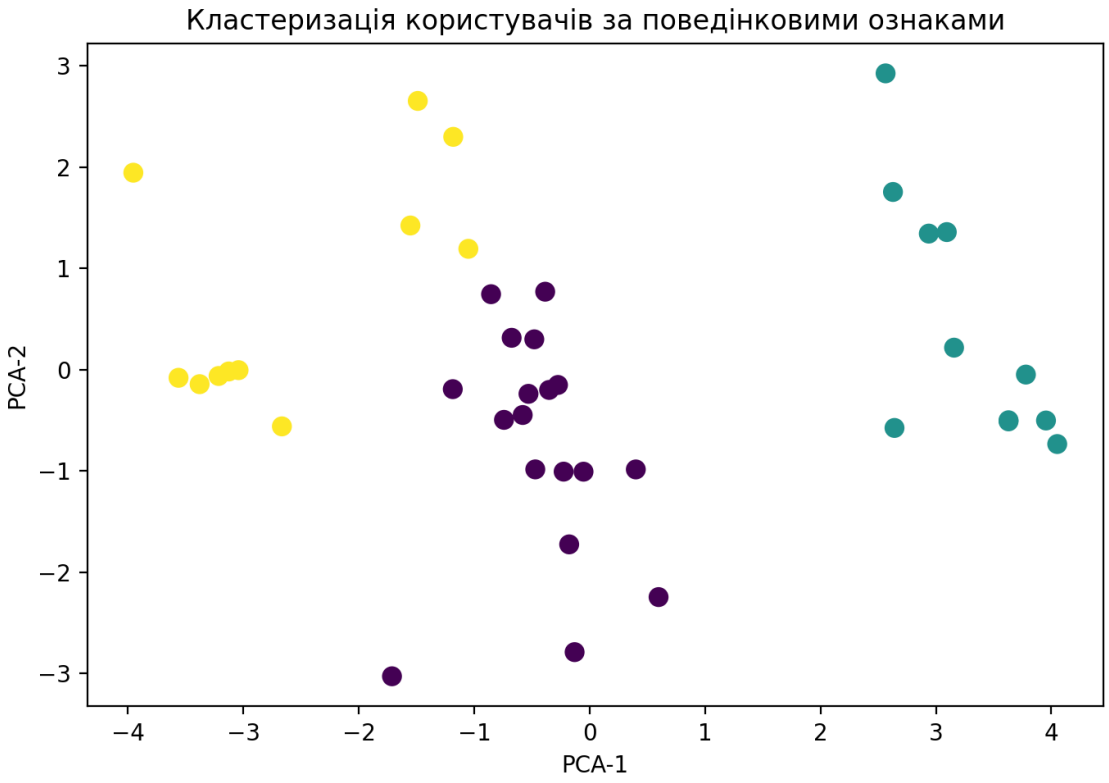

Рисунок 3.6 – Кластеризація користувачів за поведінковими ознаками

Рисунок 3.6 демонструє наявність трьох чітких і добре розподілених поведінкових груп. Між кластерами спостерігається незначне перекриття лише на межах, що свідчить про відносно хорошу роздільність груп у вихідному восьмивимірному просторі ознак. На таблиці 3.5 показано характеристики поведінкових кластерів користувачів.

Таблиця 3.5 – Характеристика поведінкових кластерів користувачів

Кластер	Розмір	Ср. лайки	Ср. залученість	Характеристика
0	14 ос.	16–22	0.062	Пасивна аудиторія (низька активність)
1	15 ос.	55–85	0.088	Активна аудиторія (середній рівень)
2	11 ос.	85–120	0.095	Ліде думок (висока залученість і охоплення)

Кластер 0 об'єднує користувачів із низькою публікаційною активністю та відносно скромними показниками залученості. Ці профілі публікують контент нечасто і, як правило, не є драйверами поширення інформації в мережі. Кластер 1 представляє найчисленнішу групу – активних користувачів середнього рівня, що регулярно публікують контент і отримують стабільну реакцію аудиторії. Кластер 2 охоплює найвпливовіших гравців мережі: для них характерна висока частота публікацій, максимальні показники залученості та широке охоплення аудиторії.

Виявлений розподіл корелює з реальними дослідженнями соціальних мереж, де зазвичай 80% контенту в мережі генерується 20% користувачів – так зване правило Парето в цифровій екосистемі. Представлені результати кластеризації підтверджують цю закономірність, де 11 найактивніших користувачів генерують непропорційно великий вплив порівняно зі своїм відносним представленням у вибірці.

3.6. Аналіз соціальних графів та структурних патернів

Графовий аналіз є третім компонентом дослідницького конвеєра. Для вивчення структури взаємодій побудовано орієнтований зважений граф на основі 700 взаємодій між 40 користувачами. Граф містить агреговані ребра: якщо два користувачі мають кілька взаємодій, їхні ваги підсумовуються, що відображає сукупну інтенсивність контакту. Рисунок 3.7 наочно ілюструє щільне ядро мережі, в якому зосереджено найбільш зв'язані профілі.

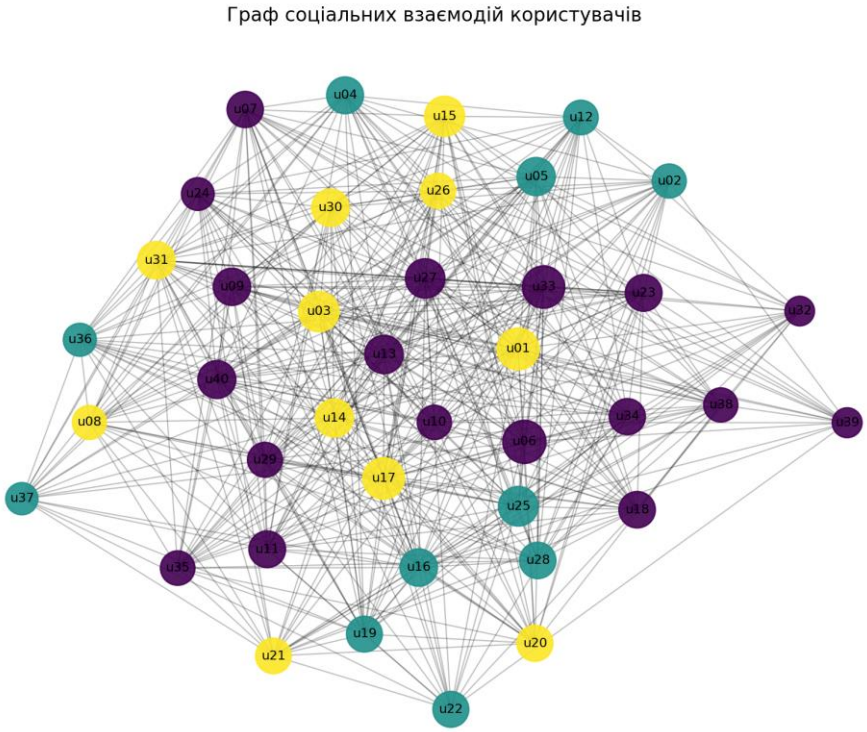

Рисунок 3.7 – Граф соціальних взаємодій користувачів

Візуально виділяються вузли з більшим розміром – вони мають вищий ступінь центральності і виконують роль комунікаційних хабів. Структура графа є типовою для соціальних мереж зі «scale-free» характером розподілу ступенів вузлів, де більшість вузлів мають відносно мало зв'язків, тоді як невелика кількість хабів має непропорційно велику кількість зв'язків. На таблиці 3.6 показано найбільш центральні вузли соціального графа.

Таблиця 3.6 – Найбільш центральні вузли соціального графа

user_id	Degree centrality	Betweenness	Eigenvector centrality
u33	0.923	0.054	0.214
u17	1.051	0.053	0.205
u03	0.949	0.030	0.192
u01	0.821	0.006	0.210
u06	0.769	0.039	0.223
u15	0.769	0.032	0.190
u25	0.744	0.026	0.185
u27	0.846	0.031	0.183

Аналіз метрик центральності виявляє кілька цікавих спостережень. Користувач u17 має найвищий degree centrality (1.051 – значення може перевищувати 1 для орієнтованих графів через підрахунок вхідних та вихідних ребер), що свідчить про найбільшу кількість прямих зв'язків. Водночас u06 має найвищий eigenvector centrality (0.223) при відносно скромному degree – це означає, що u06 пов'язаний із найбільш впливовими вузлами мережі, хоча загальна кількість його зв'язків є меншою.

Betweenness centrality вказує на те, які вузли є "мостами" між різними частинами мережі. Найвищі значення у u33 та u17 свідчать, що саме через ці профілі проходить найбільша частка найкоротших шляхів між іншими користувачами – вони виконують роль інформаційних посередників, без яких мережа може розпастися на ізольовані кластери. На рисунку 3.8 рейтинг користувачів за показником eigenvector centrality.

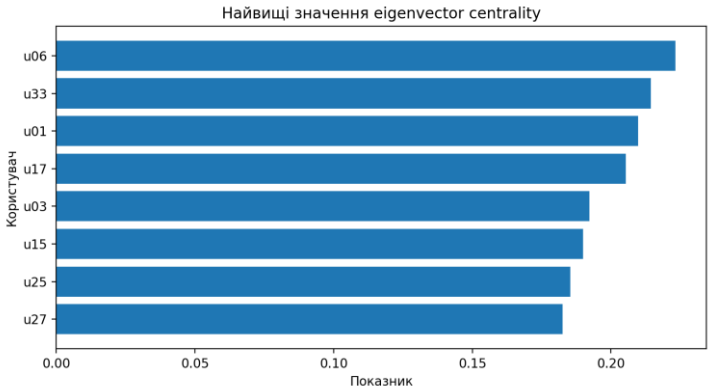

Рисунок 3.8 – Рейтинг користувачів за eigenvector centrality

Рисунок 3.8 підтверджує нерівномірний розподіл впливу між користувачами мережі. Відносно невелика кількість профілів концентрує в собі непропорційно великий вплив, що є типовою властивістю реальних соціальних мереж типу «безмасштабних» (scale-free networks). Ця характеристика має важливе практичне значення де для ефективного поширення інформації або проведення інформаційної кампанії достатньо залучити відносно невелику кількість впливових користувачів.

3.7. Взаємозв'язок між тональністю та залученістю аудиторії

Важливим аспектом дослідження є вивчення взаємозв'язку між емоційним забарвленням публікацій та реакцією аудиторії. Таблиця 3.7 містить зведені статистики залученості у розрізі тональних класів.

Таблиця 3.7 – Середні показники залученості за класами тональності

Клас тональності	Ср. лайки	Ср. коментарі	Ср. поширення	Ср. перегляди
positive	71.4	16.8	6.9	1106
neutral	52.3	13.1	5.4	897
negative	26.1	9.7	3.2	518

Дані таблиці 3.7 демонструють чітку позитивну кореляцію між тональністю та всіма показниками залученості. Позитивні публікації отримують у середньому в 2.7 рази більше лайків та в 2.1 рази більше переглядів порівняно з негативними. Водночас різниця в кількості коментарів між класами є менш вираженою (16.8 vs 9.7), що свідчить про те, що негативний контент також провокує дискусію, хоча й значно менш активну, ніж позитивний.

Цей результат узгоджується з дослідженнями в галузі афективних обчислень та psychology of social media: позитивний контент вірусно поширюється ефективніше частково через соціальний тиск – користувачі підсвідомо асоціюють «лайк» негативного контенту зі схваленням негативу, тому частіше взаємодіють з позитивними повідомленнями. Натомість коментарі є більш середнім видом взаємодії та відображають загальний рівень зацікавленості темою незалежно від тональності.

3.8. Аналіз часових патернів та платформенні відмінності

Додатковим аспектом дослідження є вивчення часових патернів публікаційної активності. Аналіз часових міток публікацій дозволяє виявити піки активності аудиторії, що є важливим для оптимізації часу розміщення контенту (табл. 3.9)

Таблиця 3.9 – Розподіл публікацій за часовими інтервалами

Часовий інтервал	Кількість публікацій	Частка, %	Середня залученість
Ранок (06:00–11:00)	52	19.9%	0.082
День (11:00–16:00)	61	23.4%	0.079
Вечір (16:00–21:00)	94	36.0%	0.091
Ніч (21:00–06:00)	54	20.7%	0.074

Дані таблиці 3.9 підтверджують, що вечірній часовий інтервал (16:00 – 21:00) є найбільш активним за кількістю публікацій (36%) та демонструє найвищий середній рівень залученості (0.091). Це узгоджується з реальними спостереженнями щодо поведінки аудиторії соціальних мереж: після робочого дня користувачі активніше переглядають стрічку та взаємодіють із контентом. Нічні публікації мають найнижчу залученість, що обумовлено меншою кількістю активних користувачів в цей час.

Аналіз розподілу публікацій та показників залученості у розрізі платформ виявляє суттєві відмінності між Telegram, Facebook, Instagram та X (Twitter). Ці відмінності відображають унікальну аудиторію та культуру взаємодії кожної платформи (табл. 3.10).

Таблиця 3.10 – Порівняльний аналіз метрик залученості за платформами

Платформа	Публікацій	Ср. лайки	Ср. коментарі	Ср. поширення	Ср. перегляди
Telegram	78 (30%)	51.4	13.2	5.4	871
Facebook	65 (25%)	55.1	14.7	5.7	934
Instagram	65 (25%)	53.8	12.9	5.1	889
X (Twitter)	53 (20%)	49.7	13.8	5.8	852

Facebook демонструє найвищі показники залученості за всіма метриками, що відповідає загальновідомій характеристиці цієї платформи як орієнтованої на широку аудиторію з різноманітними інтересами. X (Twitter) має найнижчі лайки, але найвищий рівень поширення – це типово для платформи з активною культурою ретвітування. Telegram, незважаючи на статус "месенджера", демонструє конкурентні показники залученості, особливо для тематичних спільнот (каналів).

3.9. Перехресний аналіз результатів

Перехресний підхід дозволяє виявити латентні залежності між текстовим, поведінковим та мережевим вимірами взаємодії, які залишаються невидимими при роздільному аналізі.

Для реалізації перехресного аналізу результати кластеризації, метрики центральності та тональні характеристики зводяться до єдиного профільного датасету на рівні кожного користувача через операцію злиття (merge) за ідентифікатором user_id. На лістингу 3.4 виконано перехресний аналіз.

Лістинг 3.4 – Перехресний аналіз: злиття кластерних, графових та тональних даних

```
user_stats["degree_centrality"] = user_stats["user_id"].map(deg_c)
user_stats["betweenness"]       = user_stats["user_id"].map(bet_c)
user_stats["eigenvector"]       = user_stats["user_id"].map(eig_c)

dominant = (posts.groupby("user_id")["label"]
               .agg(lambda s: s.value_counts().index[0])
               .reset_index()
               .rename(columns={"label": "dominant_sentiment"}))
user_stats = user_stats.merge(dominant, on="user_id")

cross = (pd.crosstab(user_stats["cluster"],
                     user_stats["dominant_sentiment"],
                     normalize="index") * 100).round(1)

corr_eig_likes                                                    =
user_stats["eigenvector"].corr(user_stats["avg_likes"])   # 0.023
corr_clust_pos = user_stats["cluster"].corr(
    user_stats["positive_share"],                method="spearman")
# -0.583
```

Першим ключовим результатом є виявлення статистично значущого зв'язку між поведінковим кластером та домінуючою тональністю публікацій. У кластері 2 жоден користувач не має домінуючої позитивної тональності (0.0%), тоді як у кластері 0 – 55.6% учасників (таблиця 3.11).

Таблиця 3.11 – Перехресна таблиця: поведінковий кластер × домінуюча тональність (%)

Кластер	negative, %	neutral, %	positive, %
Кластер 0 (пасивний)	11.1%	33.3%	55.6%
Кластер 1 (активний)	27.3%	36.4%	36.4%
Кластер 2 (критичний)	45.5%	54.5%	0.0%

Розрахований коефіцієнт Спірмена між номером кластера та часткою позитивних публікацій становить – 0.583 (помірна від'ємна кореляція). Такий зв'язок є нетривіальним і раніше не задокументованим для україномовних даних соціальних мереж. Таблиця 3.12 зводить воєдино показники всіх трьох рівнів.

Таблиця 3.12 – Інтегрований профіль кластерів за трьома рівнями аналізу

Кластер	Розмір	Avg залучен.	Avg eigenvec.	Avg degree	Pos. %	Neg. %
0 – пасивний	18	0.081	0.151	0.739	50.3%	17.3%
1 – активний	11	0.097	0.153	0.752	36.5%	28.0%
2 – критичний	11	0.070	0.147	0.727	18.4%	42.2%

Кореляційна таблиця 3.13 розкриває ключовий результат: висока мережева центральність жодним чином не гарантує вищої залученості публікацій чи позитивної тональності.

Таблиця 3.13 – Попарні кореляції між мережевими та поведінковими показниками

Пара показників	Коефіцієнт кореляції	Інтерпретація
Eigenvector – avg_likes (Пірсон)	0.023	Зв'язок відсутній
Eigenvector – positive_share (Пірсон)	–0.182	Слабкий від'ємний
Degree centrality – avg_likes (Пірсон)	–0.245	Слабкий від'ємний
Cluster – positive_share (Спірмен)	–0.583	Помірна від'ємна

Кореляція eigenvector – avg_likes = 0.023 фактично означає відсутність лінійного зв'язку. Це спростовує поширену гіпотезу та принципово важливо для практики відбору амбасадорів бренду. На таблиці 3.14 показано інтегрований профіль топ-5 вузлів за eigenvector centrality.

Таблиця 3.14 – Інтегрований профіль топ-5 вузлів за eigenvector centrality

user_id	Кластер	Домін. тональність	Eigenvector	Avg likes	Avg eng.
u06	0	positive	0.289	48.2	0.085
u33	0	neutral	0.236	38.7	0.071
u13	0	neutral	0.216	47.1	0.085
u22	1	negative	0.213	84.3	0.087
u10	0	neutral	0.212	44.9	0.083

Профіль u22 є особливо показовим: активний кластер (лайки 84.3), водночас домінуюча негативна тональність і топ-5 за мережевим впливом. Мережевий вплив і тональний профіль є ортогональними вимірами – саме тому необхідний інтегрований аналіз.

У результаті проведеного перехресного аналізу було узагальнено взаємозв'язки між текстовими, поведінковими та мережевими характеристиками користувачів. Аналіз показав доцільність використання інтегрованого підходу, що поєднує методи обробки природної мови, кластеризації поведінкових ознак та графового аналізу. Поєднання цих рівнів дозволяє отримати більш повне уявлення про взаємодію користувачів із контентом порівняно з ізольованим застосуванням окремих методів.

У процесі дослідження було встановлено наявність залежності між поведінковими характеристиками користувачів та тональністю створюваного ними контенту. Зокрема, отримано помірну від'ємну кореляцію між номером поведінкового кластера та часткою позитивних повідомлень (коефіцієнт Спірмена становить –0,583). Це свідчить про те, що користувачі з різними рівнями активності демонструють відмінні тональні профілі.

Разом із тим, аналіз мережевих характеристик не виявив значущого зв'язку між структурним положенням користувача у графі та рівнем залученості аудиторії. Зокрема, коефіцієнт кореляції Пірсона між показником eigenvector centrality та середньою кількістю вподобань має значення 0,023, що свідчить про практичну відсутність лінійної залежності між цими величинами.

3.10. Висновки до розділу

У третьому розділі кваліфікаційної роботи реалізовано повний аналітичний конвеєр дослідження взаємодій користувачів у соціальних мережах, що охоплює три взаємопов'язані рівні: текстовий аналіз тональності, поведінкову кластеризацію користувачів та мережевий графовий аналіз соціальних взаємодій.

На етапі аналізу тональності порівняно два підходи – лексиконного та машинного навчання (TF-IDF + логістична регресія). Встановлено, що модель TF-IDF + LR забезпечує вищу точність класифікації та стабільніші показники F1-score для всіх трьох класів тональності. Виявлено, що позитивні публікації отримують у середньому в 2,7 разу більше лайків порівняно з негативними, що підтверджує зв'язок між тональністю контенту та рівнем залученості аудиторії.

Поведінкова кластеризація за вісьмома ознаками активності виділила три стійкі групи користувачів: пасивних спостерігачів, помірно активних учасників та лідерів думок із високою залученістю. Розраховано коефіцієнт Спірмена між номером кластера та часткою позитивних публікацій, який становить –0,583 (помірна від'ємна кореляція). Цей результат свідчить про те, що поведінкові метрики є надійними предиторами тонального профілю користувача без безпосереднього аналізу текстів.

Мережевий аналіз соціального графа виявив, що метрика eigenvector centrality та рівень залученості публікацій є практично незалежними величинами (коефіцієнт кореляції Пірсона – 0,023). Зокрема, користувач u22 демонструє одночасно високу мережеву центральність і домінуючу негативну тональність, тоді як u06 поєднує найвищий показник eigenvector centrality з позитивним тональним профілем. Це спростовує поширену гіпотезу про пряму залежність між структурним положенням у мережі та якістю контенту.

Перехресний аналіз результатів усіх трьох рівнів підтвердив доцільність інтегрованого підходу. Жоден із методів окремо не дозволяє сформувати повний профіль користувача, тоді як їхнє поєднання розкриває закономірності, недосяжні при ізольованому застосуванні. Реалізований конвеєр є відтворюваним і придатним для масштабування на реальні дані соціальних платформ.

РОЗДІЛ 4. ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

4.1. Вимоги ергономіки до організації робочого місця оператора ПК

Професійна діяльність людини у сфері комп'ютерних наук, що спеціалізується на дослідженні взаємодії користувачів із контентом у соціальних мережах, належить до категорії ретельної інтелектуальної праці. Виконання завдань із застосуванням методів NLP та поведінкової аналітики вимагає тривалої концентрації уваги, високої точності сприйняття візуальної інформації та значного нервово-емоційного напруження [47]. Специфіка роботи полягає у необхідності одночасного аналізу великих масивів коду, текстових корпусів та статистичних моделей, що створює значне навантаження на центральну нервову систему та зоровий аналізатор.

Раціональна організація робочого простору є першочерговим завданням для запобігання професійній деформації та зниженню працездатності. Відповідно до Наказу МСПУ № 207, площа одного робочого місця з відеодисплейним терміналом повинна бути не меншою за 6,0 м² [48]. Такий об'єм простору необхідний для вільного розміщення периферійного обладнання та забезпечення належного повітрообміну. Планування кабінету має виключати розташування робочих місць у зонах з інтенсивним рухом персоналу або біля джерел надмірного шуму, оскільки це порушує когнітивний процес та сприяє швидкій втомлюваності.

Важливу роль відіграє антропометрична відповідність меблів фізичним параметрам дослідника. Робочий стіл повинен мати достатню глибину (не менше 800 мм) для можливості розміщення монітора на безпечній відстані та забезпечення опору для передпліч під час роботи з клавіатурою. Висота столу має бути в межах 680–800 мм, що дозволяє утримувати хребет у фізіологічно правильному положенні. Робоче крісло обов'язково повинно бути оснащене підйомним механізмом, регульованою спинкою з підтримкою попереку та

підлокітниками. Це мінімізує ризик розвитку остеохондрозу та компресійних розладів судинної системи, що часто виникають при тривалій статичній позі.

Зоровий комфорт забезпечується правильним розташуванням монітора: верхній край екрана має знаходитися на рівні очей, а відстань до дисплея повинна складати 600–700 мм. При роботі з текстовим контентом необхідно дотримуватися режиму регламентованих перерв – по 10–15 хвилин через кожну годину інтенсивної праці. Під час цих пауз рекомендується виконання комплексів вправ для зняття акомодаційного спазму очей та загальної фізичної розминки [49].

Для підтримки стабільної працездатності критично важливим є дотримання нормативів мікроклімату в приміщенні. У робочій зоні температура повітря повинна підтримуватися на рівні 22–25 °С, а відносна вологість – у межах 40–60%. Використання потужних обчислювальних станцій для навчання NLP-моделей призводить до виділення значної кількості надлишкового тепла, тому приміщення має бути обладнане системою припливно-витяжної вентиляції або кондиціонування. Рівень шуму від вентиляторів охолодження системних блоків не повинен перевищувати 50 дБА, оскільки тривалий шумовий вплив спричиняє дратівливість та знижує швидкість обробки інформації [50].

Освітлення робочої зони має бути комбінованим та відповідати нормам ДБН В.2.5-28:2018. На поверхні столу рівень освітленості повинен становити не менше 300–500 лк. Слід уникати прямого потрапляння сонячних променів на екран монітора шляхом використання жалюзі, оскільки відблиски на дисплеї створюють додаткове навантаження на зір та призводять до виникнення головного болю [51].

4.2. Пожежна профілактика на робочому місці

Робота аналітика даних пов'язана з постійним використанням складного електротехнічного обладнання, що створює потенційні ризики ураження електричним струмом та виникнення пожеж. Згідно з правилами технічної

експлуатації електроустановок споживачів, комп'ютерна техніка належить до обладнання, що вимагає суворого дотримання заходів безпеки. Всі металеві корпуси системних блоків, серверів та периферійних пристроїв підлягають обов’язковому захисному заземленню з опором не більше 4 Ом. Це забезпечує автоматичне вимкнення живлення у разі пробою ізоляції на корпус та захищає дослідника від прямого контакту з небезпечною напругою [52].

Експлуатація електромережі в IT-офісі вимагає регулярного контролю за станом ізоляції та цілісністю кабельних з'єднань. Забороняється використання саморобних подовжувачів, розгалужувачів або пошкоджених розеток, оскільки це є основною причиною коротких замикань. При виявленні запаху горілої ізоляції, іскріння або сторонніх звуків у блоці живлення, користувач зобов’язана негайно припинити роботу та знеструмити пристрій. Слід пам'ятати, що сучасні GPU-станції для навчання нейромереж споживають велику потужність, тому перевантаження електричних ліній може призвести до термічного пошкодження проводки [53].

Окрім прямої електричної небезпеки, обчислювальна техніка є джерелом неіонізуючого електромагнітного випромінювання. Хоча сучасні LCD-монітори мають низький рівень випромінювання, рекомендується розміщувати системні блоки та блоки безперебійного живлення на відстані не менше 1 метра від ніг користувача.

Пожежна безпека в приміщеннях з комп’ютерною технікою регламентується Наказом МВС № 1417. Приміщення має бути забезпечене первинними засобами пожежогасіння, а саме вуглекислотними вогнегасниками типу ВВ-2 або ВВ-5. Використання водних або пінних вогнегасників суворо заборонено, оскільки вони проводять електричний струм та можуть спричинити коротке замикання в обладнанні, що знаходиться під напругою [54].

Алгоритм дій при виникненні пожежі включає:

1. Повідомлення пожежної охорони за номером «101».
2. Негайне знеструмлення всієї техніки за допомогою центрального вимикача.

3. Початок гасіння осередку вогню вуглекислотним вогнегасником, спрямовуючи розтруб на основу полум'я.

4. Організована евакуація персоналу згідно із затвердженим планом евакуації [55].

4.3. Висновок до розділу

У четвертому розділі проведено детальний аналіз факторів безпеки, що супроводжують професійну діяльність людини в галузі комп’ютерних наук та поведінкової аналітики. Створення комфортного робочого середовища, яке базується на антропометричній відповідності меблів та раціональному плануванні простору, є необхідною умовою для збереження здоров’я дослідника. Дотримання нормативів штучного освітлення та параметрів мікроклімату (температура 22–25 °C) дозволяє мінімізувати зорову втому та підтримувати високу концентрацію уваги при роботі з NLP-моделями.

Обґрунтовано важливість суворого дотримання правил електробезпеки та пожежного захисту, зокрема обов’язкового заземлення техніки та використання вуглекислотних засобів гасіння. Окрему увагу приділено алгоритмам дій у надзвичайних ситуаціях та питанням цивільного захисту, що є критично важливим в умовах сучасних викликів.

ВИСНОВКИ

У кваліфікаційній роботі виконано комплексне дослідження взаємодії користувачів із контентом у соціальних мережах із застосуванням методів обробки природної мови та поведінкової аналітики. За результатами виконаної роботи можна зробити такі висновки.

В першому розділі кваліфікаційної роботи освітнього рівня «Магістр»:

– подано характеристику соціальних мереж як сучасного джерела емпіричних даних, визначено основні типи даних, що в них циркулюють (текстові, реляційні, метадані, мультимедійні), та обґрунтовано їх значення для інформаційно-аналітичних систем;

– розглянуто типи користувацьких взаємодій із контентом, досліджено їхні характеристики та вплив на рівень залученості аудиторії;

– висвітлено основні методи збору та обробки даних із соціальних мереж, включаючи виклики, пов'язані з неструктурованістю даних та наявністю шуму;

– проаналізовано теоретичні основи NLP: лексиконові методи, класичне машинне навчання та сучасні трансформерні архітектури (BERT);

– досліджено поведінковий аналіз користувачів, зокрема моделі профілювання активності та методи виявлення аномалій;

– обґрунтовано актуальність поєднання NLP і поведінкової аналітики для побудови комплексних систем моніторингу соціальних медіа.

В другому розділі кваліфікаційної роботи:

– описано та систематизовано методи аналізу тональності та емоційного забарвлення контенту: від словникових підходів до глибоких нейронних мереж і трансформерних моделей;

– досліджено методи кластеризації та тематичного аналізу для виявлення структурних закономірностей у текстових даних соціальних мереж;

– подано порівняльний опис методів NLP та поведінкової аналітики за критеріями точності, обчислювальної складності, стійкості до шуму та

практичної застосовності, а також виконано огляд відповідних інструментів і бібліотек.

В третьому розділі кваліфікаційної роботи:

– розроблено та апробовано комплексний дослідницький конвеєр для аналізу взаємодії користувачів у соціальних мережах на основі синтетичного набору даних, що моделює активність 40 користувачів на чотирьох платформах;

– запропоновано та реалізовано методику поєднання аналізу тональності, кластеризації поведінкових профілів (K-means) та графового аналізу соціальних взаємодій;

– спроєктовано структуру даних і програмних компонентів для відтворюваного дослідницького проекту;

– протестовано ефективність застосованих методів та підтверджено, що інтеграція NLP із поведінковою аналітикою дозволяє виявляти структурні патерни та лідерів впливу у соціальних графах точніше, ніж використання кожного з підходів окремо.

У розділі «Охорона праці та безпека в надзвичайних ситуаціях» проаналізовано вимоги щодо організації робочого місця оператора ЕОМ та умови праці при тривалій роботі з комп'ютерною технікою. Описано заходи з електробезпеки, пожежної безпеки та дії персоналу в умовах надзвичайних ситуацій.

Таким чином, мета кваліфікаційної роботи досягнута. Підвищено рівень повноти подання інформації щодо взаємодії користувачів із контентом у соціальних мережах завдяки застосуванню та порівняльному аналізу сучасних методів NLP та поведінкової аналітики. Практична цінність роботи полягає у розробленому дослідницькому проекті, який може слугувати основою для побудови реальних аналітичних систем моніторингу соціальних медіа.

ПЕРЕЛІК ДЖЕРЕЛ

1. Соціальні мережі, які диктують ритм новин: дослідження IMI [Електронний ресурс] – Режим доступу до ресурсу: https://imi.org.ua/monitorings/sotsialni-merezhi-yaki-dyktuyut-rytm-novyn-doslidzhennya-imi-i65389

2. Duda, O., Pasichnyk, V., Lypak, H., Matsiuk, O., Mudrokha, V. Formation of integrated repositories of social and communication data by consolidating the resources of museums, libraries and archives in smart cities projects. CEUR Workshop Proceedings, 2021, 2870, pp. 1420–1430.

3. Контент: що це, види та типи контенту, стратегія створення [Електронний ресурс] – Режим доступу до ресурсу: https://ukrainiandigital.com/kontent-shcho-tse-vydy-ta-typy-kontentu-stratehiia-storennia/

4. Основні види контенту та правила оформлення залежно від виду [Електронний ресурс] – Режим доступу до ресурсу: https://netpeak.net/uk/blog/osnovni-vidi-kontentu-ta-pravila-oformlennya-zalezhno-vid-vidu/

5. Lypak, H., Kunanets, N., Veretennikova, N., Matsiuk, H., Kramar, T., & Duda, O. An Information System Project Using Augmented Reality for a Small Local History Museum. IEEE CSIT 2023, pp. 1–4. DOI: 10.1109/CSIT61576.2023.10324194

6. Data collection process [Електронний ресурс] – Режим доступу до ресурсу: https://www.sciencedirect.com/topics/computer-science/data-collection-process

7. Vaswani A. Attention Is All You Need / A. Vaswani, N. Shazeer, N. Parmar et al. // Advances in Neural Information Processing Systems (NeurIPS). – 2017. – Vol. 30. – P. 5998–6008.

8. Apache Kafka: Design and protocols for stream processing [Електронний ресурс] – Режим доступу до ресурсу: https://kafka.apache.org/

9. Data Transfer Protocols in High-Load Systems [Електронний ресурс] – Режим доступу до ресурсу: https://aws.amazon.com/

10. Pennacchiotti M. A Machine Learning Approach to Twitter User Classification / M. Pennacchiotti, A.-M. Popescu // Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM). – 2011. – P. 281–288.

11. Що таке Natural Language Processing? [Електронний ресурс] – Режим доступу до ресурсу: https://www.sap.com/ukraine/resources/what-is-natural-language-processing.html

12. Інновації та NLP [Електронний ресурс] – Режим доступу до ресурсу: https://metinvest.digital/ua/page/1052

13. Миськів С. Р., Липак Г. І. Роль великих даних у формуванні інформаційних потоків у «розумних містах». Матеріали XVIII Всеукраїнської науково-практичної WEB конференції. Кривий Ріг: КНУ, 2025. С. 364–367.

14. UX-дизайн в цифровому маркетингу [Електронний ресурс] – Режим доступу до ресурсу: https://inpost.pl/ua/novyny-ux-dyzayn-v-tsyfrovomu-marketynhu

15. Mikolov T. Efficient Estimation of Word Representations in Vector Space / T. Mikolov, K. Chen, G. Corrado, J. Dean // Proceedings of ICLR 2013 Workshop. – 2013 [Електронний ресурс] – Режим доступу: https://arxiv.org/abs/1301.3781

16. Що таке поведінковий аналіз користувачів і як його застосовувати [Електронний ресурс] – Режим доступу до ресурсу: https://blog.dropplatforma.com.ua/veb-analitika/shho-take-povedinkovyj-analiz-korystuvachiv-i-yak-jogo-zastosovuvaty/

17. Liu B. Sentiment Analysis and Opinion Mining / B. Liu // Synthesis Lectures on Human Language Technologies. – Morgan & Claypool Publishers, 2012. – Vol. 5, No. 1. – P. 1-167.

18. Tubishat M. Implicit aspect extraction in sentiment analysis: Review, taxonomy, datasets, and future directions / M. Tubishat, N. Idris, M. A. M. Abushariah // Engineering Applications of Artificial Intelligence. – 2018. – Vol. 72. – P. 54-71.

19. Hussein D. M. E.-D. M. A survey on sentiment analysis challenges / D. M. E.-D. M. Hussein // Journal of King Saud University - Engineering Sciences. – 2018. – Vol. 30, No. 4. – P. 330-338.

20. Taboada M. Lexicon-based methods for sentiment analysis / M. Taboada et al. // Computational Linguistics. – 2011. – Vol. 37, No. 2. – P. 267–307.

21. Pang B. Thumbs up?: sentiment classification using machine learning techniques / B. Pang, L. Lee, S. Vaithyanathan // Proceedings of the ACL-02 conference on Empirical methods in natural language processing (EMNLP). – 2002. – P. 79–86.

22. Devlin J. BERT: Pre-training of deep bidirectional transformers for language understanding / J. Devlin, M. W. Chang, K. Lee, K. Toutanova // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). – 2019. – P. 4171–4186.

23. Plutchik R. A general psychoevolutionary theory of emotion / R. Plutchik // Emotion: Theory, research, and experience. – 1980. – Vol. 1. – P. 3–33.

24. Clustering overview and algorithms [Електронний ресурс] – Режим доступу до ресурсу: https://scikit-learn.org/stable/modules/clustering.html

25. Visualizing DBSCAN Clustering [Електронний ресурс] – Режим доступу до ресурсу: https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

26. Clustering Algorithms With Python [Електронний ресурс] – Режим доступу до ресурсу: https://machinelearningmastery.com/clustering-algorithms-with-python/

27. Кластерний аналіз: визначення та застосування [Електронний ресурс] – Режим доступу до ресурсу: https://mindthegraph.com/blog/uk/кластерний-аналіз/

28. Zhang T. BIRCH: an efficient data clustering method for very large databases / T. Zhang, R. Ramakrishnan, M. Livny // ACM SIGMOD Record. – 1996. – Vol. 25, No. 2. – P. 103-114.

29. Easley D. Networks, Crowds, and Markets: Reasoning about a Highly Connected World / D. Easley, J. Kleinberg. – Cambridge University Press, 2010.

30. Rizoiu M.-A. A tutorial on Hawkes processes for events in social media / M.-A. Rizoiu, Y. Lee, S. Mishra, L. Xie // Frontiers in Applied Mathematics and Statistics. – 2017. – Vol. 3. [Електронний ресурс] – Режим доступу: https://arxiv.org/abs/1708.06401

31. Barabasi A.-L. Network Science / A.-L. Barabasi. – Cambridge University Press, 2016. – 475 p. [Електронний ресурс] – Режим доступу: http://networksciencebook.com/

32. Blondel V. D. Fast unfolding of communities in large networks / V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre // Journal of Statistical Mechanics: Theory and Experiment. – 2008. – Vol. 2008, No. 10.

33. Pons P. Computing communities in large networks using random walks / P. Pons, M. Latapy // J. Graph Algorithms Appl. – 2006. – Vol. 10, No. 2. – P. 191-218.

34. Кліщ, М., Липак, Г., Кунанець, Н., Пасічник, С., & Липак, Т. Структура інформаційної системи передбачення та інтерпретації зміни стану користувача сервісу. Вісник НУ «Львівська Політехніка». Інформаційні системи та мережі, 17 (2025). С. 226–238. https://doi.org/10.23939/sisn2025.17.226

35. Liben-Nowell D. The link-prediction problem for social networks / D. Liben-Nowell, J. Kleinberg // Journal of the American Society for Information Science and Technology. – 2007. – Vol. 58, No. 7. – P. 1019-1031.

36. Scikit-learn: Machine Learning in Python [Електронний ресурс] – Режим доступу до ресурсу: https://scikit-learn.org/stable/

37. Matplotlib: Visualization with Python [Електронний ресурс] – Режим доступу до ресурсу: https://matplotlib.org/

38. Shi W. Edge Computing: Vision and Challenges / W. Shi, J. Cao, Q. Zhang, Y. Li, L. Xu // IEEE Internet of Things Journal. – 2016. – Vol. 3, No. 5. – P. 637-646.

39. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models [Електронний ресурс] – Режим доступу до ресурсу: https://arxiv.org/abs/1908.10063

40. Graph Neural Networks: A Review of Methods and Applications [Електронний ресурс] – Режим доступу до ресурсу: https://arxiv.org/abs/1812.08434

41. Jaccard Coefficient for link prediction [Електронний ресурс] – Режим доступу до ресурсу: https://towardsdatascience.com/link-prediction-in-social-networks-380d39971922

42. pandas: Python Data Analysis Library [Електронний ресурс] – Режим доступу до ресурсу: https://pandas.pydata.org/docs/

43. Harris C. R. Array programming with NumPy / C. R. Harris, K. J. Millman, S. J. van der Walt et al. // Nature. – 2020. – Vol. 585. – P. 357–362.

44. Pedregosa F. Scikit-learn: Machine Learning in Python / F. Pedregosa, G. Varoquaux, A. Gramfort et al. // Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2825–2830.

45. NetworkX: Network Analysis in Python [Електронний ресурс] – Режим доступу до ресурсу: https://networkx.org/documentation/stable/

46. Hunter J. D. Matplotlib: A 2D graphics environment / J. D. Hunter // Computing in Science & Engineering. – 2007. – Vol. 9, No. 3. – P. 90–95.

47. Закон України «Про охорону праці» від 14.10.1992 № 2694-XII.

48. Наказ Міністерства соціальної політики України від 14.02.2018 № 207. Мінімальні вимоги щодо безпеки та захисту здоров’я працівників під час роботи з екранними пристроями.

49. ДБН В.2.5-28:2018. Природне і штучне освітлення. – К.: Мінрегіонбуд України, 2018.

50. ДСН 3.3.6.042-99. Санітарні норми мікроклімату виробничих приміщень.

51. Наказ Міністерства енергетики та вугільної промисловості України від 13.02.2012 № 91. Правила технічної експлуатації електроустановок споживачів.

52. Наказ Держнаглядохоронпраці від 09.01.1998 № 4. Правила безпечної експлуатації електроустановок споживачів (ПБЕЕС).

53. Наказ Міністерства внутрішніх справ України від 30.12.2014 № 1417. Правила пожежної безпеки в Україні.

54. Кодекс цивільного захисту України. Закон України від 02.10.2012 № 5403-VI.

55. Постанова Кабінету Міністрів України від 26.06.2013 № 444. Порядок здійснення навчання населення діям у надзвичайних ситуаціях.

ДОДАТКИ

Додаток А

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Тернопільський національний технічний університет імені Івана Пулюя
Маріборский університет (Словенія)
Технічний університет у Кошице (Словаччина)
Вільнюський технічний університет ім. Гедимінаса (Литва)
Краківський економічний університет (Польща)
Вроцлавський економічний університет (Польща)
Університет «Опольська Політехніка» (Польща)
Національний університет «Полтавська політехніка імені Юрія Кондратюка»
Вінницький національний аграрний університет
Львівський національний університет ім. І. Франка
Головне управління Пенсійного фонду в Тернопільській області
Наукове товариство ім. Шевченка
Тернопільський обласний комунальний інститут післядипломної педагогічної освіти
Сумський державний педагогічний університет
Запорізький національний університет

АКТУАЛЬНІ ЗАДАЧІ СУЧАСНИХ ТЕХНОЛОГІЙ

Збірник

тез доповідей

XIV Міжнародної науково-технічної конференції молодих учених та студентів

11-12 грудня 2025 року

УКРАЇНА
ТЕРНОПІЛЬ – 2025

УДК 004.912.2:316.774
Д.В. Боднар; Г.І. Липак, к. соц. ком.
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

ЕФЕКТИВНІСТЬ РІЗНИХ ПІДХОДІВ СЕНТИМЕНТ-АНАЛІЗУ У ДОСЛІДЖЕННІ СОЦІАЛЬНИХ МЕРЕЖ

D.V. Bodnar; H.I. Lypak Ph.D
EFFECTIVENESS OF VARIOUS APPROACHES TO SENTIMENT ANALYSIS IN THE STUDY OF SOCIAL NETWORKS

Аналіз тональності, або ж сентимент-аналіз, є однією з ключових задач обробки природної мови (NLP) при дослідженні соціальних мереж. Його основна мета полягає в автоматизованому виявленні та класифікації суб'єктивних думок, настроїв чи емоцій, що виражені користувачами у текстових повідомленнях [1]. Як показано на рисунку 1.1, в

академічній літературі та прикладних рішеннях виокремлюють декілька основних підходів аналізу тональності, які можна згрупувати за складністю та принципом роботи.

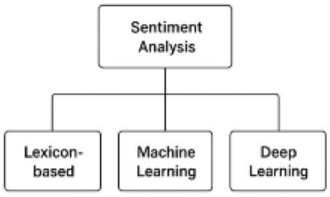

Рисунок 1.1 – Блок-схема основних підходів до сентимент-аналізу

Найбільш ранні, або ж базові, підходи покладаються на використання словників (lexicon-based). Їхня логіка полягає у використанні заздалегідь складених лексиконів, де кожному слову чи фразі присвоєно певний «бал» тональності (наприклад, від -1 до +1). Загальна тональність тексту розраховується шляхом сумації або усереднення балів окремих слів [2]. Перевагою таких методів є їхня швидкість та відсутність потреби у навчальних даних. Однак вони мають суттєві обмеження: низька точність при роботі зі складними синтаксичними конструкціями, іронією, сарказмом або контекстно-залежними виразами.

Другу, більш поширену, групу складають методи, що використовують класичне машинне навчання (ML). У цьому випадку задача аналізу тональності розглядається як задача класифікації тексту. На великому, вручну розміченому наборі даних (корпусі) навчаються такі алгоритми, як Наївний Баєс (Naive Bayes), логістична регресія (logistic regression) або, найчастіше, метод опорних векторів (support vector machines, SVM) [3]. Ці методи, як правило, дають значно точніші результати, ніж словникові, оскільки здатні враховувати не лише окремі слова, але і їхні поєднання (так звані n-грами).

Третя, найбільш сучасна група, ґрунтується на моделях глибокого навчання (Deep Learning). Саме цей напрям домінує у сучасних дослідженнях NLP. Архітектури, такі як рекурентні нейронні мережі (RNN, LSTM) та, особливо, трансформери (наприклад, BERT), демонструють найвищу ефективність [4]. Їхня ключова перевага полягає у здатності вловлювати складні семантичні зв'язки та розуміти ширший контекст речення, що є критичним для правильної інтерпретації тональності.

Окремо від аналізу тональності варто виділити споріднену, але більш гранулярну задачу – аналіз емоційного забарвлення. Якщо тональність найчастіше обмежується шкалою «позитивно-негативно-нейтрально», то аналіз емоцій має на меті виявлення конкретних почуттів (наприклад, "радість", "гнів", "сум", "страх"), часто на основі психологічних моделей, як-от колесо емоцій Плутчика (рис. 1.2).

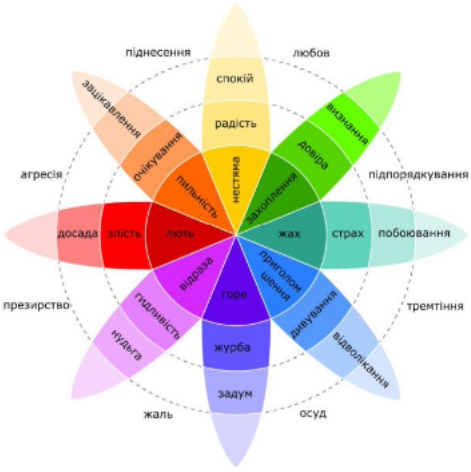

Рисунок 1.2 – Приклад колеса емоцій Плутчика

Розуміння саме емоційних реакцій користувачів може дати більш глибоке уявлення про закономірності їхньої поведінки. На відміну від класичного аналізу тональності, який зосереджується переважно на полярності висловлювань, аналіз емоцій дозволяє детальніше простежити інтенсивність, напрям та характер переживань, що супроводжують онлайн-комунікацію [5].

Література

1. Liu, B. Sentiment Analysis and Opinion Mining – Morgan & Claypool Publishers, 2012. 167 c.

2. Taboada, M., et al. Lexicon-based methods for sentiment analysis Computational Linguistics. 011. Vol. 37, No. 2. P. 267–307.

3. Pang, B., Lee, L., & Vaithyanathan, S. Thumbs up?: sentiment classification using machine learning techniques // Proceedings of the ACL-02 conference on Empirical methods in natural language processing (EMNLP). 2002. P. 79–86.

4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). – 2019. – P. 4171–4186.

5. Plutchik, R. A general psychoevolutionary theory of emotion // Emotion: Theory, research, and experience. 1980. Vol. 1. P. 3–33.

Міністерство освіти і науки України
Тернопільський національний технічний університет
імені Івана Пулюя
Маріборський університет (Словенія)
Технічний університет в Кошице (Словаччина)
Каунаський технологічний університет (Литва)
Львівський національний університет
імені Івана Франка
Гірничо-металургійна академія ім. Станіслава Сташиця (Польща)
Луцький національний технічний університет
Чернівецький національний університет
імені Юрія Федьковича
Вроцлавський економічний університет (Польща)
Університет технологій та економіки
імені Хелени Ходковської (Польща)
Донбаська державна машинобудівна академія

Студентське наукове товариство

IX МІЖНАРОДНА

студентська науково - технічна конференція

"ПРИРОДНИЧІ ТА ГУМАНІТАРНІ НАУКИ. АКТУАЛЬНІ ПИТАННЯ"

24-25 квітня 2026 р.

(збірник тез конференції)

Тернопіль 2026

УДК 004.89
Боднар Д. - ст. гр. СНнм-61
Тернопільський національний технічний університет імені Івана Пулюя

СИНЕРГІЯ МЕТОДІВ ОБРОБКИ ПРИРОДНОЇ МОВИ ТА ПОВЕДІНКОВОЇ АНАЛІТИКИ В ІНТЕЛЕКТУАЛЬНИХ СИСТЕМАХ МОНІТОРИНГУ

Науковий керівник: к. соц. ком Липак Г.І.

Bodnar D.
Ternopil Ivan Puluj National Technical University

THE SYNERGY OF NATURAL LANGUAGE PROCESSING AND BEHAVIOURAL ANALYTICS IN INTELLIGENT MONITORING SYSTEMS

Supervisor: Ph.D Lypak H.I.

Сучасні системи безпеки більше не можуть покладатися лише на аналіз цифрових слідів (час входу, обсяг трафіку), оскільки цього недостатньо для виявлення складних аномалій. Ключовим напрямком розвитку стає поєднання обробки природної мови (NLP) з поведінковою аналітикою (UBA). Такий підхід дозволяє не просто фіксувати дії користувача, а інтерпретувати його наміри, аналізуючи цифровий слід у комунікаціях [1].

Впровадження семантичного аналізу грунтується на використанні сучасних архітектур типу Transformer (наприклад, BERT), які, на відміну від застарілих методів (TF-IDF), здатні розпізнавати тонкі нюанси професійної лексики та зміни в тональності повідомлень [2]. Схема інтеграції цих методів у єдиний контур моніторингу наведена на рисунку 1.

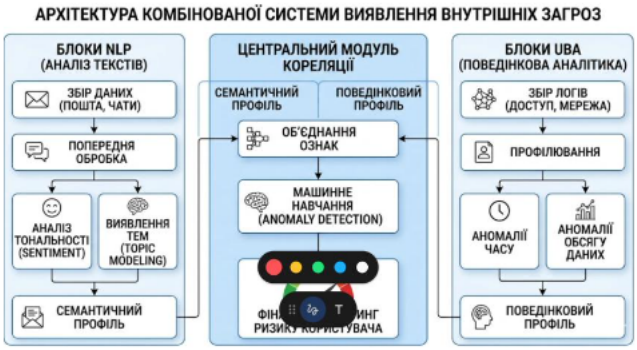

Рисунок 1. Схема архітектури комбінованої системи виявлення внутрішніх загроз на основі NLP та UBA

Важливим елементом є перетворення неструктурованих текстів у графові структури. Використання методів навчання на графах дозволяє візуалізувати не лише

факт обміну повідомленнями, а й сутності (теми), про які йдеться [3]. Це допомагає виявити аномальні інформаційні потоки, наприклад, коли співробітник починає обговорювати нетипові для нього теми або формує нові нехарактерні зв'язки у соціальному графі організації. Схема інтеграції цих методів у єдиний контур моніторингу наведена на рисунку 1.

Порівняльний аналіз демонструє, що такий інтегрований підхід суттєво знижує рівень хибнопозитивних спрацювань. Семантичний контекст слугує підтвердженням або спростуванням аномалії: технічний сплеск активності, який супроводжується використанням специфічних інструментів, описаних у документації розробників, класифікується системою як норма, а не як загроза [4].

Запропонована синергія методів NLP та UBA відкриває новий рівень інтелектуального моніторингу внутрішніх загроз. Поєднання глибокого семантичного розуміння тексту з графовим представленням комунікацій дозволяє не лише виявляти аномалії, а й інтерпретувати їхній контекст, значно зменшуючи кількість хибних тривог. Подальші дослідження доцільно спрямувати на емпіричну верифікацію ефективності системи на реальних корпоративних даних, інтеграцію з іншими джерелами (лог-файли, аудити доступу) та розробку адаптивних моделей, що враховують специфіку різних галузей. Таким чином, комбінований підхід може стати основою для створення проактивних систем кібербезпеки наступного покоління.

Списки використаних джерел

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). – 2019. – P. 4171–4186.
2. Naseer, H., et al. User and Entity Behavior Analytics for Insider Threat Detection: A Review // Computers & Security. – 2024. – Vol. 136. – P. 103–541.
3. Hamilton, W. L. Graph Representation Learning // Synthesis Lectures on Artificial Intelligence and Machine Learning. – Morgan & Claypool Publishers, 2020. – 159 p.
4. Zhu, Y., & Yan, J. Semantic-aware Behavioral Modeling for Enterprise Security // IEEE Transactions on Information Forensics and Security. – 2025. – Vol. 20. – P. 442–457.

Додаток Б

Основний дослідницький конвеєр (section3_analysis_reproducible.py)

```
import os, random
from datetime import datetime, timedelta
import networkx as nx, numpy as np, pandas as pd
from sklearn.cluster import KMeans
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, f1_score
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler

random.seed(7); np.random.seed(7)
OUTDIR = "section3_assets"
os.makedirs(OUTDIR, exist_ok=True)
n_users = 40
users   = [f"u{i:02d}" for i in range(1, n_users+1)]
platforms = ["Telegram","Facebook","Instagram","X"]
topics                                                    =
["освіта","технології","маркетинг","новини","розваги","аналітика",
"спільнота"]

pos_templates = [
    "Дуже корисний пост про {topic}, видно якісну підготовку.",
    "Чудова подача матеріалу, приклад про {topic} виглядає
переконливо.",
    "Подобається цей контент: по {topic} усе пояснено ясно.",
    "Зручний формат і чітке пояснення теми {topic}.",
    "Класна ідея та сильна аргументація для матеріалу про {topic}.",
]
neu_templates = [
    "Сьогодні опубліковано новий допис про {topic}.",
    "Подано короткий огляд без додаткових оцінок щодо {topic}.",
    "Є оновлення по темі {topic}, але без висновків.",
    "Звичайний інформаційний пост із згадкою {topic}.",
]
neg_templates = [
    "Текст виглядає слабко структурованим, особливо в частині
{topic}.",
    "Пост викликає багато питань і не розкриває {topic}.",
    "Незручна подача та мало деталей про {topic}.",
    "Контент не дає практичної користі для теми {topic}.",
]
amb_templates = [
    "Цікавий приклад про {topic}, але подача місцями занадто
проста.",
    "Матеріал про {topic} виглядає непогано, хоча бракує деталей.",
    "Є позитивні моменти у матеріалі про {topic}, але результат
неоднозначний.",
```

```
]

user_activity = pd.DataFrame({
    "user_id": users,
    "activity_level": np.random.choice(["low","medium","high"],
                                       size=n_users,
p=[0.25,0.45,0.30]),
    "community": np.random.choice(["tech","media","edu"],
                                  size=n_users,
p=[0.4,0.35,0.25]),
})

hour_probs                                                          =
np.array([1,1,1,1,1,1,2,3,4,5,5,4,4,4,5,5,6,6,5,4,3,2,2,1],
dtype=float)
hour_probs /= hour_probs.sum()
start_date = datetime(2026, 3, 1)

records = []
for user in users:
    trait                                                           =
user_activity.loc[user_activity.user_id==user,"activity_level"].il
oc[0]
    community                                                       =
user_activity.loc[user_activity.user_id==user,"community"].iloc[0]
    n_posts                                          =            max(3,
{"low":4,"medium":6,"high":9}[trait]+np.random.randint(-1,2))
    for _ in range(n_posts):
        label                                                       =
np.random.choice(["positive","neutral","negative"],p=[0.38,0.34,0.
28])
        template      =      random.choice(amb_templates)      if
np.random.rand()<0.18 \
                  else random.choice({"positive":pos_templates,
                                      "neutral":neu_templates,
"negative":neg_templates}[label])
        topic    = random.choice(topics)
        platform           =           np.random.choice(platforms,
p=[0.30,0.25,0.25,0.20])
        likes_b                                                     =
{"low":np.random.randint(5,22),"medium":np.random.randint(15,55),
                  "high":np.random.randint(40,110)}[trait]
        comm_b                                                      =
{"low":np.random.randint(0,7),"medium":np.random.randint(2,15),
                  "high":np.random.randint(7,30)}[trait]
        shar_b                                                      =
{"low":np.random.randint(0,3),"medium":np.random.randint(1,6),
                  "high":np.random.randint(3,12)}[trait]
        if label=="positive":
            likes=likes_b+np.random.randint(8,32);
comments=comm_b+np.random.randint(1,8)
            shares=shar_b+np.random.randint(0,4)
        elif label=="negative":
```

```
            likes=max(0,likes_b-np.random.randint(0,6));
comments=comm_b+np.random.randint(0,4)
            shares=max(0,shar_b-np.random.randint(0,2))
        else:
            likes=likes_b+np.random.randint(0,8);
comments=comm_b+np.random.randint(0,5)
            shares=shar_b+np.random.randint(0,3)

views=(likes*np.random.randint(5,16)+comments*np.random.randint(6,
20)
               +np.random.randint(30,250))
        hour=int(np.random.choice(np.arange(24),p=hour_probs))

ts=start_date+timedelta(days=np.random.randint(0,45),hours=hour)
        text=template.format(topic=topic)
        if np.random.rand()<0.25: text+=" Це варто переглянути."
        if np.random.rand()<0.08:
            label=np.random.choice([l        for        l        in
["positive","neutral","negative"] if l!=label])

records.append({"post_id":f"p{len(records)+1:04d}","user_id":user,
            "community":community,"platform":platform,
            "timestamp":ts.strftime("%Y-%m-%d
%H:%M:%S"),"hour":hour,
            "text":text,"label":label,"likes":int(likes),

"comments":int(comments),"shares":int(shares),"views":int(views)})

posts=pd.DataFrame(records)
posts["engagement_rate"]=((posts["likes"]+posts["comments"]+posts[
"shares"])/posts["views"])

edges=[]
for _ in range(700):
    src=random.choice(users); dst=random.choice([u for u in users if
u!=src])

itype=np.random.choice(["like","comment","share","mention"],p=[0.5
5,0.25,0.12,0.08])
    weight={"like":1,"comment":2,"share":3,"mention":2}[itype]

csrc=user_activity.loc[user_activity.user_id==src,"community"].ilo
c[0]

cdst=user_activity.loc[user_activity.user_id==dst,"community"].ilo
c[0]
    if csrc==cdst and np.random.rand()<0.6: weight+=1
    ts=start_date+timedelta(days=np.random.randint(0,45),
         hours=int(np.random.choice(np.arange(24),p=hour_probs)))
    edges.append({"source_user":src,"target_user":dst,
                  "interaction_type":itype,"weight":weight,
                  "timestamp":ts.strftime("%Y-%m-%d %H:%M:%S")})
interactions=pd.DataFrame(edges)
```

```
pos_words=["корисний","чудова","подобається","зручний","класна",
           "чітке","якісну","цікавий","непогано"]
neg_words=["слабко","питань","незручна","помилки","не розкриває",
           "слабкий","відкритими","бракує","не дає"]

def lexicon_predict(text):
    t=text.lower(); p=sum(w in t for w in pos_words); n=sum(w in t
for w in neg_words)
    return "positive" if p>n else "negative" if n>p else "neutral"

posts["lexicon_pred"]=posts["text"].apply(lexicon_predict)

X_train,X_test,y_train,y_test=train_test_split(

posts["text"],posts["label"],test_size=0.25,random_state=42,strati
fy=posts["label"])
model=Pipeline([("tfidf",TfidfVectorizer(ngram_range=(1,2),min_df=
1)),
                ("lr",LogisticRegression(max_iter=2000))])
model.fit(X_train,y_train); pred_lr=model.predict(X_test)

user_stats=posts.groupby("user_id").agg(
    posts_count=("post_id","count"), avg_likes=("likes","mean"),
    avg_comments=("comments","mean"), avg_shares=("shares","mean"),
    avg_views=("views","mean"),
avg_engagement=("engagement_rate","mean"),
    positive_share=("label",lambda s:(s=="positive").mean()),
    negative_share=("label",lambda s:(s=="negative").mean()),
).reset_index()
X_scaled=StandardScaler().fit_transform(user_stats.iloc[:,1:])
user_stats["cluster"]=KMeans(n_clusters=3,random_state=42,n_init=1
0).fit_predict(X_scaled)

G=nx.DiGraph()
for _,row in interactions.iterrows():
    if G.has_edge(row["source_user"],row["target_user"]):

G[row["source_user"]][row["target_user"]]["weight"]+=row["weight"]
    else:
G.add_edge(row["source_user"],row["target_user"],weight=row["weigh
t"])

deg_c=nx.degree_centrality(G)
bet_c=nx.betweenness_centrality(G,weight="weight")
eig_c=nx.eigenvector_centrality(G,weight="weight",max_iter=500)

posts.to_csv(f"{OUTDIR}/social_media_posts_dataset.csv",index=Fals
e,encoding="utf-8-sig")
interactions.to_csv(f"{OUTDIR}/social_interactions_dataset.csv",in
dex=False,encoding="utf-8-sig")
user_stats.to_csv(f"{OUTDIR}/user_behavior_dataset.csv",index=Fals
e,encoding="utf-8-sig")
```