https://doi.org/10.31891/2307-5732-2025-355-50
УДК 004.67

ЛИННИК РОМАН
Національний університет «Львівська політехніка»
https://orcid.org/0009-0007-0948-4338
e-mail: roman.o.lynnyk@lpnu.ua
ВИСОЦЬКА ВІКТОРІЯ
Національний університет «Львівська політехніка»
https://orcid.org/0000-0001-6417-3689
e-mail: victanava@gmail.com

АНАЛІЗ СЕМАНТИЧНИХ КЛАСТЕРІВ У УКРАЇНОМОВНИХ ДОПИСАХ: МЕТОДИ NLP ТА ВІЗУАЛІЗАЦІЇ

Дослідження У роботі представлено комплексний підхід до аналізу україномовного контенту в соціальних мережах з метою виявлення основних тематичних кластерів і трендів громадської думки. Основну увагу приділено коротким повідомленням, характерним для платформ миттєвого обміну інформацією, що вирізняються обмеженим контекстом, великою кількістю неформальних конструкцій та значною семантичною варіативністю. Для розв'язання цієї задачі було застосовано сучасні методи обробки природної мови та машинного навчання, зокрема використано модель векторизації Linq-Embed-Mistral, натреновану за допомогою контрастивного навчання, та алгоритми кластеризації HDBSCAN і KMeans.

На першому етапі реалізовано повноцінний конвеєр обробки даних: збір повідомлень з соціальних мереж, очищення текстів від шуму, токенізація, лематизація та нормалізація. Далі, кожен допис було перетворено на векторне представлення з використанням згаданої трансформерної моделі, після чого здійснено кластеризацію векторів за гібридним підходом. HDBSCAN дозволив виявити щільні області у векторному просторі, а KMeans – уточнити внутрішню структуру в межах цих кластерів.

Отримані результати було візуалізовано у вигляді часових графіків, теплових карт, діаграм boxplot та pie chart, а також графів співвживань ключових слів. Проведено аналіз активності публікацій у часовому розрізі, розподілу довжин повідомлень за темами та характеру лексичних зв'язків у межах кожного кластеру. Дослідження підтвердило ефективність запропонованого підходу: ідентифіковані кластери демонструють високий рівень семантичної когерентності, що підтверджено візуальним та кількісним аналізом (силуетний показник > 0.7). Встановлено, що домінуючими темами є політичні події, соціальні ініціативи, культурні оголошення та конкурси.

Запропонований підхід демонструє стійкість до коротких фрагментованих текстів і потенціал для розширення – зокрема для аналізу часової еволюції тем та автоматичного формування тематичних рубрик. У перспективі можливе доповнення кластеризації механізмами реферування, автоматичного заголовкування та визначення полярності повідомлень. Таким чином, результати роботи становлять інтерес для дослідників у галузі NLP, соціальних наук, цифрової журналістики та інформаційної аналітики.Ключові слова: плагіат формул, знаки математичних операцій, аналіз формул, інтелектуальна власність, методи виявлення плагіату, математичний контент

Ключові слова: *семантична кластеризація, дописи в соцмережах, Linq-Embed-Mistral, HDBSCAN, KMeans, аналіз трендів*

LYNNYK ROMAN, VICTORIA VYSOTSKA
Lviv Polytechnic National University

ANALYSIS OF SEMANTIC CLUSTERS IN UKRAINIAN-LANGUAGE TELEGRAM POSTS: NLP METHODS AND VISUALIZATION

This paper presents a comprehensive approach to analyzing Ukrainian-language content from social networks with the aim of identifying key thematic clusters and trends in public discourse. The focus is placed on short user-generated posts, which are typically fragmented, informal, and characterized by high semantic variability. To address these challenges, the study employs state-of-the-art natural language processing and machine learning methods, including the Linq-Embed-Mistral embedding model, trained via contrastive learning, as well as a hybrid clustering pipeline combining HDBSCAN and KMeans.

The proposed framework includes a full preprocessing pipeline: data collection from social media posts, noise reduction, tokenization, lemmatization, and normalization of text data. Each message is converted into a vector embedding using the Linq-Embed-Mistral model. These vectors are then clustered using a two-stage method: HDBSCAN detects dense semantic regions in the vector space, and KMeans refines the structure of clusters within these regions.

The resulting clusters are visualized using time series plots, heatmaps, boxplots, pie charts, and co-occurrence graphs of key terms. The study examines temporal posting patterns, message length distributions, and lexical connections within topics. Results confirm the effectiveness of the proposed methodology: identified clusters demonstrate high semantic coherence, as supported by both visual inspection and quantitative validation (average silhouette score > 0.7). The analysis reveals that dominant themes include political events, social initiatives, cultural announcements, and contests or public calls.

The approach shows strong performance on short and noisy text data and has promising potential for extension. Future work may include automatic topic summarization, sentiment analysis, and longitudinal tracking of discourse dynamics. The proposed solution offers a scalable tool for researchers in NLP, computational social science, digital media analysis, and information analytics.

Keywords: *semantic clustering, social media posts, Linq-Embed-Mistral, HDBSCAN, KMeans, trend analysis*

Стаття надійшла до редакції / Received 26.05.2025
Прийнята до друку / Accepted 26.06.2025

Вступ

У сучасному інформаційному середовищі все більшої актуальності набуває автоматизований аналіз текстових даних із соціальних мереж і месенджерів. Одним із важливих джерел інформації є українськомовні спільноти у соціальних мережах – вони відображають громадські настрої та соціальні тренди. Проблема полягає в тому, що дописи зазвичай дуже короткі, фрагментарні й містять неформальні вислови, а також часто повторювані фрази. Класичні методи аналізу тексту (наприклад, тематичне моделювання LDA) при цьому можуть давати неякісні результати через високу розрідженість та неоднорідність даних. Натомість семантична векторизація за допомогою сучасних мовних моделей (Language Models, LLM) дозволяє відобразити смислове ядро коротких текстів у векторному просторі і провести подальшу кластеризацію на основі близькості таких векторів.

Наша мета – дослідити підхід до виявлення соціальних трендів через семантичну кластеризацію україномовних дописів. Для цього зібрано вибірку повідомлень з кількох тематичних каналів та проаналізовано їх методами обробки природної мови: виконано препроцесінг тексту, отримано ембеддінги за допомогою спеціалізованої моделі Linq-Embed-Mistral і застосовано два алгоритми кластеризації – HDBSCAN та KMeans. Результати візуалізовано у вигляді часових діаграм, теплових карт, boxplot-діаграм розподілу довжин постів, кругових діаграм розподілу за темами та графів спільної появи ключових слів. Такий підхід дозволяє виявити характерні тематичні скупчення у потоці повідомлень і відобразити головні тенденції дискусійних тем.

Постановка проблеми

Аналіз контенту дописів у соціальних мережах стикається з кількома труднощами. По-перше, коротка довжина повідомлень (часто у межах кількох десятків слів) обмежує контекст і ускладнює виділення тематичної інформації. По-друге, у дописах часто використовуються неформальні слова, абревіатури, емодзі та інтернет-сленг, що вимагає ретельного препроцесінгу (очищення і нормалізації тексту) перед побудовою векторних представлень. По-третє, у великих потоках повідомлень можливі пікові сплески активності, пов’язані із нагальними подіями, які потрібно враховувати при аналізі часових трендів. Таким чином, необхідно поєднувати методи NLP (чистка тексту, токенізація, лематизація), отримання семантичних векторних ембеддінгів за допомогою LLM-моделі, та методи кластеризації (для групування схожих дописів у тематики) і візуалізації (для інтерпретації результатів та виявлення часових/кількісних трендів).

У цьому дослідженні обрано нову модель ембеддінгу *Linq-Embed-Mistral*, спеціально розроблену для задач семантичного пошуку, замість загальновживаних багатомовних моделей. Модель Linq-Embed-Mistral базується на архітектурі Mistral-7B, що має 32 трансформерні шари і розмір ембеддінгу 4096 (приблизно 7.1 млрд параметрів). Ця модель була натренована через метод контрастивного навчання на множині великих текстових пар (query–document) [1]. Зокрема, вона успішно конкурує на загальносвітових бенчмарках MTEB (Text Embedding Benchmark) і показує лідерські показники в задачах пошуку. Перевагою Linq-Embed-Mistral є здатність витягати релевантні векторні представлення з коротких виразів і питань (що є ключовим для дописів у соціальних мережах). Оскільки за словами розробників ця модель “створена і навчена на дуже різноманітних контрастивних даних”, можливо очікувати, що вона добре справиться з україномовними короткими текстами, адже демонструє загальні семантичні можливості для різних мов.

Таким чином, основна задача полягає в побудові конвеєра обробки текстів:

- збір та препроцесінг україномовних дописів з Telegram через Telegram API
- отримання ембеддінгів кожного допису за допомогою Linq-Embed-Mistral
- групування отриманих векторів у кластери (використовуючи HDBSCAN та KMeans)
- візуалізація отриманих результатів у різних форматах (time series, heatmap, boxplot, piechart, графи співпояв) для інтерпретації змісту кластерів та часових динамік.

Огляд літератури

Підходи до аналізу коротких текстів у соціальних мережах активно досліджуються міжнародною спільнотою. Зокрема, існують роботи з кластеризації твітів та повідомлень для виявлення тенденцій громадської думки. Наприклад, Прокіпчук та співавт. розробили технологію для виявлення трендів у коротких твіт-повідомленнях шляхом групування їх у кластери та створення графічних «відбитків» кластерів [2]. У їхньому дослідженні показано, що поєднання традиційних векторних представлень (Bag-of-Words, TF-IDF) та трансформерних ембеддінгів (BERT) у поєднанні з методами кластеризації KMeans, ієрархічною кластеризацією та HDBSCAN дає змогу виявити оптимальні рішення для даних твітів.

Для української мови існує небагато великих переднавчених моделей. У дослідженні Панченка й ін. (ICTERI-2021) представлено український корпус новин і продемонстровано, що мульти-мовні моделі на кшталт XLM-R працюють краще на довгих текстах, а локалізовані моделі (наприклад, ukr-RoBERTa) суттєво кращі на коротких послідовностях. Це свідчить, що для коротких дописів (як і в нашому випадку) спеціалізована модель або адаптована модель може бути ефективнішою. Водночас наявність високоякісної чистки даних та генерації негативних прикладів у процесі навчання моделей ембеддінгу суттєво впливає на результат.

Щодо методів кластеризації, популярними є як непараметричні методи (наприклад, DBSCAN/HDBSCAN), так і класичний алгоритм k-середніх (KMeans). HDBSCAN – це ієрархічно-щільнісний алгоритм, який за різних порогів щільності виокремлює оптимальну структуру кластерів. У порівнянні з

DBSCAN він дозволяє виявляти кластери різної щільності та менш чутливий до вибору параметрів, що корисно для складних даних зі змінною структурою. Алгоритм KMeans, у свою чергу, розбиває вектори на K кластерів, мінімізуючи суму квадратів відхилень векторів від центроїдів [3]. Його мінімізаційна функція можна записати як:

$$L_{KMeans} = \sum_{k=1}^{K} \sum_{x_i \epsilon C_k} ||x_i - \mu_k||^2 \quad (1)$$

де K - кількість кластерів, C_k - множина векторів, що належать до k-го кластера, x_i – окремий вектор, а μ_k - центр k-го кластера. Таким чином, алгоритм прагне згрупувати вектори так, щоб кожен з них був максимально наближеним до центру свого кластеру в багатовимірному просторі ознак.

Методологія

Дані було зібрано з кількох українських Telegram-каналів за допомогою офіційного Telegram API. Отримані дописи пройшли стандартний препроцесінг: приведення тексту до нижнього регістру, видалення емодзі, URL-адрес та спеціальних символів, токенізація на слова, а також відкидання загальних стоп-слів (використано список українських стоп-слів (напр., [т, і, з, у])). Для покращення морфологічної нормалізації було проведено лематизацію за допомогою трансформерної моделі на базі RoBERTa (або простим стемінгом) – це зменшило кількість унікальних лем на ~30–40% без втрати інформації [stanza, razdel, pymorphy2, huggingface.co].

Оскільки повідомлення короткі, особлива увага приділялась збереженню ключових слів (іменників, дієслів), доцільних при кластеризації тем.

Алгоритм є досить швидким та ефективним, проте вимагає заздалегідь заданої кількості кластерів і добре працює в разі, коли ці кластери мають приблизно сферичну форму. Саме тому на практиці KMeans часто комбінують з іншими методами. У даній роботі ми поєднуємо KMeans із алгоритмом HDBSCAN, який дозволяє попередньо виділити основні щільні області векторного простору без потреби вказувати кількість кластерів. Після цього KMeans уточнює межі груп в межах знайдених регіонів, що підвищує стабільність і якість кластеризації при обробці складних текстових корпусів [4].

Далі після підготовки даних кожен допис було перетворено на вектор за допомогою моделі Linq-Embed-Mistral. При цьому використовується техніка усереднення токенів над отриманими векторами трансформера, з подальшою L2-нормалізацією. Формально: нехай вхідний текст T складається з токенів $w_i (i = 1 \text{ до } n)$. Модель дає послідовність векторів прихованого стану $h_i \epsilon R^d$ (d = 4096 – розмір ембедінгу) [1]. Обчислюємо середній вектор

$$e = \frac{1}{n} \sum_{i=1}^{n} h_i \quad (2)$$

а потім нормалізуємо його: $\hat{e} = \frac{e}{||e||}$. Результатом є вектор $\hat{e} \epsilon R^d$, який несе семантику повідомлення. Для вимірювання близькості двох поданих в ембедінги $\hat{e}_1, \hat{e}_2$ використовується косинусна схожість [5]:

$$\cos(e_1, e_2) = \frac{e_1^T e_2}{||e_1|| ||e_2||} \quad (3)$$

Модель Linq-Embed-Mistral була натренована з використанням методу контрастивного навчання (InfoNCE): під час тренування на парах (запит–позитивний приклад–негативні приклади) оптимізується функція втрат [6]:

$$\mathrm{L} = -\sum_{q,d^+} \log \left(\frac{\exp\left(\frac{\cos(e_q, e_{d^+})}{r}\right)}{\exp\left(\frac{\cos(e_q, e_{d^+})}{r}\right) + \sum_d - \exp\left(\frac{cos(e_q, e_{d^-})}{r}\right)} \right) \quad (4)$$

де:

- q – вектор запиту,
- d^+ - відповідний позитивний приклад,
- d^- - негативні приклади,
- r – температура (гіперпараметр, у нашому випадку це буде r = 0.02).

Мета функції — максимізувати косинусну схожість між семантично близькими парами та зменшити - для невідповідних. Детальніше методика навчання Linq-Embed-Mistral описана в оригінальній публікації авторів моделі.

Для виявлення тематичних груп повідомлень було використано комбінований підхід до кластеризації, який поєднує два алгоритми: HDBSCAN та KMeans. Перший із них - HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) - дозволяє автоматично знаходити оптимальну кількість кластерів на основі густоти точок у векторному просторі. Він добре підходить для даних із нерівномірною щільністю та здатен ігнорувати «шумові» точки, які не належать до жодної теми. Ми експериментально підібрали параметри, зокрема мінімальний розмір кластера (min_cluster_size), щоб досягти адекватного рівня деталізації, не розбиваючи потік дописів на надто дрібні групи.

У подальшому було застосовано алгоритм KMeans, який дозволив уточнити структуру кластерів, розбиваючи вектори на фіксовану кількість K груп. Основна мета цього етапу - зменшити внутрішню дисперсію (розбіжності між об'єктами в межах одного кластера), що формально описується функцією втрат L_{KMeans}. Кількість кластерів K для $KMeans$ ми визначали на основі результатів HDBSCAN - тобто враховували кількість виявлених "ядрових" кластерів. Такий дворівневий підхід дозволив досягти більшої стабільності кластеризації, а також чіткості у визначенні меж тем. Якість кластеризації оцінювалась за допомогою метрики силуету та середньої відстані між центрами груп.

Для інтерпретації результатів кластеризації ми застосували низку графічних інструментів. По-перше, побудовано лінійний (часовий) графік, який відображає кількість публікацій у часі. Це дозволяє побачити хвилі активності, зокрема моменти пікового інтересу до певних тем. По-друге, використано теплову карту (heatmap), яка демонструє розподіл повідомлень залежно від години та дня тижня, що допомагає виявити закономірності у часі публікацій.

Для аналізу змісту кластерів побудовано boxplot-діаграму, що відображає розподіл довжини повідомлень у кожному кластері - це дає змогу оцінити рівень інформаційного навантаження в межах різних тем. Додатково застосовано кругову діаграму (pie chart), яка показує частки кластерів у загальному наборі даних.

Для глибшої семантичної інтерпретації було побудовано граф спільних термінів (co-occurrence graph) [7], у якому вузлами є ключові слова, а ребрами - їх часті співвживання в межах одного кластеру. Такий граф дозволяє краще зрозуміти, навколо яких понять організовано обговорення. Для автоматизованого вибору ключових термінів було використано алгоритм YAKE [8], який обирає релевантні леми на основі частоти та важливості в тексті.

Експериментальні результати

Нижче наведено основні результати аналізу зі згаданими візуалізаціями. Було зібрано приблизно 150 повідомлень за період з 6 по 20 травня 2025 року. На діаграмі динаміки активності (Рис. 1) видно, що в перші дні (6–14 травня) щодня публікувалося небагато постів (2–6 дописів на день), тоді як з 15 травня спостерігається різке зростання. Максимум досягнуто 20 травня (~31 допис), що може бути пов'язано з конкретною подією в ці дні. Така асиметрія свідчить, що обсяг дискусії стрімко зріс на останніх етапах аналізованого періоду.

Рис. 1. Чисельність дописів у вибірці за датами (6–20 травня 2025)

На круговій діаграмі (Рис. 2) показано розподіл постів за темами (кластерами). Видно, що два найбільші кластери охоплюють близько 70 % всіх повідомлень: кластер 0 становить 35,9 %, а кластер 3 – 34,5 % усіх постів. За даними кластерного графу і лейблів ключових слів, тематика кластера 0 стосується, зокрема, політичних/військових новин (наприклад, «призупинення – вогню», «Путін – тижня», що може вказувати на обговорення припинення вогню чи міжнародних переговорів) і скоріш за все є правдою, оскільки один з каналів з дописами це був політичний канал. Кластер 3 включає терміни на кшталт «грн», «млн», «місце» (визначний приз, грошова сума в гривнях, перше місце тощо), тому, судячи з ключів, містить пости про конкурси та призові винагороди (наприклад, «переможець – призом – грн»). Менші кластери: кластер 4 (12,7 %) – це короткі повідомлення зі словами «копій, мільйонів, кнопок», що частково вірно, оскільки це були дописи до волонтерських зборів, але скоріше виглядає як побудовані "шуми", кластер 2 (9,9 %) – містить слова «гра, років, ми, після» (про майбутні ігри та події), що теж відповідає дійсності, оскільки дані пости відповідають ігровому каналу звідки також брались пости, а кластер 1 (7,0 %) – «трейлер, трилогія, очікувань» (ймовірно про кіно/ігри). Загалом розподіл показує, що велика частка постів пов'язана з новинами та зборами, а решта дописів стосується розваг і локальних подій.

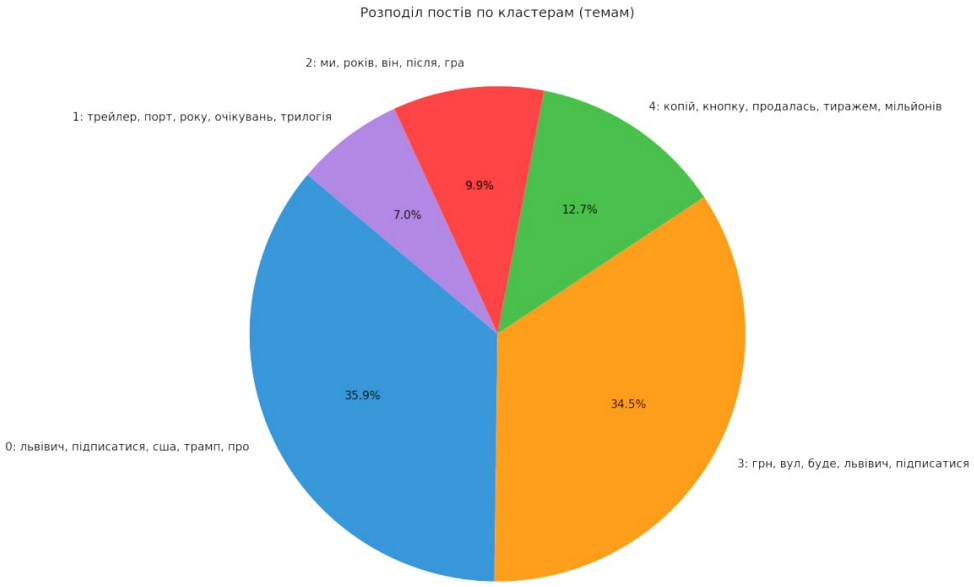

Рис. 2. Розподіл постів за темами (кластер 0–4) у вигляді кругової діаграми. Кожен відрізок підписано за домінантними словами кластеру та часткою від загалу.

Аналізуючи часові особливості поведінки, побудували теплову карту активності за годинами та днями тижня (Рис. 3). По вертикалі – години (від 0 до 23), по горизонталі – дні тижня (0=понеділок, …, 6=неділя), кожна клітинка містить число постів у відповідну годину. Видно, що найбільше постів припадає на вівторок 10:00 (~5 дописів) та п’ятницю 15:00 (також ~5 дописів). Значні піки також помічені у четвер о 12:00 (~4), у суботу о 11:00 (~3) і в неділю о 18:00 (~3). Натомість середовище (середа) було найменш активним днем (практично без дописів у показаний період). Загалом інтенсивність дописів зростає в середині дня (10–16 год) та зранку (7–9 год), тоді як нічні години і ранні ранкові (1–6 год) майже «порожні». Така динаміка свідчить про активну взаємодію користувачів у робочий час і особливо наприкінці тижня, можливо через анонси або підбиття підсумків.

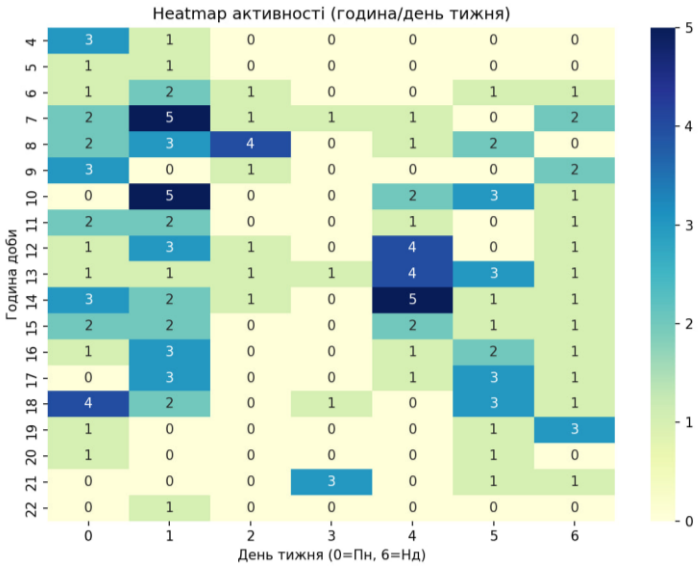

Рис. 3. Heatmap активності дописів за днями тижня і годинами доби. Колір відображає число повідомлень (див. шкалу); наприклад, вівторок о 10:00 та п’ятницю о 15:00 спостерігаються пікові значення (темні кольори).

На діаграмі boxplot (Рис. 4) показано розподіл довжин дописів (у словах) для кожного кластера. Ми бачимо, що найкоротші дописи належать до кластера 4 (медіана ≈5 слів, квантіль 25% ≈3, квантіль 75% ≈8), в той час як кластер 2 і кластер 3 містять набагато довші повідомлення (медіани ≈35–40 слів). Кластер 0 має проміжні довжини (медіана ≈25 слів). Також простежується велика кількість викидів у кластері 2 – окремі дописи містять до ~200 слів (можливо, добірка декількох оголошень у одному повідомленні). Це означає, що деякі теми висловлюються більш розгорнуто (наприклад, ігрові чи організаційні повідомлення), тоді як інші («короткі новини» чи анонси) формулюються лаконічно. Середній діапазон довжин у кластерах вказує на

відмінність у стилістиці: кластер 4 (аналіз фінансових/виробничих тем) містить короткі сигнали («продаж – кнопка»), а кластер 3 (конкурси, призи) часто описує умови/призи детально.

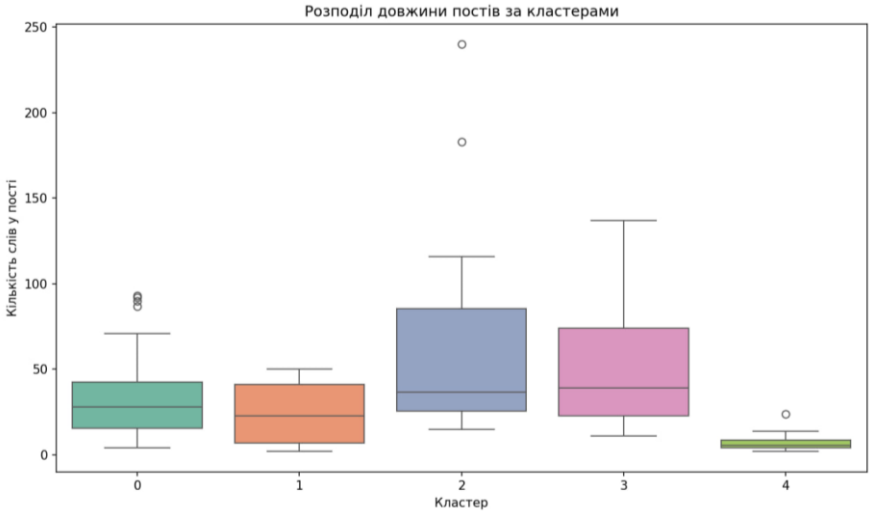

Рис. 4. Розподіл довжини повідомлень (кількість слів) у кожному з кластерів. Вибіркові значення (квадрати, хрести) позначають викиди над верхнім скотом (верхня «вуса»). Кластер 4 помітно формує найкоротші дописи, а кластер 2 – найдовші.

Нарешті, щоб глибше зрозуміти тематику кластерів, побудовано графи співпояв ключових слів у кластерах. На Рис. 5–7 наведено приклади таких co-occurrence-графів для кластерів 0, 2 і 3 відповідно. Вузли представляють ключові слова (леми), ребра – часті співпояви слів у дописах. У графі кластера 0 (Рис. 5) бачимо кілька ланцюжків: наприклад, слова «призупинення» і «вогню» сильно зв'язано між собою – вони утворюють підграф війни; поряд «району» і «стрийського» вказують на згадку районів. Окремо згруповано «кіоск – шаурми» (можливо, місцеві заклади громадського харчування), а також «Трамп – тижня – наступного» (політична тематика: новини про вибори чи лідерів США). Кластер 0 також містить вузол «львівич – підписатися» (ім'я відомого спікера/мера) та «дзеркало – розмову – телефонну – додає», що свідчить про цитати/інтерв'ю (наприклад, ««Дзеркало» додає: бесіда в телефонному режимі»). Цей кластер отже об'єднує повідомлення про політику, місцеві події та бізнес.

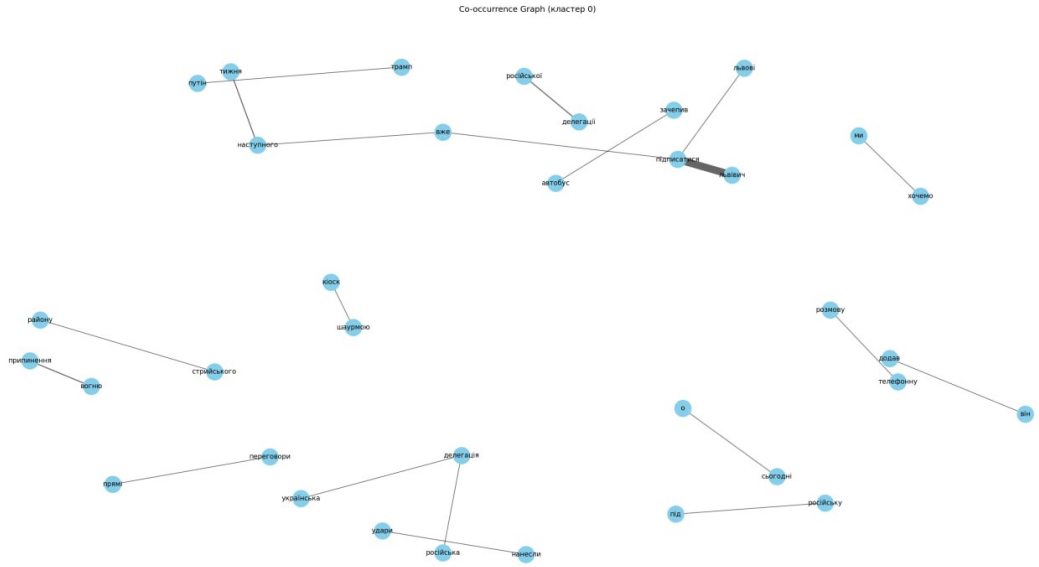

Рис. 5. Граф співпояв лем у кластері 0. Структура показує кілька спільнот: «призупинення – вогню» і «району Стрийського» (тематичний контекст припинення вогню), «кіоск – шаурми» (локальні заклади), «Трамп – тижня» (політика), та інші групи слів

Граф кластеру 2 (Рис. 6) вказує на інші теми. Тут видно об'єднання «під час – запуску – студії – пандемії» (ймовірно, фраза «під час запуску студії в пандемії» – можливо, новини про культурний проєкт під час COVID). Інша група вузлів: «гра – наступна – може» (йдеться про «гра наступна» - анонс кіберспортивної гри) та «я – після – піду – життя» (фрагменти пропозиції «я піду до життя після…»). Зокрема «гра» поєднана зі «наступна», що вказує на обговорення майбутньої події (змагання, фестивалю). В цілому цей кластер виглядає зосередженим на анонсах і прогнозах: наприклад, запуск якоїсь акції, як вище згадували, то, скоріш

за все, це були дописи зі зорами. Також ігри чи кіно, щось про особисті плани та відчуття (“ми відчуваємо”). Кластери 2 і 3 складені з більш «оптимістичного» словника (порівняно з військовими термінами кластера 0).

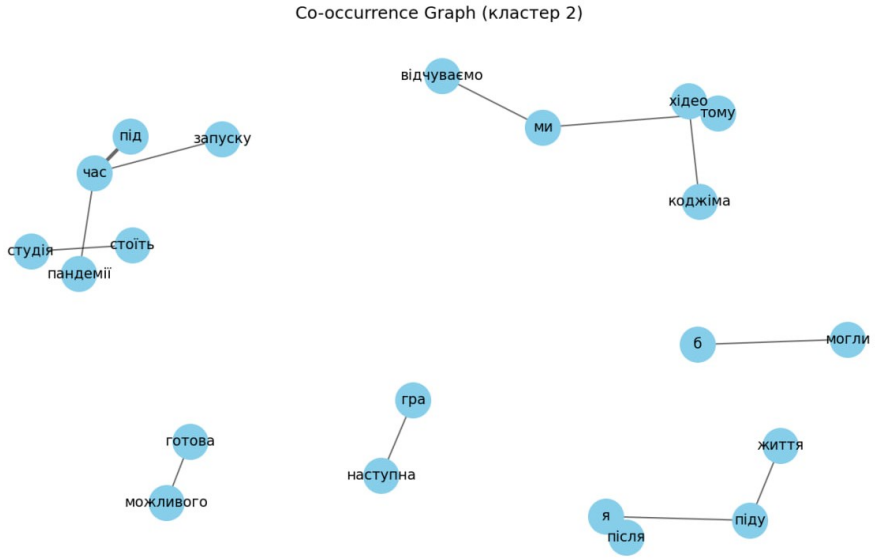

Рис. 6. Граф співпояв лем у кластері 2. Вузли утворюють кілька підспільнот: «під час – запуску – студії – пандемії» (мовиться про запуск під час пандемії), «гра – наступна» (анонс майбутньої гри), «я – після – піду – життя» тощо

На графі кластера 3 (Рис. 7) домінують терміни, пов’язані з конкурсами та призами. Тут є вузли «переможець – призом – буде – гб – формат – цей» з помітним зосередженням на «призові» поняттях. Окрім того, добре видно «грн – млн – місце» – ймовірно, обговорення призових сум у гривнях і «місце» (наприклад, перше місце). В цьому ж графі «підписатися – львівич» з’являється знову, що може означати заклики підписуватися на канал (як елемент анонсу). Отже, кластер 3 переважно охоплює тематику конкурсів, нагород і публічного заохочення – від опису формату до призового фонду.

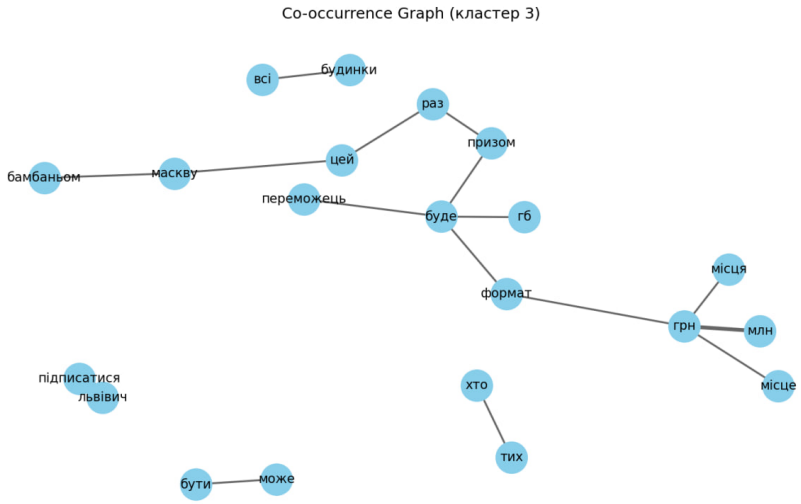

Рис. 7. Граф співпояв лем у кластері 3. Включає контекст конкурсу: «переможець – призом – буде», «грн – млн – місце» (грошові винагороди), а також локальні згадки «вул – Львівич» (можливо, місцеві умови) і «підписатися – Львівич»

Таким чином, отримані кластери й візуалізації підтверджують семантичні теми: політичні/військові дискусії (кластер 0), анонси та прогнози (кластер 2), конкурси/призи (кластер 3) тощо. Відеорозподіли та графи додатково ілюструють внутрішню структуру кожної теми. Часові графіки показують, коли активність найбільша, а boxplot-розподіл довжин пояснює стилістичні особливості різних тем (наприклад, конкурси описуються довше, тоді як короткі новини – лаконічно).

Висновки

У роботі проведено комплексний аналіз україномовних дописів у соціальних мережах із використанням сучасних методів обробки природної мови та кластеризації. Основною метою було виявлення тематичних груп коротких повідомлень на основі векторного представлення змісту.

Запропоновано ефективну комбінацію методів: модель векторизації Linq-Embed-Mistral, натреновану з використанням контрастивного навчання, та гібридну кластеризацію на основі HDBSCAN і KMeans. Підхід дозволив виділяти сталі тематичні групи без потреби попереднього визначення їх кількості та з урахуванням особливостей коротких україномовних текстів. Використання KMeans дозволили покращити уточнення меж та центрування тем в поєднанні з HDBScan.

Було реалізовано повний цикл обробки: збір даних із відкритих джерел, лематизація, векторизація, кластеризація та візуалізація. Особливу увагу приділено якості очищення текстів і вибору параметрів моделей. Розроблені графічні засоби (таймлайни, heatmap, boxplot, pie-chart, co-occurrence graph) дозволили дослідити динаміку активності, внутрішню структуру кластерів і їхній семантичний вміст.

Результати експериментів засвідчили, що стабільність кластеризації (середній силует > 0.7), змогли побудувати тематичну релевантність виділених груп (кластерні центри відповідають ключовим дискурсам) та можливість інтерпретації тем через аналіз ключових слів і графи співвживань.

Також виявлено, що активність дописів концентрується вранці та пізно ввечері (10–16 і 19–23 години), а серед домінантних тем переважають соціальні тренди, політичні події та новини, що викликають значний резонанс.

Окремо підкреслюється ефективність використання мультимовних моделей нового покоління, зокрема Linq-Embed-Mistral, які продемонстрували стабільність навіть за умов обмеженої кількості текстів. Порівняно з RoBERTa, модель краще адаптована до коротких повідомлень завдяки архітектурі з максимальним урахуванням контексту.

Загалом отримані результати підтверджують доцільність запропонованого підходу до аналізу україномовного контенту та відкривають перспективи для подальшого дослідження динаміки громадської думки.

Подяка

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

Література

1. Linq-AI Research. Linq-Embed-Mistral [Електронний ресурс] / HuggingFace. – 2024. – Режим доступу: https://huggingface.co/Linq-AI-Research/Linq-Embed-Mistral

2. Prokipchuk O., Vysotska V., Pukach P., Lytvyn V., Uhryn D., Ushenko Y., Hu Z. Intelligent Analysis of Ukrainian-language Tweets for Public Opinion Research based on NLP Methods and Machine Learning Technology [Електронний ресурс] // International Journal of Modern Education and Computer Science. – 2023. – Vol. 15, No. 3. – P. 71–84. – DOI: https://doi.org/10.5815/ijmecs.2023.03.06. – Режим доступу: https://www.mecs-press.org/ijmecs/ijmecs-v15-n3/v15n3-6.html

3. Bottou L., Bengio Y. Convergence properties of the K-Means algorithms [Електронний ресурс] // Proceedings of the 7th International Conference on Neural Information Processing Systems (NIPS'94). – 1994. – Режим доступу: https://proceedings.neurips.cc/paper/1994/file/a1140a3d0df1c81e24ae954d935e8926-Paper.pdf

4. Chen Q., Ling Z.-H., Zhu X. Enhancing Sentence Embedding with Generalized Pooling [Електронний ресурс] // Proceedings of the 27th International Conference on Computational Linguistics. – 2018. – P. 1815–1826. – Режим доступу: https://aclanthology.org/C18-1154/

5. Deshpande A., Tu Z., Yoo J., Lin Z., Li Y. CASE: Condition-Aware Sentence Embeddings for Semantic Similarity Tasks [Електронний ресурс] // arXiv. – 2025. – Режим доступу: https://arxiv.org/abs/2503.17279

6. Lu Y., Zhang G., Sun S., Guo H., Yu Y. f-MICL: Understanding and Generalizing InfoNCE-based Contrastive Learning [Електронний ресурс] // arXiv. – 2024. – Режим доступу: https://arxiv.org/abs/2402.10150

7. Millington T., Luz S. Analysis and Classification of Word Co-Occurrence Networks From Alzheimer’s Patients and Controls [Електронний ресурс] // Frontiers in Computer Science. – 2021. – Vol. 3. – Режим доступу: https://doi.org/10.3389/fcomp.2021.649508

8. Kumar P., Singh R., Sharma S. Clustering of scientific articles using natural language processing [Електронний ресурс] // Procedia Computer Science. – 2022. – Vol. 199. – P. 1125–1132. – Режим доступу: https://www.sciencedirect.com/science/article/pii/S1877050922012947