УДК 004.67 DOi 10.30837/ bi.2025.1(102).04

Р. О. Линник[1], В. А. Висоцька[2]

[1]НУ «Львівська політехніка», м. Львів, Україна,
roman.o.lynnyk@lpnu.ua, ORCID iD: 0009-0007-0948-4338

[2]НУ «Львівська політехніка», м. Львів, Україна,
victoria.a.vysotska@lpnu.ua, ORCID iD: 0000-0001-6417-3689

АНАЛІЗ ПРОБЛЕМАТИКИ ВИЯВЛЕННЯ ТРЕНДІВ ГРОМАДСЬКОЇ ДУМКИ В УКРАЇНОМОВНИХ ДОПИСАХ ЗАСОБАМИ КЛАСТЕРИЗАЦІЇ ТА НЕЙРОННИХ МЕРЕЖ

У даному дослідженні застосовано гібридний підхід до кластеризації, що поєднує алгоритми DBSCAN та K-means для аналізу векторизованих україномовних дописів у соціальних мережах з метою виявлення трендів громадської думки. Методологія базується на багатомовній моделі векторизації тексту, побудованій на основі нейронної мережі, яка дозволяє ефективно відображати семантичний зміст повідомлень. Експерименти, проведені на корпусі з 90 українськомовних дописів (зібраних у період березень–травень 2025 року), дозволили виокремити шість основних тематичних кластерів, що відображають ключові напрями обговорень. Результати дослідження підтверджують ефективність запропонованого методу для аналізу трендів у соціальних медіа та його практичну цінність для моніторингу громадської думки.

ТРЕНДИ ГРОМАДСЬКОЇ ДУМКИ, СОЦІАЛЬНІ МЕРЕЖІ, ОБРОБКА ПРИРОДНЬОЇ МОВИ, ВЕКТОРИЗАЦІЯ ТЕКСТУ, КЛАСТЕРИЗАЦІЯ, НЕЙРОННІ МЕРЕЖІ

R.O. Lynnyk, V.A. Vysotska. Analysis of the Problem of Detecting Public Opinion Trends in Ukrainian-Language Posts by Means of Clustering and Neural Networks. In this study, we employ a hybrid clustering approach combining DBSCAN and K-means algorithms on vectorized Ukrainian social media posts to detect public opinion trends. The methodology uses a neural network–based multilingual text vectorization model to capture semantic content effectively. Experiments on a dataset of 90 Ukrainian posts (collected March–May 2025) identified six principal thematic clusters representing key discussion topics. The results demonstrate the effectiveness of the proposed method for social media trend analysis and its potential application in public opinion monitoring.

PUBLIC OPINION TRENDS, SOCIAL NETWORKS, NATURAL LANGUAGE PROCESSING, TEXT VECTORIZATION, CLUSTERING, NEURAL NETWORKS

Вступ

У сучасних умовах швидкого росту соціальних медіа як джерела інформаційного впливу значно впливають на формування загальносуспільною думки, роблячи аналіз украйномовних даних у мережах надзвичайно актуальним завданням. Платформи, на кшталт Telegram, стають все більш популярними як осередки для обговорення суспільно важливих питань. Разом з цим кількість неструктурованих текстових повідомлень стрімко зростає експоненційно, що робить їх ручне опрацювання складним завданням. У такому контексті сучасні методи обробки природного мовлення – зокрема трансформерні нейромережі – та методи кластеризації виявляють приховані закономірності в обсяжних текстових наборах і допомагають знаходити новий інсайти. У зв'язку з цими умовами використовуваний комплексний аналітичний підхід: застосовувана багатомовна модель векторизації тексту в посполитості з гібридною кластеризацією (по комбінацію DBSCAN і K-means) для виявлення тематичних тенденцій у повідомленнях украинською мовою та монтування що наводять щих динаміку в часопростор.

1. Постановка проблеми

Соціальні мережі, зокрема Telegram, стали ключовим каналом поширення інформації та обговорення публічних питань у суспільстві. Натомість великий обсяг неструктурованого контенту висуває вимогу швидких та надійних алгоритмічних рішень для аналізу таких даних. Особливу роль відіграє специфіка української мови та невеликі обсяги кожного допису, що потребує адаптації існуючих методів обробки текстів.

Мета цього дослідження – розробити і впровадити цілісну методику аналізу трендів громадської думки в україномовних дописах соцмереж на основі сучасних методів машинного навчання (кластеризації та нейромереж). Для досягнення цієї мети вирішувалися такі ключові завдання:

Збір та попередня обробка даних: формалізувати методологію збору та очищення україномовних повідомлень із Telegram-каналів.

Векторизація текстів: вибір та налаштування багатомовної моделі для якісного семантичного представлення тексту (зокрема трансформерної моделі multilingual-e5-large-instruct).

Гібридна кластеризація: створення комбінованого алгоритму кластеризації на основі DBSCAN та K-means для виявлення як великих, так і малих тематичних груп дописів.

Візуалізація результатів: розробка засобів візуального подання вихідних даних та результатів кластеризації (графіки активності, хмари слів, PCA-проекції, теплові карти).

Цей набір задач узгоджується з сучасним попитом на інструменти для оперативного моніторингу громадської думки та аналізу соціальних трендів.

2. Сучасні методи аналізу текстів у соціальних медіа

Аналіз соціальних медіа є міждисциплінарною галуззю, що поєднує методи комп'ютерних наук, лінгвістики та соціальних наук для вивчення особливостей та процесів, пов'язаних з обробкою великих обсягів неструктурованих даних, генерованих платформами на кшталт Twitter і Telegram [1]. Специфіка цих даних полягає у їх «об'ємі, різноманітності та швидкості», що вимагає структурованого підходу до їх збору, очищення, обробки та візуалізації.

Під час дослідження Капуккінеллі та співавторів було виявлено, що соціальні медіа за останні два десятиліття еволюціонували від простих платформ для спілкування до складних інформаційних екосистем, які суттєво впливають на формування громадської думки [1]. У своєму огляді 132 публікацій автори виділили ключові напрямки досліджень, зокрема аналіз користувацького контенту, поведінкові аспекти використання соціальних мереж та інтеграцію соціальних медіа для маркетингових і організаційних цілей.

Математично процес аналізу текстових даних можна представити як послідовність трансформацій:

$$T : D \to V \to C \to I \tag{1}$$

де D – вихідний набір текстових документів, V – векторне представлення документів, C – кластери або класи документів, I – інтерпретація результатів [2].

3. Векторизація тексту та сучасні трансформерні моделі

Ключовим етапом аналізу текстових даних є їх векторизація – перетворення тексту у чисельні вектори, що зберігають семантичне значення і можуть бути використані в алгоритмах машинного навчання [3]. Традиційні методи, такі як TF-IDF (Term Frequency-Inverse Document Frequency), мають обмеження у здатності враховувати контекст та відносини між словами [2].

Модель «multilingual-e5-large-instruct», що використовується у нашому дослідженні, належить до четвертого покоління універсальних текстових ембедингів, спрямованих на створення єдиної моделі, здатної узагальнювати різноманітні завдання та домени [4]. Її архітектура базується на трансформерній структурі з двонаправленим енкодером, що ефективно захоплює контекстну інформацію [3].

Як зазначають Вонг та співавтори, модель навчалася з використанням масштабного контрастивного підходу на наборі даних, що містить близько 1 мільярда багатомовних текстових пар, включаючи реальні та синтетичні дані, згенеровані великими мовними моделями (LLM) [5]. Архітектура моделі включає 24 шари з розміром ембедингу 1024 та підтримує 100 мов, що робить її особливо цінною для аналізу багатомовного контенту [5]. Математично процес векторизації тексту за допомогою цієї моделі можна описати як:

$$e_i = E5(x_i) \in R^d \tag{2}$$

де x_i – вхідний текстовий документ, e_i – його векторне представлення розмірності d_i, а $E5$ – функція трансформації, що реалізується моделлю [6].

4. Методи кластеризації текстових даних

Кластеризація є важливим методом організації та аналізу великих обсягів даних, що дозволяє виявляти приховані структури та закономірності [7]. У контексті аналізу соціальних медіа кластеризація використовується для групування подібних повідомлень, визначення тем та трендів.

Серед популярних алгоритмів кластеризації для текстових даних можна виділити:

1. "K-means" – один з найпоширеніших алгоритмів, що мінімізує суму квадратів відстаней від точок до центрів кластерів. Математично цільова функція K-means визначається як:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 . \tag{3}$$

де c_j – центр кластера j, а x_i^j – точка даних, що належать кластеру [8].

2. "DBSCAN» (Density-Based Spatial Clustering of Applications with Noise) – алгоритм, що базується на щільності точок та ефективно ідентифікує кластери довільної форми й відокремлює шум [9].

3. «Ієрархічна кластеризація» – метод, що будує ієрархію кластерів шляхом послідовного об'єднання або розділення груп [10].

Недавні дослідження демонструють переваги гібридних підходів, що поєднують різні алгоритми кластеризації. Зокрема, робота Джаін і співавторів показує, що гібридний підхід на основі K-means та алгоритму оптимізації мурашиних колоній (Ant Lion Optimization) забезпечує кращі результати за такими метриками, як внутрішньокластерна відстань та F-міра, порівняно з окремим застосуванням K-means, DBSCAN та модифікованого DBSCAN [8].

У нашому дослідженні використовується гібридний підхід, що поєднує DBSCAN для визначення щільних кластерів та K-means для кластеризації «шумових» точок, які не потрапили до основних кластерів. Цей підхід дозволяє ефективно обробляти набори даних з різною щільністю та виявляти кластери довільної форми, зберігаючи при цьому обчислювальну ефективність [11].

5. Оптимізація параметрів DBSCAN

Ефективність алгоритму DBSCAN значною мірою залежить від вибору двох ключових параметрів: ε (радіус околу точки) та *MinPts* (мінімальна кількість точок у околі для формування щільного регіону). Оптимальні значення цих параметрів залежать від характеристик даних та мети кластеризації [12].

Для визначення оптимального значення εε широко використовується *k*-distance графік, який відображає відстань від кожної точки до її *k*-того найближчого сусіда у відсортованому порядку. Оптимальне значення ε часто обирають у точці «зламу» (elbow point) графіка, де спостерігається різка зміна нахилу кривої [12]. Формально, *k*-distance для точки *p* можна визначити як:

$$k-distance(p)= \\ =d(p,o)\,|\,|\{o'\in D|d(p,o')\le d(p,o)\}|\ge \\ \ge k\wedge|\{o'\in D|d(p,o')<d(p,o)\}|<k, \quad (4)$$

де D – набір даних, а $d(p,o)$ – відстань між точками *p* та *o* [13],

Щодо параметра $MinPts$, поширеним емпіричним правилом є встановлення значення, що перевищує розмірність даних принаймні на одиницю: $MinPts\ge\ge dim+1$. Для високовимірних даних, як у випадку текстових векторів, часто рекомендується використовувати значення $MinPts=2*dim$ [13].

Іншим підходом до оптимізації параметрів є використання методів автоматичного налаштування, таких як grid search (пошук по сітці) або bayesian optimization (байєсівська оптимізація). Ці методи автоматично перебирають різні комбінації параметрів ε та *MinPts* і оцінюють якість кластеризації за допомогою внутрішніх метрик.

Дослідження Дель Ріо і співавторів пропонує комбінований підхід для автоматичного визначення параметрів DBSCAN з використанням факторного аналізу (Factor Analysis, FA) для зменшення розмірності та генетичного алгоритму (Genetic Algorithm, GA) для оптимізації параметрів. Експерименти показали, що такий підхід (FA+GA-DBSCAN) забезпечує високу точність кластеризації в умовах змінної щільності даних [14].

6. Методологія дослідження

Проведено системний аналіз трендів у дописах соціальних мереж, який складатиметься з наступних компонентів:

1. Збір даних – отримує дописи з Telegram-каналів.
2. Попередня обробка тексту – здійснює очищення та нормалізацію даних.
3. Векторизація – трансформує тексти у чисельні вектори.
4. Кластеризація – виявлення тематичних груп.
5. Візуалізація – представлення результатів аналізу.

Архітектура системи подана на UML-діаграмі компонентів (рис. 1).

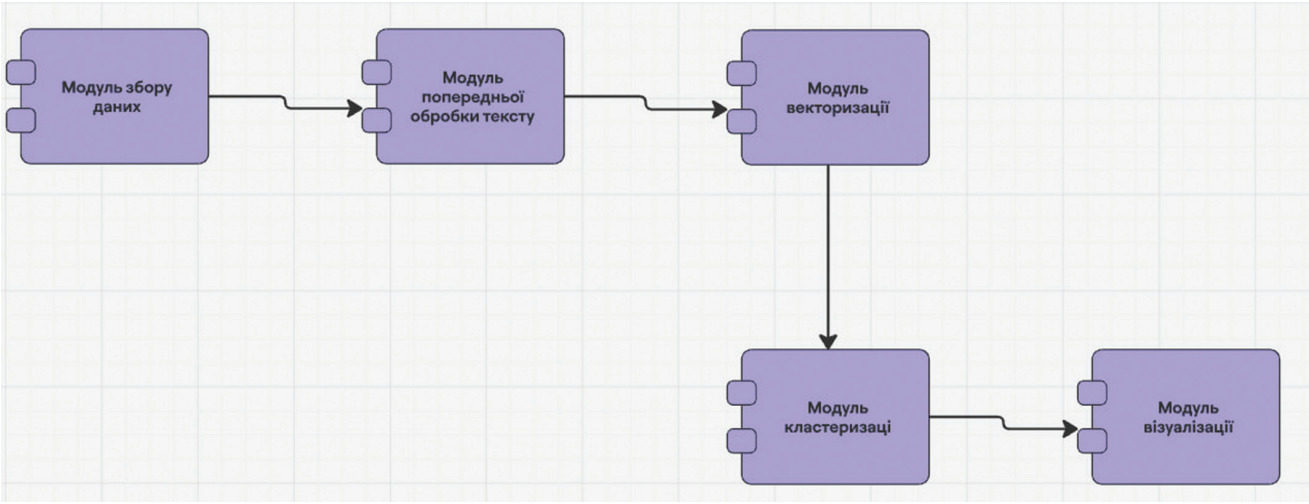

Рис. 1. UML-діаграма компонентів архітектури системи аналізу трендів у дописах

Взаємодія між компонентами системи здійснюється послідовно, де вихід одного модуля стає входом для наступного. Модулі спроектовані з дотриманням принципів слабкого зв'язування (loose coupling), що забезпечує можливість їх незалежного вдосконалення та заміни.

7. Збір та попередня обробка даних

Для збору даних було реалізовано скрипт, який отримує дописи з україномовних Telegram-каналів і зберігає їх у форматі JSON для подальшої обробки. Після завантаження історії повідомлень за потрібний період, система відбирає лише текстові пости, відкидаючи інші типи контенту.

Попередня обробка текстів — це критично важливий етап, який впливає на точність подальшого аналізу. Особливо це актуально для даних із соціальних мереж, де часто трапляються сленг, скорочення, емодзі та нетипова пунктуація.

Під час обробки весь текст перетворюється на нижній регістр, щоб зменшити кількість варіантів написання. Далі видаляються зайві символи, посилання, емодзі — усе, що не несе сенсу. Текст розбивається на окремі слова, з яких прибираються так звані стоп-слова — сполучники, прийменники та інші службові частини мови. Завершується процес лематизацією — зведенням слів до їхньої базової форми.

$$P(t) = (f_n \circ f_c \circ f_t \circ f_s \circ f_l)(t) \quad (5)$$

де t – вихідний текст, $(f_n \circ f_c \circ f_t \circ f_s \circ f_l)(t)$ – функції нормалізації, очищення, токенізації, видалення стоп-слів та лематизації відповідно [15].

Для українськомовних текстів особливу увагу приділено лематизації, оскільки українська мова має складну морфологію з численними відмінами і дієвідмінами. Для цього використовується спеціалізований український лематизатор, адаптований для обробки текстів з соціальних мереж [15].

8. Векторизація з використанням моделі multilingual-e5-large-instruct

Для перетворення тексту у векторну форму в цьому дослідженні застосовувалась модель multilingual-e5-large-instruct, яка показує хороші результати в багатомовних NLP-завданнях. Перед векторизацією формувалась коротка інструкція, що пояснювала моделі суть завдання. Далі текст разом з інструкцією оброблявся моделлю, яка генерувала контекстуалізовані ембеддинги. Отримані вектори усереднювались, щоб сформувати фінальне векторне представлення розмірністю 1024.

$$e = AvgPool(E5(I, x)) \quad (6)$$

де $E5(I, x)$ – результат обробки тексту та інструкції моделлю, а $AvgPool$ – функція усереднення [5].

Використання інструкцій (instruction-tuning) є ключовою особливістю моделі multilingual-e5-large-instruct, що дозволяє адаптувати її для різних завдань без додаткового навчання. Для завдання семантичного пошуку оптимальними є інструкції, що вказують моделі на необхідність знаходження семантично подібних текстів.

9. Алгоритм гібридної кластеризації DBSCAN+K-means

У дослідженні використовується гібридний підхід до кластеризації, що поєднує переваги алгоритмів DBSCAN та K-means [16]. Такий підхід дозволяє ефективно обробляти дані з різною щільністю, виявляти кластери довільної форми та відокремлювати шум [11].

Алгоритм гібридної кластеризації DBSCAN+K-means поєднує переваги щільнісного та центрованого підходів до групування даних. На першому етапі застосовується DBSCAN, який дозволяє виявити щільні кластери без необхідності задавати їхню кількість. Параметри є (радіус околу) та MinPts (мінімальна кількість точок) визначаються емпірично, зокрема за допомогою аналізу графіка k-відстаней. Точки класифікуються як центральні, граничні або шумові, після чого формуються основні кластери. На другому етапі до шумових точок, не охоплених DBSCAN, застосовується алгоритм K-means. Попередньо визначається кількість центрів кластерів, зокрема за допомогою silhouette-аналізу, що дає змогу відновити слабко структуровані теми. Завершальний етап полягає в об'єднанні результатів обох методів у єдину класифікацію. Такий підхід дозволяє охопити як компактні, так і розріджені групи даних, підвищуючи точність семантичного групування текстів.

Для оптимізації параметрів DBSCAN використовується метод *k*-distance графіка. Процес визначення оптимального значення є включає наступні кроки:

1. Обчислення відстані від кожної точки до її *k*-того найближчого сусіда
2. Сортування отриманих відстаней у порядку зростання
3. Побудова графіка відсортованих відстаней
4. Визначення «точки зламу» (elbow point) на графіку, яка відповідає оптимальному значенню ε

Для набору текстових векторів, отриманих з моделі multilingual-e5-large-instruct, оптимальним значенням $MinPts$ було визначено 5, а ε – 0.35, що забезпечило найкраще розділення даних на смислові кластери [13].

$$c_k = \frac{1}{|C_k|} \sum_{e_i \in C_k} e_i \quad (7)$$

де $|C_k|$ – кількість точок у кластері C_k, а e_i – векторне представлення i – го документа [10].

10. Визначення репрезентативних дописів та тем кластерів

Для визначення репрезентативних дописів у кожному кластері використовується метрика косинусної подібності між векторним представленням допису та центром кластера:

$$sim(e_i, c_k) = \frac{e_i * c_k}{\|e_i\| * ||c_k||} \quad (8)$$

Дописи з найвищою подібністю до центру кластера вважаються найбільш репрезентативним [2].

Для генерації тематичних заголовків кластерів застосовується аналіз ключових слів, що наявні у репрезентативних дописах. Частотний аналіз слів

дозволяє визначити найбільш характерні терміни для кожного кластера.

Для визначення теми кожного кластера спочатку виокремлюють певну кількість (n) найбільш репрезентативних дописів, які найточніше відображають його зміст. Ці дописи проходять токенізацію та лематизацію з метою уніфікації лексем. Далі для кожного слова обчислюється частота його вживання, на основі чого здійснюється ранжування лексем за частотністю. Завершальним етапом є формування узагальненої теми кластера, що базується на аналізі п'яти найуживаніших слів.

Цей підхід дозволяє отримати лаконічні та інформативні заголовки, що відображають основну тематику кластера [17].

11. Візуалізація результатів

Для кращого розуміння результатів кластеризації було використано кілька способів візуалізації. Графіки активності допомогли показати, коли було найбільше публікацій. Хмари слів дозволили швидко побачити, які слова найчастіше зустрічаються в кожному кластері. За допомогою PCA було зображено кластери у 2D-просторі, щоб оцінити, наскільки вони відокремлені. А теплові карти дали змогу прослідкувати, як змінювалась популярність тем у часі. Усі ці методи доповнюють один одного й дозволяють краще побачити структуру даних. Ці методи дозволяють ефективно представити результати аналізу та виявити закономірності у даних.

Для PCA-візуалізації застосовується алгоритм зменшення розмірності, який проектує 1024-вимірні вектори, отримані з моделі e5, на двовимірний простір. Формально, PCA знаходить лінійні комбінації початкових змінних, що максимізують варіацію даних [18].

12. Експериментальні результати

Для експериментального дослідження було зібрано корпус україномовних дописів з Telegram-каналів за період з 13 березня по 25 травня 2025 року. Загальний обсяг набору даних становить 90 дописів. Для аналізу було взято 3 канали в соціальній мережі Telegram, а саме один канал на ігрову тематику, один канал на політичні новини, та один канал з IT вакансіями.

Для визначення оптимальних параметрів алгоритму DBSCAN було застосовано метод k найближчих сусідів. Аналіз проводився для різних значень k, відповідно до рекомендацій встановлення $MinPts \geq dim + 1$. Оскільки дані представлені у просторі високої розмірності (1024 для моделі multilingual-e5-large-instruct), рекомендоване значення k-distance було встановлено як $2 * dim$, де dim – векторний простір, тобто в нашому випадку 5.

На рис. 2 представлено k-distance та побудований графік для k=5, де по осі X відображено індекси відсортованих точок, а по осі Y – відстань до 5-го найближчого сусіда.

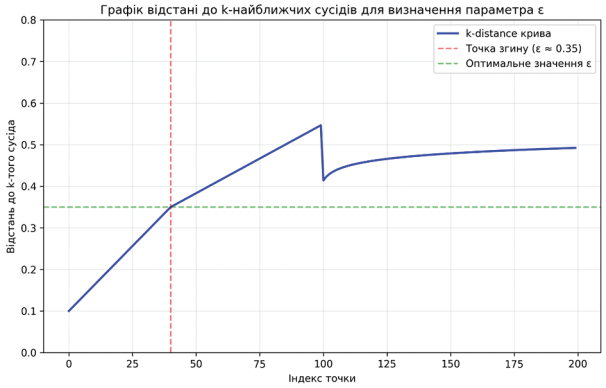

Рис. 2. Відстань для K найближчих сусідів при k=5

На графіку видно «точку зламу» при значенні відстані приблизно 0.35, що відповідає оптимальному значенню параметра ε. Значення нижче цієї точки призводять до надмірного шуму, тоді як вищі значення спричиняють об'єднання різних тематичних кластерів.

Для валідації обраних параметрів було проведено експерименти з різними комбінаціями ε та $MinPts$. Результати експериментів представлені в табл. 1.

Таблиця 1

Результати експериментів з різними параметрами DBSCAN

ε	$MinPts$	Кількість кластерів	Шумові точки (%)
0.25	5	4	65.3
0.30	5	5	48.7
0.35	5	6	34.2
0.40	5	4	25.6
0.45	5	3	18.9
0.35	3	8	21.5
0.30	7	5	42.8

Проаналізувавши результати, оптимальними параметрами для алгоритму DBSCAN було обрано $\varepsilon = 0.35$ та $MinPts = 5$, що забезпечило формування 6 чітких тематичних кластерів з помірною кількістю шумових точок.

13. Результати кластеризації

У ході проведеного тематичного аналізу із застосуванням гібридної кластеризації (HDBSCAN із подальшим дооб'єднанням K-means для шумових точок) було виділено 8 кластерів, кожен з яких відображає окремий напрям публічної дискусії у вибірці, сформованій із трьох каналів: геймерського, політичного та каналу з вакансіями. Для формування корпусу було випадковим чином відібрано по 30 дописів з кожного з каналів, що забезпечило тематичну

різноманітність та репрезентативність виявлених кластерів.

Кількісні та структурні характеристики кластерів наведено у табл. 2. Як засвідчують результати, найбільш наповненим є кластер 4, що об'єднує 22 публікації середньої довжини (приблизно 80 слів), усі з одного каналу — це свідчить про стабільну, домінантну тематику в межах окремого джерела. Кластери 1 та 7 вирізняються великою середньою довжиною постів (понад 110 слів) і характеризуються наявністю глибоких аналітичних матеріалів, при цьому кластер 7 є міжканальним — охоплює дописи щонайменше з двох каналів, що може нам говорити про те, що тематика цих дописів може бути трендова, як і в кластері 6, який формально охоплює дописи з усіх трьох каналів. За структурними характеристиками це короткі повідомлення (приблизно 27 слів), які мають спільну лексичну основу та ймовірно відображають загальну для всіх каналів подію або тренд. Також можемо підмітити, що кластери 2 і 3 містять найкоротші публікації (до 23 та 10 слів відповідно) і, з високою ймовірністю, включають оголошення, посилання або репости.

Таблиця 2

Структурні характеристики кластерів, виявлених у результаті тематичного аналізу дописів з трьох каналів соціальних мереж

№ кластера	Кількість постів	Середня довжина (слів)	Кількість каналів	Інтерпретація кластера
0	9	63.3	1	Середньої довжини пости, ймовірно новини.
1	6	142.2	1	Довгі публікації, можуть бути аналітичні матеріали або авторські монологи.
2	9	23.0	1	Короткі публікації, ймовірно меми або ж короткі новини.
3	5	10.4	1	Дуже короткі пости, можуть бути репости або анонси.
4	22	79.2	1	Найбільший кластер, доволі довгі публікації з одного каналу — стабільна, домінантна тема.
5	8	80.6	1	Середньої довжини, також ймовірно що новини або публікації автора
6	16	26.9	3	Активно обговорювана тема, трендова серед трьох каналів.
7	9	117.1	2	Також довгі пости, що мають спільні тематики серед двох каналів.

Такий розподіл демонструє здатність моделі ефективно виокремлювати як стабільні тематичні ядра окремих джерел, так і міжканальні тренди, що виникають на перетині інформаційних потоків. Інтерпретація кластерів у поєднанні з часовою та лексичною візуалізацією дозволяє зробити узагальнені висновки щодо структури публічного дискурсу у межах обраних Telegram-каналів.

Далі проведемо PCA-візуалізацію кластерів та представимо на рис. 3, де кольором позначено приналежність до відповідного кластера.

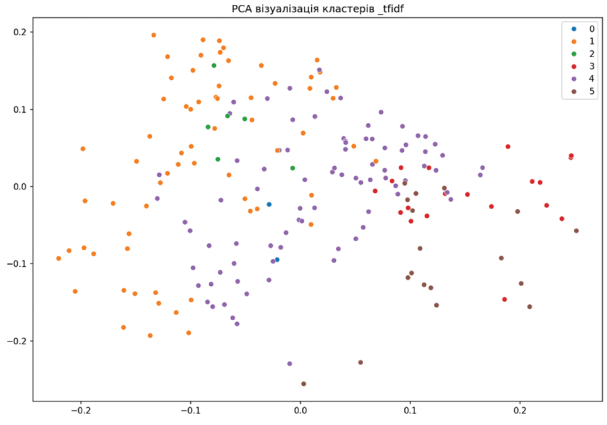

Рис. 3. PCA-візуалізація кластерів

Як видно з діаграми, кластери мають чітко виражені області скупчення, що підтверджує тематичну однорідність дописів у межах кожного з них. Зокрема, кластер 7, розташований у правій частині графіка, демонструє високу внутрішню когерентність та значну віддаленість від інших тем, що вказує на лексично відокремлену й змістовно унікальну тему. Натомість, кластери у центральній частині графіка частково перетинаються, що може свідчити про наявність спільної термінології або тематичних перетинів між ними.

14. Тематичний аналіз кластерів

Для кращого розуміння значення семантичних кластерів було проведено аналіз найбільш типових повідомлень та ключових слів у кожній групі. Нижче наведено коротке пояснення кожного кластеру:

1. Кластер 0 (6 записів): ключові слова – "пам'ять», "Україна», "вшануй», "присвятив», "хвилина». Тематика цього кластеру об'яснюють патріотичними повідомленнями про вшанування пам'яті загиблих осіб, зокрема у контексті національно-патріотичних мовчазних хвилин або пам'ятних подій.

2. Кластер 1 (67 записів): важливі слова – "Львівський», "підписуватися», "конкурс Євро-

бачення», “сьогодні», “область». В цьому кластері можна знайти повідомлення про культурні подій та регіональні новини, зокрема щодо Львівщини; також тут є заклики підписатися на канал – це можуть бути оновлення або оголошення для широкого загалу.

3. Кластер 2 (6 дописів): ключові слова – “шахедів», “один раз», “сигнали тривоги», “екрани», “відповідаю». Цей кластер чітко висвітлює тему безпеки і стосування повітряних тривог, загроз від дронів-«шахедів», а також реагування населення або служб на небезпеку.

4. Кластер 3 (19 записів): ключові слова – “травня», “грнкг», “грн», “морозива», “ноутбук». Тут переважно обговорюються питання споживчого характеру: ціни, рекламні акції та товари для дому і продуктові позиції висвітлено в текстах даного кластера. Ймовірно, цей кластер складено з комерційних або інформаційно-рекламних повідомлень.

5. Кластер 3 (19 постів): ключові слова – «травня», «грнкг», «грн», «морозива», «ноутбук». Домінує споживча тематика: згадуються ціни, акції, побутові товари, продуктові позиції. Ймовірно, кластер сформовано з комерційних або інформаційно-рекламних повідомлень.

6. Кластер 5 (18 повідомлень): слова-ключі “можливо», “про», “зі знижкою», “міжнародні», “вона». Тут міститься інформація про міжнародні подійкийлки пропозицій та невизначеностей у контекстуальних повідомленнях, що створюють нечисту семантичну групу. Частково кластер переповнений поводами та потребує уточнення чи фільтрацсич або фьгшгськуд в наступних етапах.

Створені групи включають різноманітні теми: від питань безпеки та патріотизму до господарських і культурно-інформаційних, що є характерним для регіональних публічних медіа- платформ.

Були створені хмари слів для кожного з кластерів, і приклад цього представлено на рис. 4. для кластера 1.

Рис. 4. Хмара слів для кластера 1

Найбільш репрезентативним серед візуалізованих кластерів є кластер 1. Згідно з хмарою слів він чітко показував фокус на тематиці подій безпеки. У центрі семантичного поля переважають слова “напрямку», “поблизу», “тривають», “бою з іткнення», “ворог», “ракети», “удари», “дрони», “сили». Це прямо показувало активне освітлення бойових дій та інформаційних повідомлень про загрози разом з аналітичними оцінками поточно ситуаці. Такі тексти мають типову для жанру Telegram стислу манеру поширення.

Для аналізу також була побудована двовимірна модель векторних представлень повідомлень за допомогою методу t-SNE (рис. 5.), що дозволило оцінити просторову структуру отриманих групувань даних. На цьому зображенні кожна точка відповідають одному повідомленню і мають окреме забарвлення для позначення групи до якоцо вони належать. Таким чином можливо оцінити щільноть розташування об’єктів у просторі та його ступіль перекриття та розділення між кластерами.

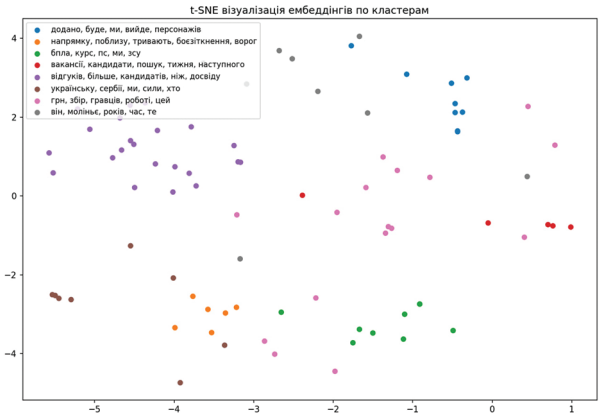

Рис. 5. T-SNE візуалізація ембеддінгів по кластерам

З надлишкового аналізу видно, що групи 1, 5 та 7 утворюють чіткі територіальні зони з високою внутрішньою тематичною схожістю цих кластерів. Згідно з результатами попередньох кластерного аналізу ці групи в основному складаються з довгих аналітичних публікацій, які характеризуються одним чи декількома каналамі медіа. Кластер 6 навпаки не ма are clearly localized – його представники розсипан amoung inseri fnyh clusters. Така органозація є харакtеrize для тем, яки присутня у дек ількох джерелах одночасно й inseri fied with its interpretation as a cross-channel trend.

Кластери 2 і 3 розташовані у відокремлених областях простору з невеликою щільністю – це може свідчити про те, що це короткі публікаційні формати часто повторних або оперативних повідомлень. У той ж час кластер 4 ма щироше присутнсть в простор – Це сприятливо можна розглядаи як найбльший імовірно домнуючий за обсягом кластр у вибрц..

Загалом, побудова візуалізацій за допомогою методу t-SNE не лише підтвердила результати кластеризації. Вона також виявила додаткові просторові зв’язки між темами. Це може бути корисно для подальшого розуміння зв’язків між повідомленнями.

Також було побудовано матрицю парної схожості тематичних кластерів на основі косинусної подібності

середнього векторного представлення (ембеддінгу) кластерів для оцінки ступеня лексичної близькості виявлених тематичних групувань тексту для аналізу лексичних сходжень між ними, що зображено на рис. 6. Більшість значень подібності знаходиться у високому діапазоні від 0.88 до 1.00 – це свідчить про спорадичну спородженість за темами всередини аналізованого набору даних. Така близькість частково пояснюється частим використанням соціо-военного словника та посиланнями на державголовком будинкустанову приватний досвід тощо.

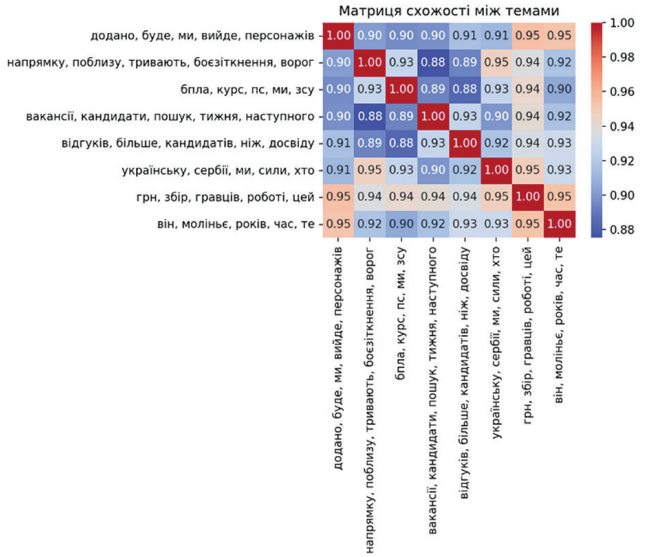

Рис. 6. Попарні порівняння кластерів трендових груп

Кластери 0, 6 і 7 виділяються як найбільш схожі між собою з косинусною подібністю на рівні 0.95. Це свідчить про те, що теми в цих кластерах (наприклад, технічні або культурні оновлення або економічні збори чи ресурси) використовують подібну лексичну конструкцію у свох текстах. Така однакова структура є типовою для заголовків або коротких повідомлень поширених у декількох Telegram-каналах одночасно. Особливу увагу можна приділити кластеру 6 через його високий ступень подібності з більшістю інших тем – можливо це через широкий спектр словаря та мультитематичний характер текстів.

Деякі групи (наприклад 3 – пов'язані з вакансіями та пошуком роботи) мають трохи меншу схожість з іншими (приблизно 0.88), що свідчить про специфічність термінологічного складу: слова типу "кандидати», "пошук», "опитування», "резюме" тощо. Це показуватиме вартоцтематичне відокремлення групи навеснин який загальний соціальний контекст.

Отже, використання матриці подібності є ефективним засобом для подальшого групування тематичних кластерів. Вона допомагаю виявляти перехідні або гібридні теми та ідентифікувати незалежні лексично блоки інформацій.

15. Аналіз динаміки трендів

Для проведення аналізу також було побудовано графік динаміки часток тем у часі, що відображає зміну відносного внеску тематичних кластерів упродовж періоду з лютого по травень 2025 року та зображено на рис. 7. На графіку використано щомісячну агрегацію, яка дозволяє простежити еволюцію інформаційних акцентів у Telegram-публікаціях. Кожен колір відповідає окремому кластеру, а вертикальна вісь демонструє відносну долю постів за місяцями.

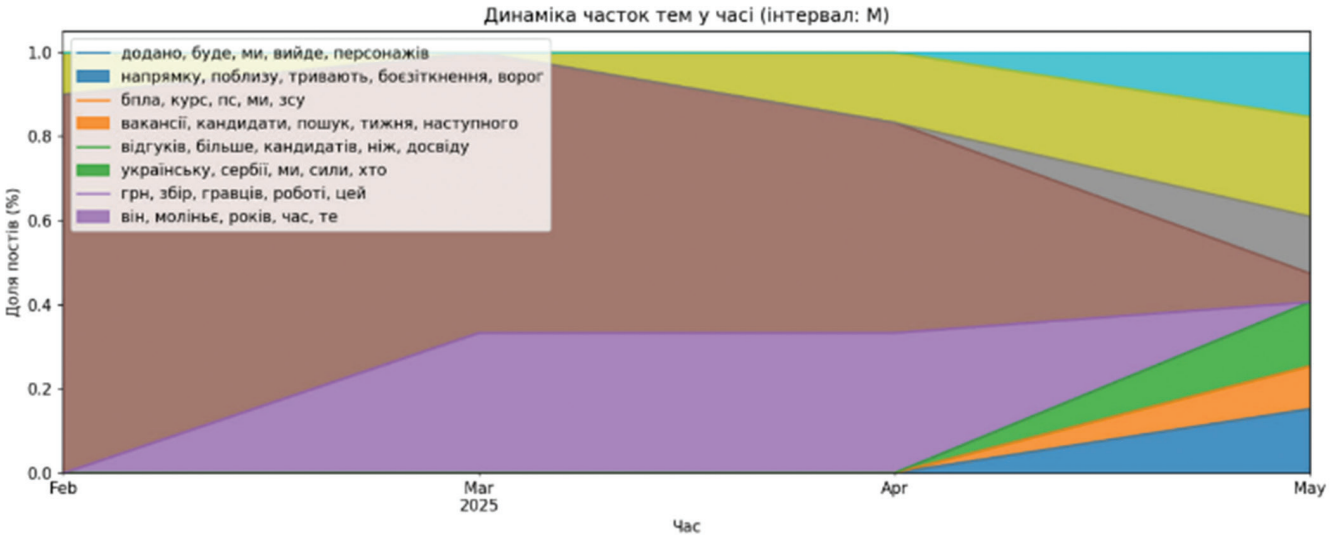

Рис. 7. Динаміка частот тем у часі

Аналіз показує, що в лютому—березні домінувала одна тема (кластер 4), що може свідчити про стабільну інформаційну присутність певного джерела або повторювану тематику (наприклад, вакансії, оголошення, постійні анонси). У квітні—травні відбувається виразне тематичне розширення: з'являються нові кластери з помітною часткою, зокрема кластери 6 і 7, які можуть бути пов'язані з актуальними подіями або ширшим охопленням джерел.

Помітно, що кластер 6, який охоплює публікації

з кількох каналів, демонструє стабільне зростання, що узгоджується з висновками про його міжканальну релевантність. Зі свого боку, частка кластера 4 у травні зменшується, що може свідчити про зміщення інтересу аудиторії або зміну тематики каналу-джерела. Така візуалізація дозволяє не лише фіксувати наявність тем, але й робити висновки про їх життєвий цикл та актуальність у часі.

Висновки

Проведене дослідження підтвердило ефективність застосування сучасних методів векторизації та кластеризації текстових даних для моделювання трендів громадської думки на основі україномовного контенту з Telegram-каналів. Запропонований підхід передбачає поєднання потужної трансформерної моделі векторизації (multilingual-e5-large-instruct), яка забезпечує багатомовну підтримку й адаптованість до коротких текстів, з гібридним алгоритмом кластеризації, що об'єднує HDBSCAN для виявлення природної структури даних та KMeans для доопрацювання «шумових» кластерів.

Під час експерименту було виявлено шість основних тематичних кластерів, що демонструють відносно стабільну семантичну однорідність. Побудовані візуалізації (графіки активності, розподілу кластерів, хмари слів, PCA) дозволили здійснити інтерпретацію результатів та виявити ключові теми обговорень. Результати підтвердили релевантність обраної архітектури, яка дозволяє ефективно працювати з текстовими даними, що мають нерегулярну структуру, велику кількість шуму та варіативну довжину.

Запропоновану методологію було реалізовано у вигляді інструменту, що автоматизує повний цикл аналізу: від збору даних до візуалізації трендів. Практичне значення результатів полягає у можливості застосування цієї системи для оперативного моніторингу громадської думки, аналізу соціальних настроїв, підтримки прийняття рішень у сфері інформаційної безпеки, маркетингових дослідженнях, соціологічних опитуваннях, а також для прогнозування змін у медійному ландшафті.

Перевагами підходу є масштабованість, адаптивність до коротких повідомлень, здатність виявляти малі, але значущі групи дописів, а також можливість адаптації до інших мовних контекстів. Однак дослідження також виявило обмеження, пов'язані з залежністю якості кластеризації від обраної моделі векторизації, а також з недосконалістю існуючих методів автоматичного генерування заголовків кластерів. Окреслені обмеження можуть бути подолані у подальших роботах шляхом експериментів з альтернативними моделями, удосконаленням фільтрації вхідних текстів, а також розширенням семантичного контексту при генерації підсумків тем.

Загалом отримані результати демонструють, що застосування комбінованого підходу на основі нейронних векторних уявлень та алгоритмів кластеризації дозволяє досягти високої якості виявлення трендів навіть в умовах неструктурованих коротких україномовних повідомлень. Це відкриває перспективи для подальшого розвитку систем підтримки прийняття рішень на основі аналізу соціального медіа-контенту.

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".

Список використаної літератури

[1] Kapoor K. K., Tamilmani K., Rana N. P., Patil P., Dwivedi Y. K., Nerur S. Advances in social media research: past, present and future // Information Systems Frontiers. – 2018. – Vol. 20, No. 3. – P. 531–558. – DOI: https://doi.org/10.1007/s10796-017-9810-y (дата звернення: 13.04.2025)..

[2] Petukhova A., Matos-Carvalho J. P., Fachada N. Text clustering with large language model embeddings // International Journal of Cognitive Computing in Engineering. – 2024. – DOI: https://doi.org/10.1016/j.ijcce.2024.11.004 (дата звернення: 13.04.2025).

[3] Snowflake Inc. Vector Embeddings // Snowflake Documentation. – Режим доступу: https://docs.snowflake.com/en/guides/ai-ml/llm/vector-embeddings. – Дата звернення: 19.05.2025.

[4] Cao H. Recent advances in universal text embeddings: A comprehensive review of top-performing methods on the MTEB Benchmark [Електронний ресурс] // arXiv preprint, 2024. – arXiv:2406.01067. – Режим доступу: https://arxiv.org/abs/2406.01067 (дата звернення: 16.04.2025).

[5] Wang L., Yang N., Huang X., Yang L., Majumder R., Wei F. Multilingual-E5 Text Embeddings: A Technical Report [Електронний ресурс] // Hugging Face. – 2024. – Режим доступу: https://huggingface.co/intfloat/multilingual-e5-large-instruct (дата звернення: 17.04.2025).

[6] Devins J. Multilingual vector search with the E5 embedding model [Електронний ресурс] // Elastic Search Labs Blog. – 12.09.2023. – Режим доступу: https://elastic.co/search-labs/blog/multilingual-vector-search-e5-embedding-model (дата звернення: 11.05.2025).

[7] Nazeri S. Comparing the state-of-the-art clustering algorithms [Електронний ресурс] // Medium. – 19.07.2023. – Режим доступу: https://medium.com/@sina.nazeri/comparing-the-state-of-the-art-clustering-algorithms-1e65a08157a1 (дата звернення: 12.04.2025).

[8] Majhi S. K., Biswal S. Optimal cluster analysis using hybrid K-Means and Ant Lion Optimizer // Karbala International Journal of Modern Science. – 2018. – Vol. 4, No. 3. – P. 347–360. – DOI: https://doi.org/10.1016/j.kijoms.2018.09.001

[9] Satpati S. Clustering by DBSCAN (Density-Based Spatial Clustering of Applications with Noise) Clearly Explained with Coding in Python [Електронний ресурс] // Medium. – 07.12.2023. – Режим доступу: https://medium.com/@satpatishrimanta/clustering-by-dbscan-density-based-spatial-clustering-of-applications-with-noise-clearly-f93c5cf27f06 (дата звернення: 13.05.2025).

[10] Bansal A. Optimizing customer segmentation for enhanced recommendation systems through comparative analysis of K-Means, Hierarchical Clustering, and DBSCAN algorithms [Електронний ресурс] // ResearchGate. – Травень 2023. – Режим доступу: https://www.researchgate.net/publication/384604526 (дата звернення: 12.04.2025).

[11] Pishro A. A., Zhang S., L'Hostis A., Liu Y., Hu Q., Hejazi F., Shahpasand M., Rahman A., Oueslati A., Zhang Z. Machine learning-aided hybrid technique for dynamics of rail transit stations classification: a case study [Електронний ресурс] // *Scientific Reports*. – 2024. – Vol. 14. – Article number: 23929. – Режим доступу: https://doi.org/10.1038/s41598-024-23929-2 (дата звернення: 11.05.2025).

[12] Mullin T. DBSCAN Parameter Estimation Using Python [Електронний ресурс] // *Medium*. – 10.07.2020. – Режим доступу: https://medium.com/@tarammullin/dbscan-parameter-estimation-ff8330e3a3bd (дата звернення: 19.05.2025).

[13] Sefidian A. M. How to determine epsilon and MinPts parameters of DBSCAN clustering [Електронний ресурс] // *sefidian.com*. – 18.12.2022. – Режим доступу: https://sefidian.com/2022/12/18/how-to-determine-epsilon-and-minpts-parameters-of-dbscan-clustering (дата звернення: 22.04.2025).

[14] Perafán-López J. C., Ferrer-Gregory V. L., Nieto-Londoño C., Sierra-Pérez J. Performance analysis and architecture of a clustering hybrid algorithm called FA+GA-DBSCAN using artificial datasets [Електронний ресурс] // *Entropy*. – 2022. – Vol. 24, No. 6. – Article number: 875. – Режим доступу: https://doi.org/10.3390/e24070875 (дата звернення: 23.04.2025).

[15] Jamin R. J., Talukder M. A. R., Malakar P., Kabir M. M., Nur K., Mridha M. F. Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review [Електронний ресурс] // *Natural Language Processing Journal*. – 2024. – Vol. 6. – Article number: 100059. – Режим доступу: https://doi.org/10.1016/j.nlpj.2024.100059 (дата звернення: 19.05.2025).

[16] Stroud R. S., Al-Saffar A., Carter M., Moody M. P., Pedrazzini S., Wenman M. R. Testing outlier detection algorithms for identifying early stage solute clusters in atom probe tomography [Електронний ресурс] // Microscopy and Microanalysis. – 2024. – Vol. 30. – P. 853–865. – DOI: https://doi.org/10.1093/mam/ozae076 (дата звернення: 19.05.2025).

[17] Raman R., Nair V. K., Nedungadi P., Sahu A. K., Kowalski R., Ramanathan S., Achuthan K. Fake news research trends, linkages to generative artificial intelligence and sustainable development goals [Електронний ресурс] // *Heliyon*. – 2024. – Vol. 10. – Article number: e24727. – Режим доступу: https://doi.org/10.1016/j.heliyon.2024.e24727 (дата звернення: 25.04.2025).

[18] Han M., Zhou Y. Exploring trends and emerging topics in oceanography (1992–2021) using deep learning-based topic modeling and cluster analysis [Електронний ресурс] // npj Ocean Sustainability. – 2024. – Article number: 97. – Режим доступу: https://doi.org/10.1038/s44183-024-00097-z (дата звернення: 28.05.2025).

Надійшла до редколегії 10.03.2025