УДК 004.8:37.018.43 **DOI: https://doi.org/10.30837/PMW.2026.T1.118**

МЕТОДИЧНІ ЗАСАДИ ТЕМАТИЧНОГО МОДЕЛЮВАННЯ ТЕКСТОВИХ КОРПУСІВ У ДОСЛІДЖЕННЯХ ВИКОРИСТАННЯ ШТУЧНОГО ІНТЕЛЕКТУ В ДИСТАНЦІЙНІЙ ОСВІТІ

Слісаренко Р.В.
аспірант, кафедра «Медіасистеми та технології»,
Харківський національний університет радіоелектроніки
ORCID ID: 0009-0009-3286-4333

Анотація. *У розділі розглянуто теоретичні та методичні засади тематичного моделювання текстових корпусів, у яких відображено використання штучного інтелекту в дистанційній освіті. Здійснено порівняльний аналіз класичних і сучасних підходів до виявлення тематичних структур у текстових даних цієї предметної області. Обґрунтовано доцільність використання гібридної методики на основі інформованих пріорів для Latent Dirichlet Allocation та формалізованого фреймворку динамічного контекстуального тематичного моделювання з використанням великих мовних моделей. Показано, що запропонований підхід підвищує змістову релевантність, інтерпретованість і аналітичну придатність результатів аналізу.*

Ключові слова: *тематичне моделювання, дистанційна освіта, штучний інтелект в освіті, освітні текстові корпуси, latent dirichlet allocation, dynamic contextual topic modeling.*

Вступ

Сучасний етап розвитку дистанційної освіти характеризується активним використанням засобів штучного інтелекту в навчальному процесі, підтримці здобувачів освіти, аналізі освітнього контенту та цифровій організації освітнього середовища. Це зумовлює появу значних масивів текстових даних, у яких фіксуються різні аспекти впровадження, оцінювання та осмислення можливостей штучного інтелекту в дистанційному навчанні. До таких даних належать наукові публікації, аналітичні огляди, описи цифрових платформ, результати опитувань, форуми, відкриті відповіді учасників освітнього процесу та інші текстові джерела, що відображають тенденції розвитку цієї сфери [1, 2].

За таких умов особливої ваги набувають методи аналізу текстових корпусів, здатні виявляти приховані змістові закономірності, стійкі тематичні напрями та проблемні аспекти досліджуваної предметної області. Одним із найбільш перспективних інструментів такого аналізу є тематичне моделювання, яке дозволяє автоматизовано виявляти теми у великих масивах текстів без попереднього ручного маркування [1, 2]. У межах цього дослідження тематичне моделювання розглядається як метод аналізу текстових матеріалів, у яких відображено використання штучного інтелекту в дистанційній освіті, а не як

інструмент безпосереднього впровадження штучного інтелекту в освітню практику.

Разом із тим застосування тематичного моделювання до таких корпусів супроводжується низкою методичних труднощів. Тексти, присвячені використанню штучного інтелекту в дистанційному навчанні, характеризуються термінологічною варіативністю, міждисциплінарністю, різними рівнями формалізації та часовою динамікою змісту. Унаслідок цього класичні підходи, зокрема Latent Dirichlet Allocation (LDA) та Non-negative Matrix Factorization (NMF), не завжди забезпечують достатню чутливість до контексту, стійкість результатів і точність відображення змістових зв'язків у таких текстах [1, 3, 4]. У зв'язку з цим актуальним є не лише порівняльний аналіз наявних методів тематичного моделювання, а й розробка таких методичних рішень, які дозволяють підвищити змістову релевантність, інтерпретованість і контекстуальну чутливість результатів аналізу.

Одним із перспективних напрямів у цьому контексті є використання гібридної методики на основі інформованих пріорів для LDA та формалізованого фреймворку Dynamic Contextual Topic Modeling with Large Language Models (DCTM-LLM) [5].

Мета та задачі дослідження

Метою дослідження є обґрунтування й розробка методичних засад тематичного моделювання текстових корпусів, у яких відображено використання штучного інтелекту в дистанційній освіті [1, 5].

Для досягнення поставленої мети передбачено проаналізувати класичні та сучасні підходи до тематичного моделювання текстових корпусів і визначити їх придатність для дослідження матеріалів, присвячених використанню штучного інтелекту в дистанційній освіті [1-4].

Окрему увагу зосереджено на встановленні переваг та обмежень методів LDA, NMF і class-based Term Frequency–Inverse Document Frequency (c-TF-IDF) у контексті аналізу текстів, що відображають напрями, особливості та тенденції застосування штучного інтелекту в дистанційному навчанні [1, 3, 4].

Подальшим завданням є обґрунтування доцільності використання гібридної методики тематичного моделювання на основі інформованих пріорів для LDA з метою підвищення якості аналізу текстів зазначеної предметної області [1, 5].

Водночас важливо подати формалізований фреймворк DCTM-LLM для дослідження тематичної структури корпусів, присвячених використанню штучного інтелекту в дистанційній освіті [5].

Завершальним завданням є характеристика переваг запропонованого підходу та визначення його аналітичної придатності для виявлення змістових закономірностей у текстових матеріалах досліджуваної тематики [1, 2, 5].

Основна частина

Порівняльний аналіз методів тематичного моделювання в дослідженні наукового дискурсу щодо штучного інтелекту в дистанційній освіті

Тематичне моделювання доцільно розглядати як інструмент виявлення прихованих змістових структур у великих масивах текстових даних, що дозволяє переходити від аналізу окремих документів до узагальненого виявлення тематичних ліній у межах певної предметної області [1, 2]. У контексті цього дослідження особливого значення набувають інтерпретованість результатів, здатність працювати з неоднорідними текстами та можливість виявлення стійких тематичних закономірностей у корпусах, присвячених використанню штучного інтелекту в дистанційній освіті [1, 2, 5].

Для дослідження цієї предметної області доцільно зіставити методи, що репрезентують різні підходи до побудови тематичних структур, а саме LDA, NMF і c-TF-IDF. LDA належить до класичних імовірнісних моделей і виходить із припущення, що документ може бути представлений як суміш тем, а кожна тема – як розподіл імовірностей слів [1, 2]. NMF репрезентує лінійно-алгебраїчний підхід до тематичного аналізу та ґрунтується на факторизації матриці «термін × документ» на дві невід'ємні матриці меншої розмірності [4, 6]. c-TF-IDF, на відміну від двох попередніх підходів, орієнтований на побудову репрезентативних тематичних описів для попередньо виділених кластерів документів і особливо ефективний у поєднанні із сучасними методами векторного представлення текстів [7, [8].

Узагальнення теоретичної основи та принципу роботи розглянутих моделей подано в табл. 1.

Таблиця 1 – Теоретична основа та принцип роботи тематичних моделей

Модель	Коротка сутність	Основний механізм
LDA	Генеративна імовірнісна модель	Документи моделюються як суміші тем (розподіл Діріхле). Теми є розподілами слів
NMF	Лінійно-алгебраїчний розклад	Розклад матриці «термін × документ» у невід'ємні фактори W та H (ранг K)
c-TF-IDF	Кластеризація на ембеддингах	Інформаційно-пошукова вага терміна для кластера або «метадокумента» після кластеризації за допомогою UMAP/HDBSCAN

Як видно з табл. 1, розглянуті моделі відрізняються не лише технічною реалізацією, а й способом формування тематичного представлення тексту. LDA формує теми в межах імовірнісної генеративної схеми, NMF спирається на матричну факторизацію, тоді як c-TF-IDF працює в межах кластерно-орієнтованого підходу. Саме ця відмінність методологічної основи визначає різницю в інтерпретованості результатів, стійкості моделей та їх придатності до роботи з короткими й неоднорідними текстами [3, 5, 7, 8].

Порівняльний аналіз якісних характеристик розглянутих моделей дає змогу точніше оцінити їх придатність до дослідження корпусів, у яких

висвітлюється використання штучного інтелекту в дистанційній освіті. Для таких текстів суттєвими є не лише формальна здатність моделі виокремлювати теми, а й когерентність тематичних описів, рівень їх інтерпретованості та відповідність людському сприйняттю змісту. Узагальнення цих характеристик подано в табл. 2.

Таблиця 2 – Якісні характеристики, інтерпретованість та оцінка тем

Критерій	LDA	NMF	c-TF-IDF (BERTopic)
Когерентність / Інтерпретованість	Середня. Висока лише на довших текстах. Чутлива до вибору K	Висока. Формує «частинно-орієнтовані» теми	Найвища. Завдяки контекстним ембедингам та c-TF-IDF [8]
Робота з короткими текстами	Низька. Базова LDA програє; потрібні спеціалізовані версії	Добре. Демонструє переваги на розріджених даних	Відмінно. Розроблено для коротких і різнорідних даних [3]
Людиноцентрична оцінка	Не завжди узгоджується з людським сприйняттям на коротких даних	Часто добре сприймається експертами через «гостріші» теми	Найкраща. Найвища відповідність людським судженням у нових роботах [8]

Як видно з табл. 2, LDA зберігає значення як класичний інтерпретований підхід, однак його результати суттєво залежать від довжини текстів і параметричної конфігурації моделі [3, 8].

NMF у багатьох випадках формує чіткіші тематичні профілі, особливо для розріджених даних, однак не позбавлений проблеми залежності від ініціалізації та конфігурації факторизації [3, 4, 6, 8].

c-TF-IDF у поєднанні з embedding-based підходами демонструє найкращі результати щодо когерентності, інтерпретованості й людиноцентричної оцінки, що робить його особливо перспективним для аналізу коротких та неоднорідних текстів [3, 7, 8].

Таким чином, проведений порівняльний аналіз засвідчує, що жоден із розглянутих методів окремо не забезпечує повного врахування специфіки текстових корпусів, у яких висвітлюється використання штучного інтелекту в дистанційній освіті. LDA забезпечує добру статистичну основу та інтерпретованість, але слабше враховує контекст; NMF дає більш сфокусовані профілі, проте залишається чутливим до конфігурації моделі; c-TF-IDF підвищує виразність тематичних описів, але залежить від якості попередньої кластеризації та векторного представлення текстів [3, 4, 7, 8]. Це зумовлює необхідність пошуку таких методичних рішень, які дозволяють поєднати інтерпретованість, контекстуальну чутливість і змістову релевантність тематичного аналізу.

Передумови вдосконалення тематичного моделювання для задач дистанційної освіти

Проведений порівняльний аналіз LDA, NMF і c-TF-IDF показав, що кожен із цих підходів має власні переваги, проте жоден із них окремо не забезпечує

повного врахування специфіки текстових корпусів, у яких висвітлюється використання штучного інтелекту в дистанційній освіті. Для цієї предметної області характерними є короткі та неоднорідні тексти, міждисциплінарна термінологія, швидке оновлення проблематики та потреба в інтерпретованому представленні тематичної структури. Саме тому питання вдосконалення тематичного моделювання в такому контексті набуває не лише методичного, а й прикладного значення [3, 5, 8].

Однією з ключових передумов удосконалення є обмеженість статичних неконтекстуальних моделей при роботі з короткими текстами. Для LDA проблема полягає в розрідженості даних і залежності якості тем від наперед заданої кількості компонент, тоді як для NMF суттєвими залишаються чутливість до ініціалізації та залежність від параметризації факторизації [3, 4, 6]. Хоча c-TF-IDF демонструє вищу якість тематичних описів, його результативність значною мірою визначається якістю попередньої кластеризації та характеристиками векторного простору документів [7, 8].

Важливою умовою підвищення якості тематичного аналізу є також коректна попередня обробка корпусу. Уніфікована нормалізація тексту, видалення стоп-слів, лематизація та фільтрація надто коротких і надто довгих документів дозволяють зменшити варіативність корпусу, знизити вплив шумової інформації та сформувати більш однорідне семантичне середовище для подальшого тематичного структурування. Унаслідок цього попередню обробку доцільно розглядати не як суто технічний етап, а як одну з методичних передумов підвищення якості моделювання [3, 7, 9, 10].

Не менш важливою є потреба у врахуванні контексту. У традиційних моделях, побудованих на принципі «мішка слів», семантичні зв’язки між словами істотно спрощуються, через що модель гірше відображає змістові нюанси освітніх текстів. Для корпусів, присвячених використанню штучного інтелекту в дистанційній освіті, це особливо критично, оскільки одна й та сама лексика може належати до різних тематичних площин: педагогічної, технологічної, етичної або організаційної. Саме тому сучасні підходи дедалі частіше спираються на контекстні векторні представлення текстів, які дозволяють точніше відображати семантичну близькість документів і формувати більш когерентні тематичні кластери [5, 7].

Ще однією передумовою є необхідність переходу від жорстко параметризованих моделей до підходів, у яких тематична структура виводиться з геометрії простору ембедингів. Аналіз поведінки базових моделей показує, що якість тематичного моделювання суттєво залежить від параметризації, а вибір кількості тем не є тривіальним. Це зумовлює зростання інтересу до BERTopic-подібних конвеєрів, які поєднують контекстні представлення текстів, зниження розмірності за допомогою Uniform Manifold Approximation and Projection (UMAP) та кластеризацію Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), після чого тематичні описи уточнюються через c-TF-IDF [7, 9, 10].

Додатково залежність якості тематичного моделювання від вибору кількості тем доцільно проілюструвати на прикладі когерентності NPMI для LDA та NMF (рис. 1). Саме така залежність показує, що параметризація класичних моделей не має універсального значення і потребує окремого обґрунтування для конкретного корпусу. У цьому контексті важливо, що зміна кількості тем безпосередньо впливає на змістову узгодженість отриманих тематичних структур, а отже – і на їх подальшу інтерпретованість.

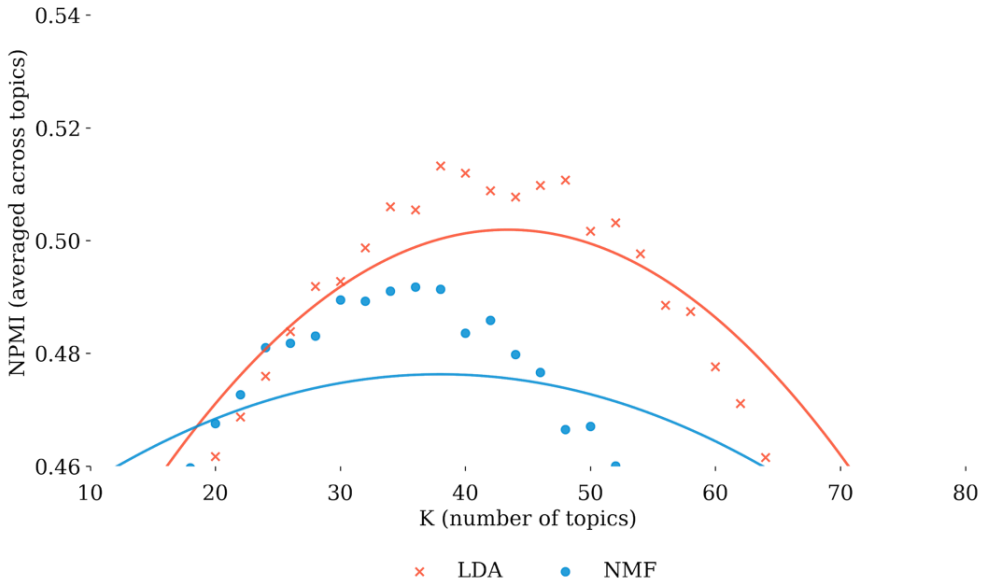

Рисунок 1 – Когерентність (NPMI) проти кількості тем (LDA проти NMF)

Як видно з рис. 1, когерентність тематичних моделей змінюється залежно від кількості тем, причому характер цієї залежності для LDA та NMF є різним. Це підтверджує, що вибір кількості тем не може розглядатися як суто технічне налаштування, оскільки він безпосередньо впливає на якість тематичної репрезентації корпусу. Для задач аналізу текстових матеріалів, присвячених використанню штучного інтелекту в дистанційній освіті, така чутливість до параметризації є особливо важливою, оскільки неоднорідність і міждисциплінарність корпусу ускладнюють застосування фіксованих конфігурацій. Саме це додатково обґрунтовує доцільність переходу до підходів, у яких тематична структура визначається не лише параметрами моделі, а й контекстною близькістю документів та їх просторовою організацією.

Отже, чутливість класичних тематичних моделей до параметризації додатково підтверджує доцільність переходу до підходів, у яких тематична структура визначається не лише наперед заданими параметрами, а й контекстною близькістю документів та їх просторовою організацією. Така архітектура виявляється особливо продуктивною для коротких і різнорідних корпусів, оскільки дозволяє уникнути частини обмежень, властивих класичним тематичним моделям [3, 7, 8].

Узагальнення характеристик, пов’язаних зі стабільністю, масштабованістю та чутливістю до гіперпараметрів, подано в табл. 3.

Таблиця 3 – Стабільність, масштабованість та чутливість до гіперпараметрів

Критерій	LDA	NMF	c-TF-IDF (BERTopic)
Стабільність / Робастність	Схильна до варіацій від ініціалізації; потрібна оцінка стабільності (перезапуски/бутстреп)	Залежить від рангу K та ініціалізації W, H	Залежить від якості ембедингів і параметрів UMAP/HDBSCAN; можливі дрібні кластери / «викиди», потрібна агрегація [7, 9, 10]
Ефективність / Масштабованість	Існують online-варіанти та наближені схеми інференції, але якість чутлива до налаштувань [1, 2]	Добре векторизується та зручна для матричних обчислень [4, 6]	Найбільші витрати припадають на побудову ембедингів, UMAP і HDBSCAN; побудова c-TF-IDF є відносно дешевою [7, 9, 10]
Чутливість до гіперпараметрів	Висока: кількість тем і параметри апріорних розподілів [1, 3]	Висока: ранг K та початкова ініціалізація [4, 6]	Дуже висока: вибір моделі ембедингів та параметрів UMAP/HDBSCAN [7, 9, 10]

Як видно з табл. 3, проблема вдосконалення тематичного моделювання полягає не лише в підвищенні когерентності тем, а й у забезпеченні їх стійкості, масштабованості та керованості. Для задач дистанційної освіти це особливо важливо, оскільки аналізовані корпуси можуть бути як короткими й фрагментованими, так і часово протяжними, коли потрібно фіксувати зміни тематичної структури в динаміці [3, 5, 8].

Отже, сукупність виявлених обмежень і вимог формує логічні передумови для переходу до гібридної методики тематичного моделювання, а далі – до формалізованих контекстуально-динамічних рішень.

Саме в цій логіці доцільно розглядати фреймворк DCTM-LLM як подальший етап розвитку засобів аналізу текстових корпусів, присвячених використанню штучного інтелекту в дистанційній освіті [5, 7, 8].

Формалізований фреймворк динамічного контекстуального тематичного моделювання

Гібридна методика на основі інформованих пріорів для LDA підвищує змістову релевантність тематичних структур, проте не усуває повністю обмежень статичного тематичного моделювання. У зв’язку з цим доцільним є перехід до фреймворку DCTM-LLM, який поєднує контекстне представлення текстів, динамічне тематичне структурування та семантичне узагальнення результатів [5, 11-14]. Цей підхід орієнтований на аналіз часово-мічених текстових корпусів, у яких важливо не лише виявити змістові напрями, а й простежити їх еволюцію в часі.

Перш за все формалізується постановка задачі. Нехай задано часово-мічений текстовий корпус:

$$D = \{(d_i, t_i)\}_{i=1}^{N}, \quad (1)$$

де d_i – текст і-го документа;

t_i – його часова мітка;

N – загальна кількість документів у корпусі.

Метою є побудова множини динамічних тематичних траєкторій:

$$\mathrm{T} = \{\mathrm{T}_m\}_{m=1}^{M}, \tag{2}$$

де T_m – m-та динамічна траєкторія, яка описує еволюцію окремого тематичного напряму в часі;

M – загальна кількість таких траєкторій.

Кожна траєкторія надалі доповнюється короткою назвою, розгорнутим наративним описом і набором ключових термінів, що характеризують її зміст на різних часових інтервалах. У такій постановці тема розглядається не як ізольований кластер, а як послідовність змістових станів, пов'язаних часовою логікою розвитку.

Концептуально DCTM-LLM реалізується як послідовність взаємопов'язаних етапів: попередня обробка корпусу, побудова контекстних векторних представлень документів, тематичне структурування із застосуванням процедур зниження розмірності та кластеризації, формування лексичних профілів тем і подальше семантичне узагальнення результатів. Така організація дає змогу перейти від статичного переліку ключових слів до динамічних тематичних траєкторій, що відображають розвиток змістових ліній у часі [7, 9, 10, 15, 16]. Особливість цього підходу полягає в тому, що контекстуальна близькість документів визначається не лише частотними характеристиками слів, а й їхнім розміщенням у семантичному просторі, сформованому на основі сучасних мовних моделей.

Структурну логіку реалізації запропонованого підходу доцільно подати у вигляді схеми, що відображає перехід від вхідного корпусу до інтерпретованих тематичних траєкторій і наративних описів (рис. 2). У межах такого конвеєра попередньо оброблений корпус трансформується в множину контекстних векторних представлень, на основі яких формуються локальні тематичні утворення, що далі пов'язуються між собою в часовій перспективі. Результатом є не лише тематичне структурування документів, а й можливість простежити зміни змістових акцентів у межах окремих тематичних напрямів.

Як показує рис. 2, запропонований підхід реалізує завершений аналітичний конвеєр: від попередньо обробленого корпусу та побудови його контекстного представлення – до формування динамічних тематичних утворень, їх лексичної репрезентації та автоматизованого синтезу наративів засобами великої мовної моделі. Саме така структурна організація дозволяє поєднати контекстну чутливість, часовий аналіз і подальше інтерпретаційне узагальнення.

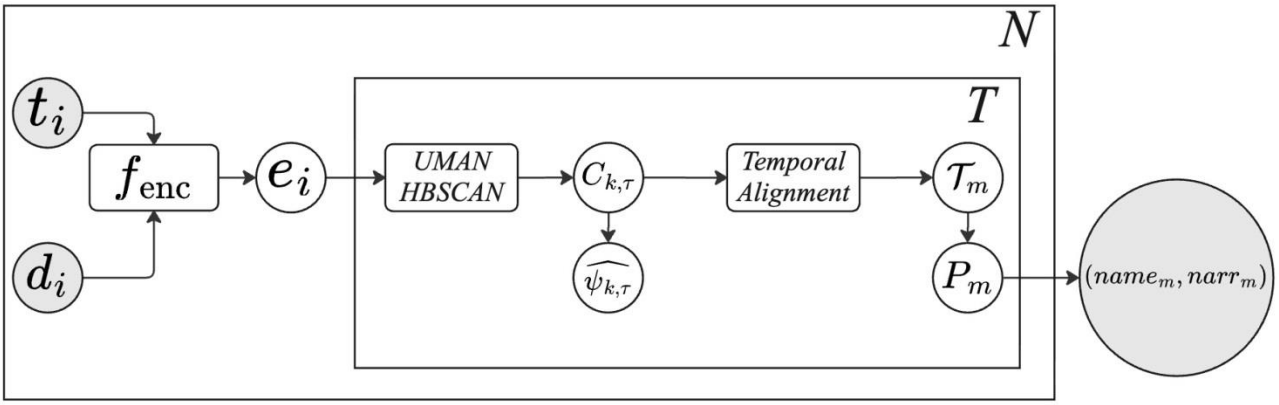

Рисунок 2 – Структурна схема реалізації методу DCTM-LLM

Структурна схема реалізації методу DCTM-LLM на рисунку 2 відображає послідовність трансформації даних від вхідного корпусу (d_i, t_i) до отримання інтерпретованих тематичних траєкторій T_m та наративних описів $(name_m, narr_m)$/

Особливістю цього підходу є те, що велика мовна модель (LLM) використовується не як заміна тематичного моделювання, а як засіб інтерпретаційного доопрацювання вже виявлених тематичних структур. Генеративний етап можна подати як відображення:

$$(name_m, narr_m) = g_\Theta(P_m), \tag{3}$$

де $name_m$ – коротка назва m-тої тематичної траєкторії;

$narr_m$ – її наративний опис;

P_m – сформований для неї prompt;

g_Θ – відображення типу «prompt → текстова відповідь», реалізоване великою мовною моделлю з фіксованими параметрами Θ.

Унаслідок цього формальні кількісні характеристики тем трансформуються в людиноінтерпретовані описи, придатні для подальшого аналітичного тлумачення [13, 14]. Отже, фреймворк поєднує алгоритмічну строгість тематичного структурування з інтерпретаційною гнучкістю генеративних моделей.

З методичного погляду DCTM-LLM забезпечує поєднання трьох взаємопов'язаних вимірів аналізу: контекстної чутливості, часової динаміки та семантичного узагальнення. Це дозволяє істотно розширити аналітичні можливості тематичного моделювання для корпусів, присвячених використанню штучного інтелекту в дистанційній освіті, де змістові акценти можуть швидко змінюватися під впливом розвитку технологій, освітніх практик і суспільних очікувань [5, 11, 12]. Саме така інтеграція створює підґрунтя для подальшого розгляду переваг, інноваційних відмінностей і результатів оцінювання запропонованої методики.

Переваги, інноваційні відмінності та оцінка результатів методики DCTM-LLM

Переваги фреймворку DCTM-LLM доцільно розглядати в порівнянні з підходами, що вже були проаналізовані в попередніх підрозділах. На відміну від статичних тематичних моделей, запропонована методика поєднує контекстне представлення текстів, динамічне структурування тем і наративне узагальнення результатів. Саме така інтеграція визначає її головну інноваційну відмінність: тема в межах DCTM-LLM постає не як статичний набір ключових слів, а як динамічна тематична траєкторія, що дозволяє простежувати еволюцію тематичних ліній у часі [5, 7, 8, 11].

Для оцінювання методики застосовано комплекс взаємодоповнювальних показників, які відображають як якість тематичної структури, так і рівень інтерпретованості результатів. До них належать когерентність тем як базовий критерій змістової узгодженості, структурні метрики кластеризації для оцінювання відокремленості та стабільності тематичних утворень, Topic Diversity@10 для вимірювання лексичної унікальності топ-термінів, а також метрика BERTScore F1 для оцінювання семантичної відповідності згенерованих наративів референтним описам. Така багатокритеріальна рамка дає змогу аналізувати не лише формальну якість тематичного моделювання, а й аналітичну придатність отриманих результатів. Важливо, що вибір саме такої системи показників дозволяє оцінювати DCTM-LLM одночасно в кількох вимірах. Якщо когерентність і Topic Diversity@10 характеризують якість тематичного представлення та лексичного розмежування тем, то Silhouette і ARI відображають структурну чіткість і стабільність кластеризації. Своєю чергою BERTScore F1 виводить оцінювання за межі суто статистичних характеристик і дає змогу врахувати, наскільки згенеровані наративні описи справді зберігають семантичний зміст тематичних утворень. У підсумку така система метрик є релевантною саме для DCTM-LLM, оскільки цей фреймворк поєднує структурне тематичне моделювання з інтерпретаційним узагальненням результатів.

Узагальнені результати порівняння DCTM-LLM із Dynamic BERTopic та гібридним підходом LDA+NMF+c-TF-IDF наведено в табл. 4.

Таблиця 4 – Показники якості моделей тематичного моделювання (корпус cs.AI, 2015-2025)

Метрика	Dynamic BERTopic	LDA+NMF+c-TF-IDF	DCTM-LLM
NPMI ↑	0.48	0.51	0.53
Silhouette ↑	0.59	0.61	0.62
ARI ↑	0.45	0.52	0.55
Topic Diversity@10 ↑	0.82	0.85	0.88
BERTScore F1 ↑	0.62	0.65	0.89

Примітка: стрілкою ↑ позначено метрики, для яких більше значення означає кращу якість

Як видно з табл. 4, DCTM-LLM перевищує альтернативні підходи як за структурними, так і за семантичними показниками. Це означає, що запропонована методика не лише формує більш узгоджені тематичні структури, а й забезпечує вищу якість їх інтерпретаційного представлення. Особливо

показовою є різниця за BERTScore F1, що свідчить про суттєве покращення наративного узагальнення результатів.

Такий розподіл результатів дозволяє зробити важливий методичний висновок: перевага DCTM-LLM не зводиться до покращення лише одного окремого аспекту аналізу. Навпаки, запропонований підхід демонструє більш збалансовану якість у межах усієї системи оцінювання. Підвищення NPMI свідчить про кращу змістову зв'язність тем, вищі значення Silhouette та ARI - про якісніше групування документів і стабільніше відтворення тематичних утворень, а зростання Topic Diversity@10 та BERTScore F1 - про більш виразну й аналітично корисну інтерпретацію результатів. Саме ця сукупність ознак дозволяє розглядати DCTM-LLM як більш придатний інструмент для дослідження складних міждисциплінарних корпусів.

Для кількісної оцінки внеску окремих компонентів фреймворку було проведено абляційний аналіз. Його результати щодо структурної якості кластеризації подано в табл. 5.

Таблиця 5 – Вплив компонентів DCTM-LLM на показники структурної кластеризації

Configuration	Noise	Silhouette	ARI
DCTM-LLM	8.5%	0.62	0.55
no-LLM-refinement	8.5%	0.60 (Δ -0.02)	0.52 (Δ -0.03)
no-UMAP	14.2%	0.40 (Δ -0.22)	0.35 (Δ -0.20)
HDBSCAN → k-means	0.0%	0.54 (Δ -0.08)	0.48 (Δ -0.07)
c-TF-IDF → TF-IDF	8.5%	0.62 (Δ 0.00)	0.55 (Δ 0.00)

Таблиця 5 показує, що найбільш критичний вплив на структурну якість моделі мають компоненти, пов'язані з геометрією простору та кластеризацією. Відмова від UMAP спричиняє найвідчутніше погіршення структурних показників, а заміна HDBSCAN на k-means також знижує якість відокремлення тематичних утворень. Це підтверджує, що блок просторового подання та кластеризації є одним із визначальних для стабільності DCTM-LLM. Водночас вилучення LLM-refinement не руйнує структуру кластерів, але знижує загальну узгодженість результатів, що особливо помітно на рівні їх інтерпретації. Іншими словами, результати абляційного аналізу дозволяють розмежувати внесок окремих компонентів фреймворку. UMAP і HDBSCAN у більшій мірі забезпечують геометричну та структурну якість тематичного простору, тоді як LLM-refinement істотно впливає на завершальний етап інтерпретації та подання результатів. У цьому контексті DCTM-LLM доцільно розглядати не як механічне поєднання кількох інструментів, а як узгоджену систему, у якій кожен компонент відповідає за окремий рівень аналітичної якості. Саме така багаторівнева організація і пояснює, чому вилучення будь-якого з елементів фреймворку позначається на загальному результаті.

Семантичний внесок окремих компонентів відображено в табл. 6.

Таблиця 6 – Вплив компонентів DCTM-LLM на семантичні метрики та якість інтерпретації

Configuration	NPMI	Topic Diversity@10	BERTScore F1
DCTM-LLM	0.53	0.88	0.89
no-LLM-refinement	0.51 (Δ -0.02)	0.86 (Δ -0.02)	0.62 (Δ -0.27)
no-UMAP	0.45 (Δ -0.08)	0.80 (Δ -0.08)	0.75 (Δ -0.14)
HDBSCAN → k-means	0.49 (Δ -0.04)	0.83 (Δ -0.05)	0.81 (Δ -0.08)
c-TF-IDF → TF-IDF	0.44 (Δ -0.09)	0.81 (Δ -0.07)	0.80 (Δ -0.09)

Як видно з табл. 6, найбільш різке погіршення BERTScore F1 відбувається при вилученні LLM-refinement. Це означає, що саме цей компонент забезпечує головний внесок у формування якісних наративних описів і перехід від списків ключових термінів до змістовно зв'язаних тематичних інтерпретацій. Водночас відмова від UMAP і c-TF-IDF негативно впливає на когерентність і тематичну різноманітність, що підтверджує їхню важливість для структурної та лексичної організації моделі. Таким чином, аналітична цінність DCTM-LLM формується не одним окремим елементом, а сукупною дією взаємопов'язаних компонентів. У цьому контексті інтерпретованість результатів DCTM-LLM доцільно розглядати як інтегральну властивість фреймворку, що виникає внаслідок послідовного поєднання кількох рівнів аналітичної обробки. Формування тематичної структури, її просторово-кластерна організація та подальше наративне узагальнення утворюють єдиний ланцюг, у межах якого кожен етап впливає на якість кінцевої інтерпретації. З огляду на це зниження BERTScore F1 у разі вилучення LLM-refinement слід тлумачити як індикатор послаблення інтерпретаційної спроможності всього підходу, а не як ізольовану зміну окремого показника. Для аналізу текстових корпусів, присвячених використанню штучного інтелекту в дистанційній освіті, така властивість є принципово важливою, оскільки забезпечує можливість переходу від формального тематичного структурування до змістовно обґрунтованих наукових висновків.

Однією з ключових переваг методики DCTM-LLM є можливість відображення зміни змістових акцентів у межах окремого тематичного напряму в часовій перспективі. На відміну від статичного тематичного моделювання, такий підхід дозволяє працювати не лише з фіксованим набором ключових слів, а й з послідовною еволюцією тематичного ядра. Це особливо важливо для аналізу наукового дискурсу щодо використання штучного інтелекту в освіті, де дослідницькі пріоритети змінюються під впливом розвитку технологій, методологічних підходів та суспільних запитів. Така аналітична можливість ілюструється на прикладі тематики Explainable Artificial Intelligence (XAI), у межах якої простежується поступове зміщення уваги від локальних інструментів інтерпретації до каузальних, контрфактичних і етично орієнтованих підходів (рис. 3).

Як показано на рис. 3, динамічне тематичне моделювання дозволяє виявити не лише наявність окремих змістових ліній у межах XAI, а й зміну їх відносної ваги в часі. Якщо на ранніх етапах домінують локальні підходи до

інтерпретації, пов’язані з LIME та SHAP, то в подальшому посилюється увага до візуалізації механізмів transformer-моделей, каузальних і контрфактичних пояснень, а також до проблематики справедливості й етичного аудиту штучного інтелекту.

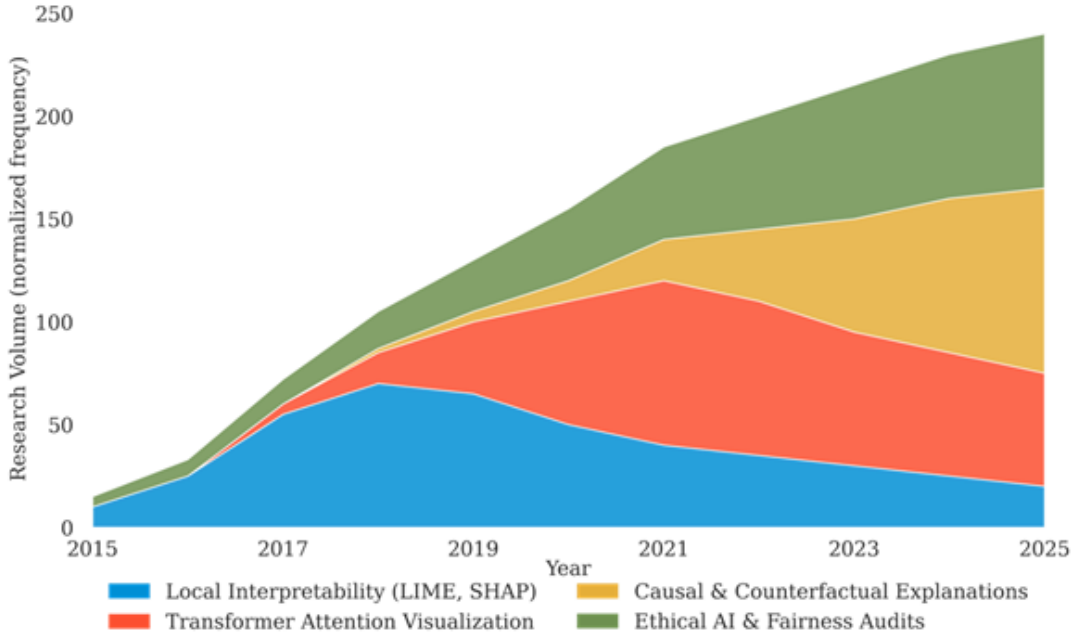

Рисунок 3 – Тематична еволюція «Пояснюваного штучного інтелекту (XAI)» з часом

Це підтверджує, що DCTM-LLM є придатним не лише для тематичного структурування корпусу, а й для виявлення діахронічної еволюції дослідницьких акцентів. У цьому виявляється його принципова методична перевага над статичними моделями, які, як правило, фіксують лише синхронний тематичний зріз корпусу. Натомість DCTM-LLM уможливлює перехід до аналізу тем як динамічних утворень, внутрішня логіка яких розкривається через зміну домінантних понять, інтерпретаційних рамок і дослідницьких пріоритетів у часовій перспективі. Для вивчення використання штучного інтелекту в дистанційній освіті це має особливу аналітичну цінність, оскільки дозволяє не лише констатувати наявність певних тематичних напрямів, а й виявляти закономірності їх розвитку, переорієнтації та змістового ускладнення. Такий підхід істотно розширює інтерпретаційний потенціал тематичного моделювання, переводячи його з рівня опису до рівня змістового пояснення еволюції предметної області.

Узагальнюючи результати, подані в табл. 5-6, можна зробити висновок, що всі базові компоненти фреймворку – UMAP, HDBSCAN, c-TF-IDF і LLM-refinement – роблять істотний і взаємодоповнювальний внесок у якість моделі. Вилучення будь-якого з них призводить до погіршення або структурних, або семантичних показників, тоді як повна конфігурація DCTM-LLM забезпечує найкращий баланс між когерентністю, різноманітністю та інтерпретованістю результатів. Саме це дозволяє розглядати DCTM-LLM як методику, що формує вищий рівень аналітичної придатності для дослідження текстових корпусів, присвячених використанню штучного інтелекту в дистанційній освіті.

Висновки

Обґрунтовано доцільність використання тематичного моделювання як методу аналізу текстових корпусів, у яких відображено використання штучного інтелекту в дистанційній освіті. Показано, що для цієї предметної області характерними є короткі та неоднорідні тексти, міждисциплінарна термінологія, контекстуальна варіативність і часова динаміка змісту, що ускладнює застосування класичних тематичних моделей без додаткового методичного вдосконалення. Це дає підстави розглядати тематичне моделювання в даному контексті не як універсальний інструмент, а як аналітичний підхід, ефективність якого безпосередньо залежить від урахування специфіки корпусу, характеру тематичної структури та вимог до інтерпретованості результатів.

У результаті порівняльного аналізу встановлено, що LDA, NMF і c-TF-IDF репрезентують різні підходи до тематичного моделювання та мають різні аналітичні можливості. LDA зберігає значення як класична імовірнісна модель, але виявляє обмеження при роботі з короткими й розрідженими текстами. NMF формує більш сфокусовані тематичні профілі, однак залишається чутливою до ініціалізації та параметризації. c-TF-IDF у поєднанні з embedding-based підходами є особливо придатним до аналізу коротких і неоднорідних текстів, хоча залежить від якості попередньої кластеризації. У підсумку показано, що жоден із зазначених підходів окремо не забезпечує повного врахування особливостей текстових корпусів, присвячених використанню штучного інтелекту в дистанційній освіті, що зумовлює потребу в подальшому розвитку методичних засобів тематичного аналізу.

На основі виявлених обмежень обґрунтовано доцільність переходу до гібридної методики на основі інформованих пріорів, а далі - до фреймворку DCTM-LLM. Запропонований підхід поєднує контекстне представлення текстів, динамічне тематичне структурування та автоматизоване наративне узагальнення, що дозволяє перейти від статичних тематичних описів до інтерпретованих тематичних траєкторій. Його методична перевага полягає в тому, що тематична структура корпусу розглядається не як фіксований результат одноразового моделювання, а як динамічне утворення, зміст якого може змінюватися в часовій перспективі. Це розширює аналітичні можливості дослідження та створює підґрунтя для більш змістового осмислення еволюції тематичних пріоритетів у відповідній предметній області.

Узагальнення результатів оцінювання показало, що DCTM-LLM забезпечує кращий баланс між когерентністю, структурною стабільністю, різноманітністю тем і якістю інтерпретації, ніж менш інтегровані підходи. Абляційний аналіз підтвердив, що всі ключові компоненти фреймворку – UMAP, HDBSCAN, c-TF-IDF і LLM-refinement – роблять істотний і взаємодоповнювальний внесок у якість моделі. Це свідчить про те, що аналітична результативність DCTM-LLM формується не окремими локальними покращеннями, а узгодженою взаємодією кількох рівнів обробки даних:

структурного, просторово-кластерного та інтерпретаційного. З огляду на це запропоновану методику доцільно розглядати як перспективний інструмент аналізу текстових корпусів, присвячених використанню штучного інтелекту в дистанційній освіті, який забезпечує не лише формальне виявлення тематичних структур, а й створює умови для їх подальшого наукового осмислення.

Список літератури.

1. Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, (3), 993-1022.

2. Blei, D.M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84. https://doi.org/10.1145/2133806.2133826.

3. Fan, Y., Shi, L., & Yuan, L. (2023). Topic modeling methods for short texts: A survey. Journal of Intelligent and Fuzzy Systems, 45(2), 1971-1990. https://doi.org/10.3233/JIFS-223834.

4. Gillis, N. (2020). Nonnegative matrix factorization. SIAM. https://doi.org/10.1137/1.9781611976410.

5. Wu, X., Nguyen, T., & Luu, A. T. (2024). A survey on neural topic models: Methods, applications, and challenges. Artificial Intelligence Review, 57(2), 18. https://doi.org/10.1007/s10462-023-10661-7.

6. Lee, D.D., & Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, (401), 788-791. https://doi.org/10.1038/44565.

7. Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794. https://arxiv.org/abs/2203.05794.

8. Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. Frontiers in Sociology, (7), Article 886498. https://doi.org/10.3389/fsoc.2022.886498.

9. McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. Journal of Open Source Software, 3(29), 861. https://doi.org/10.21105/joss.00861.

10. McInnes, L., & Healy, J. (2017). Accelerated hierarchical density clustering. arXiv preprint arXiv:1705.07321. https://arxiv.org/abs/1705.07321.

11. Wu, X., Dong, X., Pan, L., Nguyen, T., & Luu, A. T. (2024). Modeling dynamic topics in chain-free fashion by evolution-tracking contrastive learning and unassociated word exclusion. Findings of the Association for Computational Linguistics: ACL 2024, 3088-3105. https://doi.org/10.18653/v1/2024.findings-acl.183.

12. Fil, N. Yu., Slisarenko, R.V., Deineko, Zh.V., & Morozova, L.Yu. (2025). Trends in artificial intelligence research on education: Topic modeling using latent Dirichlet allocation. Bulletin of Kharkiv National Automobile and Highway University, (108), 17-24. https://doi.org/10.30977/BUL.2219-5548.2025.108.0.17.

13. Khandelwal, T. (2025). Using LLM-based approaches to enhance and automate topic labeling. arXiv preprint arXiv:2502.18469. https://arxiv.org/abs/2502.18469.

14. Jenner, S., Raidos, D., Anderson, E., Fleetwood, S., Ainsworth, B., Fox, K., Kreppner, J., & Barker, M. (2025). Using large language models for narrative analysis: A novel application of generative AI. Methods in Psychology, (12), 100183. https://doi.org/10.1016/j.metip.2025.100183.

15. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805.

16. Logan, C.H.A., & Fotopoulou, S. (2020). Unsupervised star, galaxy, QSO classification: Application of HDBSCAN. Astronomy and Astrophysics, (633), A154. https://doi.org/10.1051/0004-6361/201936648.