



КРИТЕРІЇ АНАЛІТИЧНОЇ ПРИДАТНОСТІ ТЕМАТИЧНИХ МОДЕЛЕЙ ДЛЯ АНАЛІЗУ ТЕКСТОВИХ КОРПУСІВ У ДИСТАНЦІЙНІЙ ОСВІТІ

Дейнеко Ж.В., зав. кафедри МСТ, ХНУРЕ
Слісаренко Р.В., аспірант, кафедра МСТ, ХНУРЕ

Abstract. *The study examines the analytical suitability of topic modeling methods for exploring text corpora that reflect the use of artificial intelligence in distance education. It is shown that the effectiveness of topic modeling in this domain should be assessed not only by the ability to identify topics, but also by coherence, interpretability, robustness to parameter settings, and applicability to short and heterogeneous texts. The paper argues that the analytical value of topic models in this field depends on their ability to support both thematic structuring and the subsequent interpretation of results in educational and scientific contexts.*

Keywords: *topic modeling, distance education, artificial intelligence in education, text corpora, analytical suitability, interpretability.*

Сучасний розвиток дистанційної освіти супроводжується зростанням обсягу текстових матеріалів, у яких відображаються напрями, проблеми та результати використання штучного інтелекту в освітньому середовищі. До таких матеріалів належать наукові публікації, аналітичні огляди, описи цифрових платформ, результати опитувань, відкриті відповіді учасників освітнього процесу та тексти, пов'язані з питаннями академічної доброчесності й інформаційної безпеки [1, 2]. Тематичне моделювання доцільно розглядати не лише як інструмент автоматизованого виявлення тем, а як засіб аналітичного структурування предметної області, у межах якої поєднуються педагогічні, технологічні, етичні та організаційні аспекти. Такий підхід є особливо важливим для освітньої сфери, де текстові дані не лише фіксують результати впровадження нових технологій, а й відображають зміни в уявленнях про роль викладача, здобувача освіти та цифрових інструментів у навчальному процесі.

Метою роботи є визначення критеріїв аналітичної придатності тематичних моделей для аналізу текстових корпусів, присвячених використанню штучного інтелекту в дистанційній освіті. Особливу увагу приділено тому, наскільки різні підходи до тематичного моделювання забезпечують когерентність, інтерпретованість, стійкість результатів та придатність до аналізу коротких і неоднорідних текстів.

Для корпусів такого типу характерні міждисциплінарність, термінологічна варіативність, нерівномірність довжини документів і виражена часова динаміка змісту. Це ускладнює застосування класичних тематичних моделей без додаткового методичного обґрунтування. Зокрема, Latent Dirichlet Allocation (LDA) зберігає значення як базова імовірнісна модель, однак виявляє обмеження при роботі з короткими та розрідженими текстами. Non-negative Matrix Factorization (NMF) дозволяє отримувати більш сфокусовані тематичні профілі, але залишається чутливою до ініціалізації та параметризації. Натомість class-based Term Frequency–Inverse Document Frequency (c-TF-IDF) у поєднанні з контекстними векторними представленнями текстів демонструє вищу



інтерпретованість і кращу придатність до аналізу коротких та неоднорідних корпусів [1, 3]. Це свідчить про те, що в освітніх текстових корпусах вирішальне значення має не лише механізм виокремлення тем, а й здатність моделі зберігати змістові зв'язки між фрагментованими, термінологічно варіативними та контекстно залежними висловлюваннями.

У цьому контексті тематичні моделі доцільно оцінювати не лише за формальною здатністю виокремлювати теми, а й за сукупністю критеріїв аналітичної придатності. До таких критеріїв належать когерентність тематичних описів, стійкість результатів до параметризації, здатність працювати з короткими текстами, рівень людиноцентричної інтерпретованості та можливість відображення часової динаміки тематичної структури. Саме така постановка дозволяє перейти від формального порівняння окремих моделей до змістовного аналізу того, наскільки отримані результати придатні для подальшого наукового узагальнення. Інакше кажучи, аналітична придатність тематичної моделі визначається не лише точністю алгоритмічного поділу корпусу, а й тим, наскільки результати цього поділу можуть бути змістовно інтерпретовані в межах конкретної предметної області.

Окремого значення набуває застосування тематичного моделювання до текстів, у яких висвітлено етичні, педагогічні та організаційні аспекти використання штучного інтелекту в дистанційній освіті. Тематична модель має не лише фіксувати наявність відповідних змістових кластерів, а й забезпечувати можливість подальшого інтерпретаційного аналізу змін дослідницьких акцентів, пов'язаних із безпечним, відповідальним і педагогічно обґрунтованим використанням засобів штучного інтелекту [2]. Це особливо важливо в умовах, коли обговорення штучного інтелекту в освіті дедалі частіше виходить за межі суто технологічної проблематики та охоплює питання академічної доброчесності, довіри до автоматизованих рішень, етичних меж використання генеративних систем і педагогічної доцільності їх інтеграції.

Таким чином, для дослідження текстових корпусів, присвячених використанню штучного інтелекту в дистанційній освіті, тематичне моделювання доцільно розглядати як багаторівневий аналітичний інструмент. Його результативність має визначатися не лише точністю тематичної сегментації, а й здатністю формувати змістовно узгоджені, інтерпретовані та аналітично придатні тематичні структури. Це створює підґрунтя для подальшої розробки методик аналізу освітніх текстових корпусів з урахуванням їхньої контекстної, динамічної та міждисциплінарної специфіки.

Список літератури

1. 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>.
2. Слісаренко, Р., & Дейнеко, Ж. (2024). Забезпечення академічної доброчесності з урахуванням засобів штучного інтелекту в інформаційній безпеці. *Наука і техніка сьогодні*, 9(37), 842-854. [https://doi.org/10.52058/2786-6025-2024-9\(37\)-842-854](https://doi.org/10.52058/2786-6025-2024-9(37)-842-854).
3. 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>.