УДК 004.89

СТРУКТУРНИЙ АНАЛІЗ НАУКОВИХ КЛАСТЕРІВ У ДОСЛІДЖЕННЯХ AIED: ІЄРАРХІЧНА КЛАСТЕРИЗАЦІЯ ТЕМАТИЧНИХ МОДЕЛЕЙ

Слісаренко Р.В., Дейнеко Ж.В.
e-mail: roman.slisarenko@nure.ua; zhanna.deineko@nure.ua
Харківський національний університет радіоелектроніки, каф. МСТ,
м. Харків, Україна

This study explores thematic structures in Artificial Intelligence in Education research through hierarchical clustering of topic models. Using data from Web of Science, Latent Dirichlet Allocation was applied to identify dominant research themes. Cosine similarity-based clustering revealed five key topics, including AI applications in medical education, machine learning in education, and large language models in learning. Future work will compare LDA with Non-negative Matrix Factorization and enhance visualization using MDS, t-SNE, and UMAP to refine topic differentiation and thematic interpretation.

У сучасному світі штучний інтелект (ШІ) радикально трансформує життя людей, від особистих застосувань до високоорганізованих систем, охоплюючи різні сфери діяльності, зокрема й освіту. Інтеграція штучного інтелекту у сферу освіти є не лише революційним кроком для сучасного навчання, але й стратегічним рішенням для підготовки молоді до професій майбутнього [1]. Як зазначає у своєї роботі Luckin [2], штучний інтелект може стати потужним інструментом для персоналізації навчання, забезпечуючи адаптивні освітні траєкторії, які враховують індивідуальні потреби студентів.

Одним із ключових напрямів досліджень у сфері Artificial Intelligence in Education (AIED) є створення та впровадження інтелектуальних навчальних систем, що використовують алгоритмічні методи машинного навчання для персоналізації освітнього процесу. Сучасні підходи до тематичного моделювання, такі як Latent Dirichlet Allocation (LDA) та Non-negative Matrix Factorization (NMF), дозволяють аналізувати великі корпуси наукових текстів, виокремлюючи основні тренди в дослідженнях AIED.

У цьому дослідженні проведено структурний аналіз наукових кластерів у сфері AIED за допомогою ієрархічної кластеризації тематичних моделей. Вхідний корпус текстових даних було сформовано шляхом агрегування наукових публікацій із бази Web of Science, застосовуючи релевантні запити, які охоплюють фундаментальні аспекти і прикладні напрями AIED. Первинний препроцесінг текстів включав класичні методи обробки природної мови, зокрема токенізацію, лематизацію та фільтрацію стоп-слів, що забезпечило уніфіковане векторне представлення документів.

Для тематичного моделювання застосовано метод LDA, який є одним із найефективніших методів для виявлення латентних тем у великих текстових корпусах, що робить його ідеальним інструментом для аналізу наукових публікацій [4]. Визначення оптимальної кількості тем

виконувалося на основі обчислення метрик перплексії та когерентності, що дозволило сформувати стабільні та інтерпретовані тематичні групи.

Подальший аналіз структури наукових кластерів здійснювався за допомогою ієрархічної кластеризації, де в якості міри відстані використовувалася косинусна подібність. Такий підхід забезпечує точну оцінку семантичної спорідненості між тематичними векторами, отриманими на основі TF-IDF. На основі ієрархічної кластеризації та аналізу тематичних моделей було виокремлено такі ключові теми:

– «ШІ у медичній освіті» – охоплює дослідження ролі AIED у сфері медичної підготовки та охорони здоров'я (розробка симуляційних платформ, аналізу клінічних даних, створення адаптивних навчальних програм). Основні терміни: health, medical, patient, care, education, study, clinical, ethical.

– «ШІ в освіті» – включає дослідження інтеграції ШІ у навчальні процеси. Охоплюється широкий спектр застосувань, від автоматизації адміністративних процесів до розробки інтелектуальних систем оцінювання та підтримки прийняття рішень у навчальних закладах. Основні терміни: education, technology, research, artificial intelligence, development.

– «Алгоритмічні методи та машинне навчання в освіті» – аналізує алгоритмічні підходи у навчальних середовищах, зокрема нейронні мережі та інтелектуальні освітні технології. Важливість машинного навчання для створення адаптивних навчальних систем, здатних аналізувати великі обсяги даних для оптимізації навчальних результатів. Основні терміни: system, model, use, base, datum, method, algorithm, learning, intelligence, network.

– «Роль великих мовних моделей у навчанні» – розглядає використання генеративного ШІ та чат-ботів у навчальному процесі, мовні моделі для створення інтерактивних середовищ [5]. Основні терміни: chatgpt, question, response, use, study, chatbot, model, language.

– «Студентоцентричне навчання та персоналізація освітніх технологій» – присвячена адаптивному навчанню та індивідуалізованому підходу. Тема підкреслює важливість створення освітніх технологій, які адаптуються до індивідуальних потреб студентів, забезпечуючи персоналізовані навчальні траєкторії та підвищуючи залученість у навчальний процес. Основні терміни: student, learn, education, study, teaching, artificial, teacher, intelligence, learning, course.

Аналіз отриманих матриць відстаней і зв'язків свідчить про існування чітко визначених дослідницьких кластерів у сфері AIED. Зокрема, косинусна подібність вказує на високу семантичну спорідненість між Темами 1 і 2, що свідчить про їхню схожість у дослідницькому контексті. Це може бути наслідком перетинних методологічних підходів або схожих аналітичних моделей. У той же час, Тема 3 має найбільшу косинусну відстань до інших тем, що вказує на її концептуальну відособленість та потенційну унікальність у структурі AIED-досліджень (рисунок 1).

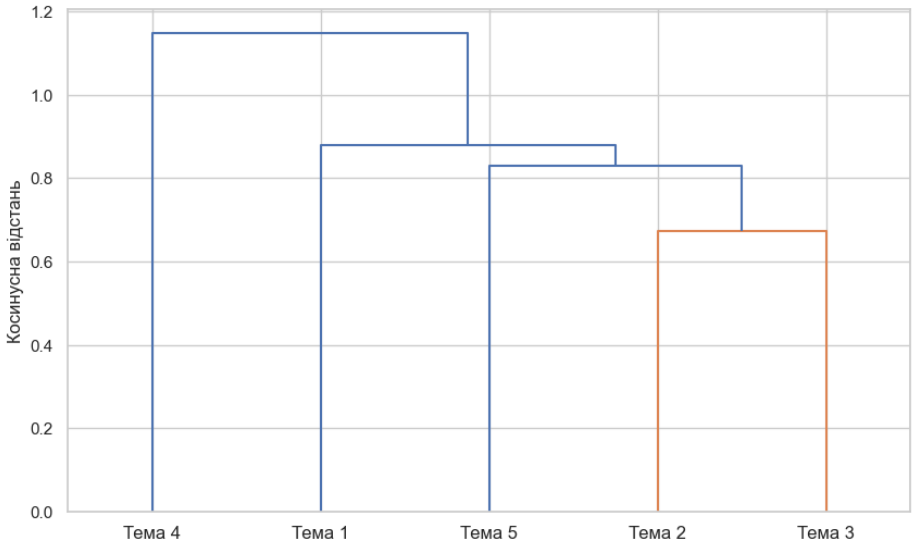

Рисунок 1 – Ієрархічна кластеризація тематичних моделей у дослідженнях AIED на основі косинусної відстані

Ідентифіковані наукові кластери дозволяють провести глибинний аналіз внутрішньої структури досліджень AIED, виявляючи домінуючі тенденції та потенційні точки публікаціях. Для покращення інтерпретації результатів було застосовано багатовимірне масштабування (MDS), t-SNE та UMAP, що дозволять точніше відобразити міжтематичні зв'язки міждисциплінарної інтеграції.

Дослідження показують, що t-SNE є ефективним методом попередньої обробки даних, покращуючи точність класифікаторів, таких як k-Nearest Neighbors (kNN) та Support Vector Classifier (SVC) [5]. UMAP краще зберігає як локальні, так і глобальні структури даних, працює швидше та є більш масштабованим. Семантично ізольовані теми можуть стати основою для подальших досліджень, зокрема таких, що потребують емпіричного аналізу та формалізації нових концептуальних підходів.

Список використаних джерел:

1. Palamar, S., & Naumenko, M. (2024). Штучний інтелект в освіті: використання без порушення принципів академічної чесності. Educological discourse, 1(44), 68-83.

2. Luckin, R. (2017). Towards artificial intelligence-based assessment systems. Nature Human Behaviour, 1(3), 1-3. https://doi.org/10.1038/s41562-016-0028.

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

4. Brown, T. B., Mann, B., Ryder, N. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

5. Кондрук, Н. Е. (2023). Аналіз технік зменшення розмірності в машинному навчанні. Науковий вісник Ужгородського університету. 42 (1), 181–187. https://doi.org/10.24144/2616-7700.2023.42(1).181-187.