УДК 655.5+004.942
DOI https://doi.org/10.32782/IT/2026-1-6

Жанна ДЕЙНЕКО
кандидат технічних наук, доцент, завідувачка кафедри медіасистем та технологій, Харківський національний університет радіоелектроніки, просп. Науки, 14, м. Харків, Україна, 61166
***ORCID:** 0000-0003-0175-4181*
***Scopus Author ID:** 57199330199*

Роман СЛІСАРЕНКО
аспірант кафедри медіасистем та технологій, Харківський національний університет радіоелектроніки, просп. Науки, 14, м. Харків, Україна, 61166, roman.slisarenko@nure.ua
***ORCID:** 0009-0009-3286-4333*

Бібліографічний опис статті: Дейнеко, Ж., Слісаренко, Р. (2026). Тематичне моделювання як інструмент аналізу сприйняття текстових корпусів у дистанційному навчанні. *Information Technology: Computer Science, Software Engineering and Cyber Security*, doi: https://doi.org/10.32782/IT/2026-1-6

ТЕМАТИЧНЕ МОДЕЛЮВАННЯ ЯК ІНСТРУМЕНТ АНАЛІЗУ СПРИЙНЯТТЯ ТЕКСТОВИХ КОРПУСІВ У ДИСТАНЦІЙНОМУ НАВЧАННІ

У роботі розглянуто застосування тематичного моделювання (LDA, NMF, BERTopic) для аналізу освітніх текстових корпусів дистанційного навчання, що характеризуються різножанровістю, контекстною варіативністю та значною часткою коротких повідомлень. Виконано порівняльне дослідження моделей за уніфікованим протоколом підготовки даних і метриками якості/стабільності; виявлені відмінності узагальнено як компроміси між зрозумілістю тем, робастністю та масштабованістю.

***Мета роботи.** Методологічно обґрунтувати придатність LDA, NMF і BERTopic з c-TF-IDF для аналізу корпусів дистанційного навчання та систематизувати критерії їх порівняння з акцентом на відтворюваність і чутливість до налаштувань на коротких гетерогенних текстах.*

***Методологія.** Застосовано уніфікований препроцесинг із контролем словника та два типи подань: «документ-термін» для LDA/NMF і документні ембеддинги для BERTopic. Якість оцінено за NPMI, UMass, Diversity@10, Silhouette, ARI та середньою JSD; стабільність перевірено повторними перезапусками з фіксацією випадковості.*

***Наукова новизна.** Запропоновано та апробовано відтворюваний багатокритеріальний протокол порівняння генеративних, факторизаційних та embedding-орієнтованих моделей для освітнього контенту. У проведених експериментах BERTopic переважно демонструє вищу когерентність, тоді як NMF – кращу стабільність.*

***Висновки.** Результати порівняння засвідчили відсутність універсальної тематичної моделі для коротких гетерогенних освітніх повідомлень, тому вибір має ґрунтуватися на відтворюваному протоколі та контролі стабільності. BERTopic забезпечив найвищу структурну узгодженість кластерів (Silhouette = 0,35), але є чутливим до параметрів embedding-конвеєра та кластеризації. LDA зберігає високе лексичне різноманіття (Diversity@10 = 0,70), проте демонструє низьку узгодженість призначень між перезапусками (ARI = 0,00). Найбільш стабільні результати отримано для NMF, що робить її доцільною за пріоритету відтворюваності. Подальші дослідження доцільно спрямувати на використання LLMs як інтерфейсу до кількісних тематичних артефактів і аналіз динаміки тем у часі.*

***Ключові слова:** дистанційне навчання, тематичне моделювання, LDA, NMF, BERTopic, c-TF-IDF, відтворюваність.*

Zhanna DEINEKO
Candidate of Technical Sciences, Associate Professor, Head of the Department of Media Systems and Technology, Kharkiv National University of Radio Electronics, 14, Nauky Ave., Kharkiv, Ukraine, 61166, zhanna.deineko@nure.ua
***ORCID:** 0000-0003-0175-4181*
***Scopus Author ID:** 57199330199*

Roman SLISARENKO
Postgraduate Student at the Department of Media Systems and Technology, Kharkiv National University of Radio Electronics, 14, Nauky Ave., Kharkiv, Ukraine, 61166, roman.slisarenko@nure.ua
***ORCID:** 0009-0009-3286-4333*

To cite this article: Deineko, Z., Slisarenko, R. (2026). Tematychne modeliuvannia yak instrument analizu spryiniattia tekstovykh korpusiv u dystantsiinomu navchanni [Topic modeling as a tool for analyzing the perception of text corpora in distance learning]. *Information Technology: Computer Science, Software Engineering and Cyber Security,* doi: https://doi.org/10.32782/IT/2026-1-6

TOPIC MODELING AS A TOOL FOR ANALYZING THE PERCEPTION OF TEXT CORPORA IN DISTANCE LEARNING

The paper examines the application of topic modeling (LDA, NMF, BERTopic) to analyze educational text corpora in distance learning, characterized by multi-genre content, contextual variability, and a substantial share of short messages. A comparative study of the models is conducted using a unified data-preparation protocol and quality/stability metrics; the observed differences are summarized as trade-offs between topic clarity, robustness, and scalability.

***Purpose of the work.** To methodologically justify the suitability of LDA, NMF, and BERTopic with c-TF-IDF for analyzing distance-learning corpora and to systematize comparison criteria with an emphasis on reproducibility and sensitivity to settings for short heterogeneous texts.*

***Methodology.** Unified preprocessing with vocabulary control is applied using two representations: document–term matrices for LDA/NMF and document embeddings for BERTopic. Quality is evaluated with NPMI, UMass, Diversity@10, Silhouette, ARI, and average JSD; stability is assessed via repeated reruns with fixed randomness.*

***Scientific novelty.** A reproducible multi-criteria protocol for comparing generative, factorization, and embedding-oriented models for educational content is proposed and tested. In the experiments, BERTopic tends to demonstrate higher coherence, whereas NMF shows better stability.*

***Conclusions.** The comparative results confirm that there is no universal topic model for short heterogeneous educational messages; therefore, model selection should rely on a reproducible protocol and explicit stability control. BERTopic achieved the highest structural consistency of clusters (Silhouette = 0.35) but is sensitive to embedding-pipeline and clustering parameters. LDA preserves high lexical diversity (Diversity@10 = 0.70) yet shows low agreement of assignments across reruns (ARI = 0.00). The most stable results were obtained for NMF, making it a suitable choice when reproducibility is a priority. Further research should focus on using LLMs as an interface to quantitative topic artifacts and on extending the analysis toward temporal topic dynamics.*

***Key words:** distance learning, topic modeling, LDA, NMF, BERTopic, c-TF-IDF, reproducibility.*

Актуальність проблеми. Системи дистанційного навчання генерують значні обсяги неструктурованих текстових даних (відкриті відповіді, форуми, коментарі до завдань, звернення до підтримки тощо). Оперативний аналіз цих повідомлень потрібний для моніторингу якості освітнього процесу, виявлення типових труднощів, оцінювання релевантності контенту та підтримки управлінських рішень в освітній аналітиці, зокрема для аналізу сприйняття освітнього контенту учасниками.

У межах роботи корпус розглядається як сукупність документів різних жанрів і довжини, об'єднаних контекстом дистанційного навчання. Гетерогенність, доменна термінологія та контекстна варіативність ускладнюють узгоджений вибір тематичної моделі й трактування тематичних сигналів як індикаторів сприйняття, а короткі та розріджені повідомлення послаблюють сигнал «мішка слів» (Bag-of-Words) (Devlin et al., 2019; Murshed et al., 2022; Fan et al., 2023; Yan et al., 2013).

Одним із поширених підходів до автоматизації тематичного аналізу є тематичне моделювання – методи виявлення латентних семантичних структур без навчання з учителем. У прикладних системах доцільно зіставляти взаємодоповнювальні підходи: Latent Dirichlet Allocation (LDA) як імовірнісний бенчмарк (Blei et al., 2003; Blei, 2012), Non-negative Matrix Factorization (NMF) як факторизаційний підхід з адитивними компонентами (Lee & Seung,

1999; Gillis, 2020) і BERTopic як embedding-орієнтований конвеєр з описом тем через class-based TF-IDF (c-TF-IDF) (Grootendorst, 2022; Egger & Yu, 2022). Водночас інженерно коректний вибір моделі вимагає відтворюваного багатокритеріального оцінювання (якість тем, стабільність, чутливість до налаштувань і витрати), оскільки оптимізація лише когерентності є недостатньою (Terragni et al., 2021; Hosseiny et al., 2024).

Аналіз останніх досліджень і публікацій. Для освітньої аналітики тематичне моделювання є базовим інструментом узагальнення змісту повідомлень і побудови інтерпретованих тематичних представлень. LDA залишається класичним імовірнісним орієнтиром, але для коротких контекстно варіативних повідомлень може поступатися у відтворюваності через обмеження Bag-of-Words та стохастичність інференції (Blei et al., 2003; Blei, 2012; Hosseiny et al., 2024). NMF часто дає читабельні адитивні теми, проте залежить від рангу та ініціалізації (Lee & Seung, 1999; Gillis, 2020). Embedding-орієнтовані конвеєри на кшталт BERTopic підсилюють роботу з короткими/гетерогенними текстами завдяки семантичним поданням і c-TF-IDF, але залишаються чутливими до параметрів конвеєра (Grootendorst, 2022; Egger & Yu, 2022). У сучасних роботах наголошується на потребі відтворюваного багатокритеріального порівняння та явного контролю стабільності, що підтримується підходами OCTIS і оглядами стабільності (Terragni et al., 2021; Hosseiny et al., 2024).

Мета дослідження. Науково обґрунтувати доцільність застосування тематичного моделювання для аналізу сприйняття текстових корпусів у дистанційному навчанні та розробити узгоджений підхід до вибору моделі, що базується на багатокритеріальному оцінюванні LDA, NMF і BERTopic з c-TF-IDF з урахуванням інтерпретованості, стабільності й обчислювальної ефективності.

Виклад основного матеріалу дослідження. Об’єктом аналізу є корпус дистанційного навчання, сформований з повідомлень різних жанрів і довжини (відкриті відповіді, коментарі, дискусії, звернення), що відображають сприйняття освітнього процесу. Гетерогенність даних визначає потребу в узгодженому конвеєрі тематичного моделювання для коротких контекстно варіативних текстів. Попередня обробка охоплює нормалізацію, токенізацію, видалення стоп-слів, лематизацію або стемінг, а також формування контрольованого словника для зниження шуму та підвищення якості тематичних представлень. Для класичних моделей використано Bag-of-Words або Term Frequency–Inverse Document Frequency (TF-IDF), тоді як для embedding-орієнтованих підходів застосовано контекстні векторні подання документів.

У дослідженні реалізовано три підходи до тематичного моделювання: імовірнісний (LDA), факторизаційний (NMF) та embedding-орієнтований (BERTopic з c-TF-IDF). LDA моделює документи як суміші тем і слугує базовим імовірнісним орієнтиром (Blei et al., 2003; Blei, 2012), NMF відновлює адитивні тематичні компоненти через факторизацію матриці термів і документів (Lee & Seung, 1999; Gillis, 2020), тоді як BERTopic інтегрує контекстні ембеддинги, зниження розмірності, кластеризацію та побудову лексичного опису тем за c-TF-IDF (Grootendorst, 2022; Egger & Yu, 2022). Поєднання цих підходів забезпечує репрезентативне охоплення різних парадигм виявлення латентних структур у текстових даних.

Для забезпечення коректності порівняння моделей використовується уніфікований експериментальний протокол, що передбачає фіксацію ключових гіперпараметрів, повторні перезапуски моделей і аналіз варіативності результатів. Це є критично важливим для тематичного моделювання, оскільки результати можуть суттєво залежати від стохастичних компонентів і початкових умов (Hosseiny et al., 2024).

Оцінювання якості тематичних моделей здійснюється в багатокритеріальній постановці з використанням метрик когерентності, різноманітності, узгодженості кластерів та стабільності між перезапусками. Такий підхід дозволяє врахувати як якість тематичних словників, так і їхню відтворюваність та інженерну придатність для використання в автоматизованих системах освітньої аналітики. Зокрема, застосування лише однієї метрики (наприклад, когерентності) не забезпечує повної оцінки якості моделей, що узгоджується з сучасними рекомендаціями щодо оцінювання тематичного моделювання (Terragni et al., 2021; Hosseiny et al., 2024).

Порівняння тематичних моделей LDA, NMF та підходу BERTopic з c-TF-IDF виконано в межах багатокритеріальної рамки оцінювання, що охоплює якість і інтерпретованість тем, придатність до гетерогенних повідомлень різної довжини, практичну інтерпретованість для аналітика/викладача (за даними порівняльних робіт), стабільність при перезапусках, обчислювальну ефективність і чутливість до гіперпараметрів. Такий формат зіставлення узгоджується зі стандартизованою логікою порівняння за множиною метрик і конфігурацій (Terragni et al., 2021) та з рекомендаціями щодо явного контролю

відтворюваності висновків у тематичному моделюванні (Hosseiny et al., 2024). Усі кількісні оцінки отримано в межах експериментального протоколу, описаного в попередньому розділі, із уніфікованим препроцесингом, контрольованим словником і серією повторних запусків.

Стислий технічний профіль моделей наведено в табл. 1 і використовується як «інженерна оптика» для інтерпретації компромісів у результатах: LDA є Bag-of-Words генеративним бенчмарком (Blei et al., 2003; Blei, 2012), NMF формує адитивні теми через невід'ємну факторизацію (Lee & Seung, 1999; Gillis, 2020), а BERTopic поєднує ембеддинги, кластеризацію та лексичний шар c-TF-IDF (Grootendorst, 2022; Egger & Yu, 2022).

Таблиця 1

Теоретична основа та принцип роботи тематичних моделей

Модель	Коротка сутність	Основний механізм
LDA	Генеративна імовірнісна модель	Документи подаються як суміші тем, теми як розподіли слів, інференція виконується наближеними методами (Blei et al., 2003; Blei, 2012).
NMF	Невід'ємна матрична факторизація	Розклад матриці «документ × термін» на невід'ємні фактори W та H рангу K з адитивною структурою тем (Lee & Seung, 1999; Gillis, 2020).
BERTopic з c-TF-IDF	Ембеддинги + кластеризація + лексичний шар	Кластеризація семантичних подань документів, далі опис тем через c-TF-IDF як ваги термінів для кластерів (Grootendorst, 2022; Egger & Yu, 2022).

Якісна інтерпретація результатів у прикладних сценаріях зводиться до такого. LDA забезпечує прозору імовірнісну інтерпретацію, однак для гетерогенних корпусів часто потребує стабілізації налаштувань і контролю словника; NMF зазвичай дає «гостріші» адитивні теми та добре працює на розріджених матрицях, але залежить від рангу та ініціалізації (Gillis, 2020); BERTopic частіше зберігає читабельність тем на різножанрових і контекстно варіативних повідомленнях завдяки семантичним ембеддингам, але чутливий до параметрів UMAP (Uniform Manifold Approximation and Projection) / HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) та політики роботи з викидами (Egger & Yu, 2022).

Таблиця 2

Кількісні метрики якості тем і стабільності

Метрика	LDA	NMF	BERTopic (c-TF-IDF)
NPMI ↑	0,19	0,18	0,24
UMass (ближче до 0 – краще)	–1,10	–1,20	–0,80
Silhouette ↑	0,33	0,30	0,35
ARI ↑	0,00	0,02	0,01
Diversity@10 ↑	0,70	0,50	0,63
Avg. JSD ↓	0,73	0,65	0,70

Примітка: для Normalized Pointwise Mutual Information (NPMI), Silhouette, Adjusted Rand Index (ARI), Diversity@10 більші значення кращі; для UMass coherence (UMass) – значення ближчі до 0; для середньої Jensen–Shannon divergence (JSD) менші значення кращі.

Отримані значення демонструють виразну структуру компромісів. BERTopic з c-TF-IDF забезпечив найвищу когерентність (NPMI = 0,24; UMass = –0,80) і максимальну структурну узгодженість (Silhouette = 0,35), що відповідає embedding-орієнтованій природі методу (Grootendorst, 2022; Egger & Yu, 2022). LDA продемонструвала найбільше різноманіття топлексики (Diversity@10 = 0,70), однак відтворюваність за кластеризацією документів залишається низькою (ARI ≈ 0), що обґрунтовує потребу в повторних запусках і контролі стабільності, а не лише в оптимізації когерентності (Hosseiny et al., 2024). NMF показала найкращу стабільність лексичних профілів (Avg. JSD = 0,65) та найвище ARI (0,02), проте поступилася за когерентністю і різноманіттям тем, що відображає компроміс між «гостротою» адитивних компонент і шириною лексичного покриття (Gillis, 2020; Hosseiny et al., 2024).

Для компактного узагальнення експлуатаційних профілів на рис. 1 наведено радар-діаграму за сукупністю критеріїв. Візуалізація відображає якісні оцінки у шкалі 1 → 5 (1 – найнижчий, 5 – найвищий рівень прояву характеристики); для чутливості до гіперпараметрів застосовано інверсію, щоб більші значення відповідали кращій керованості.

Профілі засвідчують, що BERTopic потребує контролю embedding-конвеєра та кластеризації, NMF – стабілізації рангу/ініціалізації, тоді як для LDA визначальними є повторні перезапуски та контроль налаштувань (Egger & Yu, 2022; Hosseiny et al., 2024).

Узагальнюючи, вибір тематичної моделі для аналізу сприйняття в корпусах дистанційного навчання доцільно здійснювати на основі багатокритеріальної логіки й відтворюваного протоколу порівняння. BERTopic з c-TF-IDF доцільний, коли пріоритетом є когерентність і структурна

Рис. 1. Радар-діаграма порівняння LDA, NMF та BERTopic з c-TF-IDF за сукупністю критеріїв

розділеність тем, але потребує контролю параметрів ембеддингів/кластеризації; NMF є практичним варіантом за обмежених ресурсів і підвищених вимог до стабільності, але чутлива до рангу та ініціалізації; LDA варто використовувати як прозорий імовірнісний бенчмарк, але з обов'язковими повторними перезапусками та стабілізацією налаштувань (Blei, 2012; Gillis, 2020; Hosseiny et al., 2024).

Висновки. У статті обґрунтовано використання тематичного моделювання для аналізу сприйняття текстових корпусів у дистанційному навчанні та показано, що вибір моделі в прикладній освітній аналітиці має спиратися на відтворюваний багатокритеріальний протокол, а не лише на когерентність.

Результати демонструють компроміси: BERTopic забезпечує найкращу когерентність і структурну узгодженість, але є чутливим до параметрів конвеєра; NMF дає найстабільніші лексичні профілі, однак залежить від рангу та ініціалізації; LDA зберігає прозору інтерпретацію і високе лексичне різноманіття, проте потребує ретельного налаштування й контролю стабільності між перезапусками.

Подальші дослідження доцільно спрямувати на моніторинг змін сприйняття у часі через тематичні траєкторії, автоматизацію контролю стабільності/параметричної чутливості та використання Large Language Models (LLM) як інтерфейсу для пояснення кількісних тематичних артефактів без заміни власне тематичного моделювання.

ЛІТЕРАТУРА:

1. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003. Vol. 3. P. 993–1022.

2. Blei D. M. Probabilistic topic models. Communications of the ACM. 2012. Vol. 55, No. 4. P. 77–84. DOI: https://doi.org/10.1145/2133806.2133826.

3. Lee D. D., Seung H. S. Learning the parts of objects by non-negative matrix factorization. Nature. 1999. Vol. 401. P. 788–791. DOI: https://doi.org/10.1038/44565.

4. Gillis N. Nonnegative Matrix Factorization. Philadelphia: SIAM, 2020. DOI: https://doi.org/10.1137/1.9781611976410.

5. Yan X., Guo J., Lan Y., Cheng X. A biterm topic model for short texts. Proceedings of the 22nd International Conference on World Wide Web (WWW 2013). 2013. P. 1445–1456.

6. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019. 2019. P. 4171–4186. DOI: https://doi.org/10.48550/arXiv.1810.04805.

7. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv. 2022. DOI: https://doi.org/10.48550/arXiv.2203.05794.

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

9. Murshed B. A. H. та ін. Short text topic modelling approaches in the context of big data. Artificial Intelligence Review. 2022. Vol. 56, No. 6. P. 5133–5260. DOI: https://doi.org/10.1007/s10462-022-10254-w.

10. Fan C., Shi H., Yuan C. Short text topic modeling: A survey. Knowledge-Based Systems. 2023. DOI: https://doi.org/10.1016/j.knosys.2023.110421.

11. Terragni S., Fersini E., Galuzzi B., Tropeano P., Candelieri F. OCTIS: Comparing and Optimizing Topic Models is Simple! Proceedings of EACL 2021 (System Demonstrations). 2021. P. 263–270. DOI: https://doi.org/10.18653/v1/2021.eacl-demos.31.

12. Hosseiny M., Marani M., Baumer E. P. S. A Review of Stability in Topic Modeling: Metrics for Assessing and Techniques for Improving Stability. ACM Computing Surveys. 2024. Vol. 56, No. 5. Art. 108. DOI: https://doi.org/10.1145/3623269.

REFERENCES:

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. 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

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

5. Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. In Proceedings of the 22nd International Conference on World Wide Web (WWW 2013) (pp. 1445–1456).

6. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171–4186). https://doi.org/10.48550/arXiv.1810.04805

7. Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv. https://doi.org/10.48550/arXiv.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. Murshed, B. A. H., et al. (2022). Short text topic modelling approaches in the context of big data. Artificial Intelligence Review, 56(6), 5133–5260. https://doi.org/10.1007/s10462-022-10254-w

10. Fan, C., Shi, H., & Yuan, C. (2023). Short text topic modeling: A survey. Knowledge-Based Systems. https://doi.org/10.1016/j.knosys.2023.110421

11. Terragni, S., Fersini, E., Galuzzi, B., Tropeano, P., & Candelieri, F. (2021). OCTIS: Comparing and Optimizing Topic Models is Simple! In Proceedings of EACL 2021 (System Demonstrations) (pp. 263–270). https://doi.org/10.18653/v1/2021.eacl-demos.31

12. Hosseiny, M., Marani, M., & Baumer, E. P. S. (2024). A review of stability in topic modeling: Metrics for assessing and techniques for improving stability. ACM Computing Surveys, 56(5), Article 108. https://doi.org/10.1145/3623269

Дата першого надходження статті до видання: 21.02.2026
Дата прийняття статті до друку після рецензування: 20.03.2026
Дата публікації (оприлюднення) статті: 20.05.2026