УДК 655.5:004.8

СПРИЙНЯТТЯ ТЕКСТОВОЇ ІНФОРМАЦІЇ У КОРПУСІ «ВИДАВНИЧА СПРАВА – ДИСТАНЦІЙНА ОСВІТА» НА ОСНОВІ МЕТОДІВ ТЕМАТИЧНОГО МОДЕЛЮВАННЯ

Ж. В. Дейнеко[1], Р. В. Слісаренко[2]

[1] Харківський національний університет радіоелектроніки, просп. Науки, 14, м. Харків, 61166, Україна, e-mail: zhanna.deineko@nure.ua
[2] Харківський національний університет радіоелектроніки, просп. Науки, 14, м. Харків, 61166, Україна, e-mail: roman.slisarenko@nure.ua

Запропоновано відтворюваний протокол тематичного моделювання для аналізу наукового корпусу, релевантного перетину дистанційної освіти та видавничо-поліграфічної проблематики, із фокусом на контролі доменної «протікальності» як ризику тематизації лексично перетинних піддоменів. Протокол реалізує багатоступеневу фільтрацію: жорсткі та доменно-специфічні виключення, вилучення «publication-polish», застосування порога poly_score ≥ 2.0 і обмеження мінімальної довжини (≥ 40 токенів). У результаті сформовано доменно-сфокусований core-набір (262 → 49; suspects_n = 0). Тематичну структуру визначено методом NMF на поданні TF-IDF, що забезпечило виокремлення 6 стійких тем із локалізованими термінологічними ядрами, релевантними практикам типографіки, верстання, набору, видавничого виробництва та цифровим характеристикам дистанційного контенту (web-based подання, usability, доступність). Узгодженість структури підтверджено кластеризацією в просторі Sentence-BERT (k = 4). Протокол може використовуватися як аналітична основа для систематизації дискурсу й підтримки редакційних рішень.

Ключові слова: *дистанційна освіта, видавнича справа, поліграфія, тематичне моделювання, NMF, TF-IDF, доменна фільтрація, доменна «протікальність», Sentence-BERT, кластеризація.*

Постановка проблеми. У дистанційній освіті зростає обсяг навчально-методичних матеріалів, які створюються та оновлюються в різних форматах, зокрема в системах управління навчанням (Learning Management System, LMS), електронних конспектах, інтерактивних модулях і методичних вказівках, а підготовка контенту часто відбувається у розподілених командах без єдиного редакційно-видавничого контролю та узгоджених стандартів оформлення. За таких умов непослідовність у застосуванні норм типографіки, верстання, редакторського опрацювання й підготовки макетів призводить до зниження якості електронних і друкованих навчальних видань, погіршення читабельності та доступності, фрагментації стилю, а також до зростання витрат часу на виправлення й перевипуск матеріалів. Водночас організації, що підтримують дистанційне навчання, накопичують великі масиви текстів, у яких педагогічний зміст поєднується з елементами видавничої

справи та поліграфії (структурування сторінки, вимоги до макета, типографічні правила, виробничі та цифрові обмеження). Традиційні ручні методи огляду й систематизації такого корпусу є трудомісткими, слабо масштабуються та не забезпечують відтворюваності висновків, що ускладнює формування узгоджених редакційних політик і технологічних рішень.

У цих умовах актуальним стає застосування тематичного моделювання як інструмента аналітики, що дозволяє виявляти та структурувати тематичні напрями, пов'язані з видавничо-поліграфічними аспектами дистанційної освіти. Однак на практиці виникає критичний методичний ризик доменної «протікальності»: через перетин термінології (design, layout, publishing тощо) у результати тематизації можуть потрапляти документи із суміжних, але нерелевантних піддоменів, що призводить до змішування тематик, зниження інтерпретованості та некоректних прикладних висновків. Це зумовлює необхідність розроблення відтворюваного процесу тематичного аналізу, який поєднує формалізовану фільтрацію корпусу, інтерпретоване тематичне виділення та незалежну семантичну перевірку узгодженості групувань для отримання публікаційно-придатних і доменно релевантних результатів.

Аналіз останніх досліджень та публікацій. Тематичне моделювання є одним із базових інструментів структурування великих текстових масивів, зокрема для виявлення латентних тематичних напрямів у наукових публікаціях. Класичним підходом є Latent Dirichlet Allocation (LDA), генеративна ймовірнісна модель, у якій тема описується розподілом слів, а документ подається як суміш тем [1], [2]. Практика застосування LDA вказує на її придатність для корпусів із достатньо довгими документами та відносно однорідним лексичним простором; натомість у випадку коротких фрагментів (назви, анотації, ключові слова) і високого перетину загальної лексики між суміжними доменами зростає ризик «розмиття» тем та домінування загальних термінів, що ускладнює доменну інтерпретацію, що узгоджується з висновком про фрагментарність дослідницького поля та необхідність узагальнювального тематичного аналізу для формування цілісної картини [3]. Альтернативою для підвищення інтерпретованості є факторизаційні методи, зокрема Non-negative Matrix Factorization (NMF), які формують локалізовані термінологічні ядра тем у просторі Term Frequency–Inverse Document Frequency (TF-IDF) ознак [4], [5]. Для задач, де важлива не лише лексична, а й семантична близькість документів, застосовуються трансформерні семантичні векторні подання речень (sentence embeddings), які відображають зміст текстів у щільному векторному просторі та підтримують кластеризацію і валідацію тематичних структур поза межами частотної лексики [6]. Окрему групу становлять сучасні конвеєри тематизації, які поєднують ембеддинги, зниження розмірності та кластеризацію з подальшим формуванням переліків ключових термінів тем, зокрема в підході BERTopic (BERT-based Topic Modeling), що підсилює роботу з короткими текстами та підвищує керованість тематичних контурів у неоднорідних корпусах [7]. У видавничій справі та поліграфії дослідження фокусуються на типографічних рішеннях, правилах макетування, читабельності, технологічних етапах підготовки видань і виробничих обмеженнях [8], [9], однак у контексті дистанційної освіти

бракує відтворюваних методик, які б одночасно забезпечували тематичну систематизацію великих корпусів і контроль доменної релевантності. Це визначає доцільність поєднання інтерпретованого тематичного виділення (на основі TF-IDF/NMF) із семантичним групуванням документів у просторі ембеддингів та формалізованими правилами доменної фільтрації для мінімізації «протікальності» та отримання стійких тем, придатних до видавничо-поліграфічної інтерпретації у дистанційному навчанні.

Мета статті. Основну мету дослідження спрямовано на розроблення та обґрунтування відтворюваного процесу тематичного моделювання для аналізу корпусів дистанційної освіти, у яких присутні змішані фрагменти навчального змісту та термінології видавничої справи й поліграфії. Передбачається формалізувати послідовність етапів відбору релевантних документів, контролю доменної «протікальності» та побудови інтерпретованих тематичних структур шляхом поєднання моделей на основі TF-IDF (NMF) із семантичним кластеруванням документів за векторними поданнями. Реалізація зазначеної мети має забезпечити отримання стабільних і змістовно узгоджених тем, придатних для подальшої інтерпретації у контексті видавничо-поліграфічних процесів підготовки освітніх матеріалів, а також для формування рекомендацій щодо підвищення якості редакційно-видавничого супроводу ресурсів дистанційного навчання.

Виклад основного матеріалу дослідження. Дослідження виконано на корпусі наукових публікацій, релевантних одночасно дистанційній освіті та видавничо-поліграфічному контексту. Критичною умовою валідної тематизації в таких даних є контроль доменної «протікальності»: через спільну лексику (design, layout, publishing тощо) до вибірки можуть потрапляти документи із суміжних, але нерелевантних піддоменів. Для мінімізації цього ефекту застосовано поетапну селекцію документів: вилучення за жорстким списком виключень (hard blacklist), доменно-специфічні виключення (CORE exclusions) та додаткові виключення «publication-polish», після чого виконано відбір in-domain за порогом poly_score ≥ 2.0 і фільтрацію за мінімальною довжиною тематизованого тексту L(topic_text) ≥ 40 токенів, де poly_score – агрегований індикатор доменної релевантності документа до видавничо-поліграфічного контексту (використовується як пороговий критерій включення), а topic_text – нормалізований текстовий вхід для тематизації (поля документа після очищення та відсікання шуму); поріг L(topic_text) відсікає надто короткі записи, для яких тематичні моделі формують нестабільні або випадкові компоненти. Підсумковий «core»-набір сформовано як компактний доменно сфокусований корпус для інтерпретованої тематизації та подальшої семантичної перевірки узгодженості.

Вибір тематичної моделі здійснювався на основі емпіричного зіставлення альтернатив у спільному експериментальному протоколі з єдиними правилами попередньої обробки, векторизації та оцінювання результатів. Параметр target_min_docs у табл. 1 задає мінімально бажаний розмір доменного підкорпусу для стабільної тематизації (служить орієнтиром при налаштуванні правил селекції, а не «жорсткою» умовою включення). Параметри vectorizer min_df/max_df визначають пороги відсікання термінів під час побудови словника: min_df = 2 вилучає

надрідкісні терміни, тоді як max_df = 0.95 пригнічує надчастотні (загальні) слова, що зменшує лексичний шум і підвищує інтерпретованість тем у NMF.

Таблиця 1

Статистики фільтрації корпусу та ключові параметри протоколу

Показник	Значення
Вхідний обсяг корпусу, ()	262
Вилучено hard blacklist	33 (12.6%)
Вилучено CORE exclusions	33 (12.6%)
Вилучено publication-polish exclusions	4 (1.5%)
Відібрано як in-domain	173 (66.0%)
Після фільтра довжини	136 (51.9%)
Фінальний core-набір, ()	49 (18.7%)
Поріг довжини topic-text, tokens	40
Обраний поріг	2.0
Цільовий мінімум документів ()	60
Розмір словника після векторизації	874
vectorizer /	2 / 0.95
Кількість тем NMF	6
Кількість кластерів (ембеддинги)	4
Підозрілі документи ()	0 (0.0%)

На початковому етапі оцінено ймовірнісну модель LDA як базовий підхід для наукових корпусів [1], [2]; однак для даних зі значною часткою коротких і стилістично неоднорідних текстів (назви, анотації, ключові слова) та із перетином загальної лексики між суміжними доменами було зафіксовано зниження інтерпретованості: теми виявлялися надмірно «розмитими», чутливими до вибору кількості тем і гіперпараметрів, а також схильними до домінування загальнонаукових термінів, що ускладнювало відокремлення видавничо-поліграфічних аспектів і підвищувало ризик доменної «протікальності». Практично це проявлялося у слабкій тематичній диференціації та появі тем із близькими наборами високочастотних слів, які складно інтерпретувати в прикладному контексті типографіки, верстання та виробничих процесів.

У зв'язку з цим виконано перехід до факторизаційного підходу NMF на TF-IDF [4], [5], який посилює роль дискримінативних термінів, зменшує вплив частотної «загальності» лексики та, як наслідок, формує більш локалізовані термінологічні ядра тем у змішаних корпусах (табл. 2). Фінальну конфігурацію (6 тем) прийнято як компроміс між компактністю тематичної структури та її змістовою диференційованістю для поліграфічного контексту; додатково її узгодженість перевірено семантичним групуванням документів у просторі Sentence-BERT-ембеддингів [6], що забезпечує незалежну валідацію стабільності смислових «ядер» поза межами частотного подання TF-IDF.

Таблиця 2

Теми NMF (6 тем): top-words та внесок у корпус (за домінуванням у документах)

Тема	Узагальнювальна інтерпретація	Top-words (з CSV)	Частка теми у core-наборі (за score)
0	Типографіка та графічний дизайн	typography, design, graphic, graphic design, type, …	0.232
1	Макет/верстка та структурні елементи сторінки	page, layout, page layout, document, document layout, …	0.186
2	Набір/LaTeX та математична верстка	typesetting, mathematical, word, latex, construct, …	0.130
3	Візуальна ієрархія/ естетика (layout-saliency)	textual layout, textual, layout, saliency, aesthetic, …	0.070
4	Друк/видавництво (історико-технол. вимір)	printing, movable, movable type, publishing, book, …	0.163
5	Web-based/юзабіліті та навчальний контент	based, web based, usability, web, learners, course, …	0.219

Семантична інтерпретація виділених тем свідчить про наявність двох взаємодоповнювальних підсистем. Перша підсистема має виробничо-видавничий характер і охоплює типографіку, макетування, набір і друк. Друга підсистема відображає цифрово-освітній вимір та пов'язана з web-орієнтованими навчальними середовищами, вимогами usability і доступності. Така структура є закономірною для корпусу, в якому поліграфічні практики репрезентовані як у традиційних видавничо-друкарських процесах, так і в цифрових формах дистанційної освіти, зокрема через веборієнтоване подання (web-based), вимоги до зручності користування (usability) та стандарти доступності.

Для незалежної верифікації узгодженості тематичної структури виконано семантичне групування документів у просторі ембеддингів Sentence-BERT [6] двома альтернативними алгоритмами кластеризації: агломеративною ієрархічною кластеризацією (Agglomerative clustering) та методом k-середніх (k-means), із фіксацією кількості кластерів на рівні чотирьох (k = 4). Робастність групувань оцінювали за відтворюваністю смислових «ядер», коли поява співставних кластерів за різних алгоритмів розглядалася як незалежне підтвердження консистентності тематичної організації корпусу. Інтерпретацію кластерів здійснено на підставі найбільш репрезентативних термінів кластера та домінантних тем NMF у документах, що входять до відповідного кластера. Такий підхід дає змогу розмежувати змістові групи, що підтверджують основні тематичні компоненти, і потенційні доменні «витоки», тобто нерелевантні включення. Під «підозрілими» документами у межах даного конвеєра розуміли записи з позадоменними репрезентативними термінами або зі слабким узгодженням між домінантною NMF-темою та семантичним оточенням кластера.

У межах сформованого core-набору таких документів не виявлено (0), що відповідає suspects_n = 0 у табл. 1 та узгоджується з результатами попередньої доменної селекції.

Кластерні профілі демонструють відтворення ключових змістових «ядер», релевантних поліграфічному контексту дистанційної освіти: верстка/сторінка (структурування матеріалу та макетні рішення), типографіка/графічний дизайн (візуальна форма подання та читабельність), друк/видавництво (виробничо-технологічний вимір), а також цифрова взаємодія й навчальний контент (вимоги платформ, користувацький досвід (User Experience, UX) і доступність). Для підвищення читабельності результати подано окремо для Agglomerative (табл. 3) та k-means (табл. 4): Agglomerative використано як перевірку робастності групувань, тоді як основна інтерпретація ґрунтується на k-means. Показник узгодження кластера з домінантною NMF-темою n_dom/n_docs кількісно фіксує ступінь семантичної однорідності групи та узгодження між семантичним групуванням (ембеддинги) і лексично інтерпретованими компонентами NMF (TF-IDF); це дає підстави розглядати виділені групи як стабільні смислові сегменти корпусу. Отримані оцінки підтверджують, що «цифрово-освітній» сегмент (кластер 3 у k-means, домінантна тема 5) формує окреме семантичне ядро, релевантне висновкам щодо цифрової трансформації видавничих практик у дистанційній освіті.

У сукупності наведені результати підсилюють інтерпретацію NMF-тем як змістовно узгоджених компонентів і водночас демонструють, що «цифрово-освітня» підсистема (тема 5) не розчиняється серед виробничо-видавничих тематик, а відокремлюється в самостійну групу документів у k-means, що має прикладне значення для редакційно-видавничих рішень щодо структурування та подання навчального контенту в онлайнових середовищах.

Таблиця 3

Семантичні кластери (Sentence-BERT, Agglomerative): розмір, Top-5 термінів і узгодження з NMF

Cluster_id	n_docs	Top-terms (Top-5)	Домінантна NMF-тема (id)	n_dom	Узгодження, %
0	14	layout, page, visual, page layout, design	1	6	42.9
1	15	printing, publishing, typography, media, design	4	6	40.0
2	10	typography, typographic, design, type, graphic	0	9	90.0
3	10	typesetting, word, mathematical, document, image	2	4	40.0

Таблиця 4

Семантичні кластери (Sentence-BERT, k-means): розмір, Top-5 термінів і узгодження з NMF

Cluster_id	n_docs	Top-terms (Top-5)	Домінантна NMF-тема (id)	n_dom	Узгодження, %
0	11	layout, page, page layout, document, image …	1	7	63.6
1	23	typography, design, typographic, type, graphic …	0	13	56.5
2	6	publishing, typesetting, indesign, mathematical, publication …	4	3	50.0
3	9	interactive, classroom, learners, journalism, course …	5	7	77.8

Для наочного підтвердження узгодженості семантичних групувань у просторі Sentence-BERT-ембеддингів подано «карту щільності» документів у двовимірній проєкції: поверх точок додано контури ядерної оцінки щільності, які підкреслюють компактність та часткове перекриття змістових груп. На рис. 1 наведено топографічну візуалізацію розподілу документів core-набору з маркуванням за кластерною належністю (k-means), де контурні «поля» відображають області підвищеної семантичної концентрації. Отримане представлення узгоджується з табл. 4 (k-means) і візуально підтверджує структуру семантичних груп, описану через їх розміри, топ-терми та домінантні NMF-теми: виробничо-видавничі напрями (типографіка/верстка/набір/друк) формують окремі ядра, відокремлені від цифрово-освітнього сегмента (web-based/usability), що додатково підтримує інтерпретацію виділених тем як змістовно стійких і доменно релевантних компонентів.

Отримана тематична структура може бути інтерпретована як індикаторний «зріз» технологічних і редакційних пріоритетів у середовищах дистанційної освіти. Теми 0–2 прямо відносяться до допрес-логіки (типографіка, макет, набір/формули), що визначає вимоги до якості навчальних матеріалів, їх читабельності та відтворюваності в цифрових і друкованих форматах. Тема 4 відображає видавничо-друкарський сегмент (publishing/printing), який є релевантним для формування видавничого портфеля та технологічних рішень (зокрема щодо друку навчальних матеріалів, архітектури видань, вибору форматів). Тема 5 відображає web-based та usability-вимір, що безпосередньо пов’язаний з редакційними політиками для дистанційного контенту (структурування, доступність, інтерфейсні обмеження, мультимодальність) і підсилює необхідність узгодженого редакційно-видавничого супроводу цифрових освітніх ресурсів.

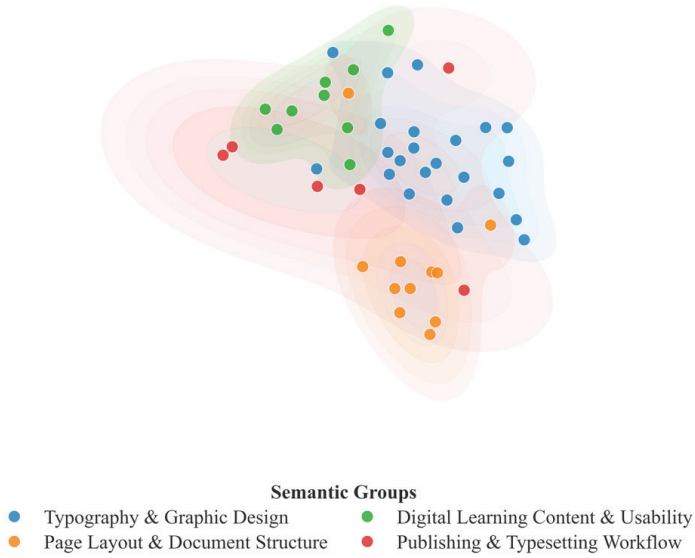

Рис. 1. Топографічна карта щільності документів core-набору у просторі Sentence-BERT-ембеддингів (k-means): контури щільності та точки документів за змістовими групами

Висновки. У роботі сформовано та апробовано відтворюваний протокол тематичного аналізу наукового корпусу, що відображає перетин дистанційної освіти та видавничо-поліграфічної проблематики, з акцентом на контроль доменної «протікальності» як ключового ризику тематизації змішаних корпусів. Поетапна селекція документів за формалізованими правилами (жорсткі та доменно-специфічні виключення, поріг доменної релевантності poly_score ≥ 2.0. обмеження мінімальної довжини тематизованого тексту) забезпечила формування компактного, але доменно сфокусованого «core»-набору = 49, придатного для інтерпретованого виділення тем та подальшої семантичної валідації. Практичний результат цього етапу полягає в мінімізації позадоменного шуму й стабілізації термінологічного поля, що підвищує публікаційність тематичних висновків у поліграфічному контексті.

На основі NMF на TF-IDF отримано 6 стійких тем, які формують змістовно узгоджену структуру з двома взаємодоповнювальними підсистемами. Перша відображає допрес-логіку та виробничо-видавничі процеси (типографіка/графічний дизайн, макет/верстка й структурні елементи сторінки, набір/LaTeX і математична верстка, а також сегмент друку/видавництва). Друга репрезентує цифровий вимір навчального контенту (web-based/usability), що є критичним для дистанційної освіти і водночас напряму пов’язаний із редакційними політиками щодо структурування, доступності, мультимодальності та керованості навчальних матеріалів. Таким чином, тематична структура слугує індикатором того, що якість освітніх ресурсів у дистанційному навчанні визначається не лише педагогічним змістом, а й дотриманням видавничих норм оформлення та технологічної придатності контенту для багатоканальної дистрибуції.

Узгодженість тематичних компонентів підтверджено семантичним групуванням документів у просторі Sentence-BERT-ембеддингів: обидві схеми кластеризації (Agglomerative та k-means, по 4 кластери) відтворюють близькі за змістом смислові «ядра», а двовимірне проєктування ембеддингів із поданням топографічної карти щільності наочно демонструє їх компактність, часткове перекриття суміжних підтем та відокремлення цифрово-освітнього сегмента. Відсутність підозрілих доменних відхилень у фінальному наборі (suspects_n = 0) узгоджується з етапами доменної селекції та підтверджує доцільність поєднання формалізованої фільтрації, інтерпретованого тематичного моделювання та незалежної семантичної валідації. Практична цінність отриманих результатів полягає у можливості використовувати теми та кластери як аналітичну основу для систематизації наукового дискурсу, а також для обґрунтування редакційно-технологічних рішень щодо підготовки й оновлення навчально-методичних матеріалів (контроль типографіки та верстки, стандартизація метаданих і версій, підсилення читабельності, забезпечення сумісності з цифровими платформами й, за потреби, з друкованими форматами). Подальші дослідження доцільно спрямувати на розширення корпусу, введення часової компоненти (динаміки тем), а також на формалізацію критеріїв «протікальності» у вигляді кількісних індикаторів, інтегрованих у конвеєр підготовки корпусу та валідації тем.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation // *Journal of Machine Learning Research*. 2003. Vol. 3. P. 993-1022. DOI: 10.5555/944919.944937.
2. Griffiths T. L., Steyvers M. Finding Scientific Topics // *Proceedings of the National Academy of Sciences*. 2004. Vol. 101, Suppl. 1. P. 5228–5235. DOI: 10.1073/pnas.0307752101.
3. Філь Н. Ю., Слісаренко Р. В., Дейнеко Ж. В., Морозова Л. Ю. Тенденції розвитку досліджень у сфері штучного інтелекту в освіті: тематичне моделювання за допомогою латентного розподілу Діріхле // *Вісник Харківського національного автомобільно-дорожнього університету*. 2025. № 108. С. 17–24. DOI: 10.30977/BUL.2219-5548.2025.108.0.17.
4. Lee D. D., Seung H. S. Learning the parts of objects by non-negative matrix factorization // *Nature*. 1999. Vol. 401. P. 788–791. DOI: 10.1038/44565.
5. Gillis N. *Nonnegative Matrix Factorization*. SIAM, 2020. DOI: 10.1137/1.9781611976410.
6. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // *Proceedings of EMNLP-IJCNLP*. 2019. DOI: 10.18653/v1/D19-1410.
7. Grootendorst M. *BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics* // Zenodo. 2022. DOI: 10.5281/zenodo.4719700.
8. Ren H., Liu Y., Naren G., Lu J. The impact of multidirectional text typography on text readability in word clouds // *Displays*. 2024. Vol. 83. Art. 102724. DOI: 10.1016/j.displa.2024.102724.
9. Vollenwyder B., Petralito S., Iten G. H., Brühlmann F., Opwis K., Mekler E. D. How compliance with web accessibility standards shapes the experiences of users with and without disabilities // *International Journal of Human-Computer Studies*. 2022. Art. 102956. DOI: 10.1016/j.ijhcs.2022.102956.

REFERENCES

1. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022. https://doi.org/10.5555/944919.944937.
2. Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(Suppl. 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101.
3. 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.
4. 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.
5. Gillis, N. (2020). *Nonnegative matrix factorization*. SIAM. https://doi.org/10.1137/1.9781611976410.
6. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. https://doi.org/10.18653/v1/D19-1410.
7. Grootendorst, M. (2022). *BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics* [Software]. Zenodo. https://doi.org/10.5281/zenodo.4719700.
8. Ren, H., Liu, Y., Naren, G., & Lu, J. (2024). The impact of multidirectional text typography on text readability in word clouds. *Displays, 83*, 102724. https://doi.org/10.1016/j.displa.2024.102724.
9. Vollenwyder, B., Petralito, S., Iten, G. H., Brühlmann, F., Opwis, K., & Mekler, E. D. (2022). How compliance with web accessibility standards shapes the experiences of users with and without disabilities. *International Journal of Human-Computer Studies*, 102956. https://doi.org/10.1016/j.ijhcs.2022.102956.

doi: 10.32403/1998-6912-2026-1-72-39-49

PERCEPTION OF TEXTUAL INFORMATION IN THE CORPS “PUBLISHING – DISTANCE EDUCATION” BASED ON THEMATIC MODELING METHODS

Zh. V. Deineko[1], R. V. Slisarenko[2]

[1] Kharkiv National University of Radio Electronics,
14 Nauky Ave., Kharkiv, 61166, Ukraine, e-mail: zhanna.deineko@nure.ua
[2] Kharkiv National University of Radio Electronics,
14 Nauky Ave., Kharkiv, 61166, Ukraine, e-mail: roman.slisarenko@nure.ua

The paper proposes a reproducible topic-modeling protocol for analyzing a scientific corpus located at the intersection of distance education and publishing/printing studies, with a focus on controlling domain “leakage” as a key risk when thematizing lexically overlapping subdomains. The protocol implements multi-stage filtering that combines

hard exclusions, domain-specific exclusions, and the removal of "publication polish" fragments, followed by a domain-relevance threshold (poly_score ≥ 2.0) and a minimum thematic-text length constraint (≥ 40 tokens). As a result, the initial dataset is reduced from = 262 to a compact domain-focused core set = 49, while no suspicious out-of-domain items are detected (suspects_n = 0). Topic components are extracted using Non-negative Matrix Factorization (NMF) with a TF-IDF representation, yielding six stable topics with localized terminological cores. The obtained topics align with publishing production practices (typography, layout, typesetting, and publishing workflows) and with digital characteristics of distance-learning content, including web-based representation, usability considerations, and accessibility requirements. The coherence of the resulting thematic structure is additionally validated through document clustering in a Sentence-BERT embedding space (k = 4), where semantic groups reproduce the corpus "core" and do not demonstrate domain drift. The proposed protocol can be used as a reproducible analytical basis for structuring the scientific discourse at this intersection and for supporting editorial and publishing decision-making in distance education environments.

Keywords: *distance education, publishing industry, printing, topic modeling, NMF, TF-IDF, domain filtering, domain leakage, Sentence-BERT, clustering.*

Стаття надійшла до редакції 02.03.2026.
Submitted: 02.03.2026.
Прийнято до друку: 20.03.2026.
Accepted: 20.03.2026.
Опубліковано: 30.05.2026.
Published: 30.05.2026.