

УДК 81'33

DOI <https://doi.org/10.32782/2412-933X/2025-XXIV-19>

АНАТОМІЯ ХУДОЖНЬОГО ТЕКСТУ: ПОГЛЯД КРІЗЬ ПРИЗМУ ЦИФРОВОЇ ФІЛОЛОГІЇ (НА ПРИКЛАДІ РОМАНУ Т. ДРАЙЗЕРА «ФІНАНСИСТ»)

Саліонович Людмила

кандидат філологічних наук, доцент,
доцент кафедри міжкультурної комунікації та іноземної мови
Національний технічний університет «Харківський політехнічний інститут»
ORCID ID: 0000-0001-9005-9273
Researcher ID: KDR-0826-2024

Землякова Олена

кандидат філологічних наук, доцент,
доцент кафедри міжкультурної комунікації та іноземної мови
Національний технічний університет «Харківський політехнічний інститут»
ORCID ID: 0000-0002-4668-8264
Web of Science Researcher ID: KGM-0065-2024

Рубцова Вікторія

доцент кафедри міжкультурної комунікації та іноземної мови
Національний технічний університет «Харківський політехнічний інститут»
ORCID ID: 0000-0003-1120-5514

У статті розглянуто потенціал тематичного моделювання як інструменту кількісного аналізу художнього тексту в контексті сучасної цифрової філології. Метою дослідження є порівняння ефективності двох підходів до автоматичного виокремлення тем у художньому дискурсі: класичного алгоритму LDA (Latent Dirichlet Allocation) та сучасної моделі BERTopic, що базується на використанні контекстуальних векторних представлень і здатна вловлювати глибші семантичні зв'язки в тексті. Матеріалом дослідження слугує роман Теодора Драйзера «Фінансист», на основі якого сформовано корпус для тематичного аналізу. Розглянуто стратегії попереднього оброблення тексту: корпус підготовлено за допомогою стандартних процедур очищення, лематизації та фільтрації частин мови із залученням інструментів NLP. Тексти розбито на сегменти, дібрано оптимальні параметри для кожної моделі. У процесі моделювання аналізувалася кількість отриманих тем, когерентність, їх змістова насиченість, інтерпретованість і адекватність художньому змісту твору. Особливу увагу приділено методам налаштування моделей (визначення кількості тем, фільтрація частин мови, оброблення власних імен), візуалізації результатів для покращення інтерпретації та впливу попереднього оброблення тексту на якість тем. Порівняльний аналіз показав, що модель BERTopic виявляє глибші семантичні зв'язки, ніж LDA, та в більшості випадків формує теми, які точніше передають смислові домінанти тексту, демонструє вищу здатність до побудови змістово згаджених тематичних структур, зберігаючи при цьому контекстуальні зв'язки між словами. Результати дослідження можуть бути використані для подальшого вивчення семантичної структури художніх творів та розроблення методик для автоматизованого аналізу художнього дискурсу. Перспектива подальших досліджень вбачається в проведенні порівняльного тематичного аналізу кількох романів одного автора з метою виявлення еволюції смислових домінант у межах авторського стилю.

Ключові слова: тематичне моделювання, оброблення природної мови, латентне розміщення Діріхле, BERTopic, кількісний аналіз, семантична структура, художній текст.

Salionovych Liudmyla

Candidate of Philological Sciences, Associate Professor,
Associate Professor at the Department of Cross-Cultural Communication and Foreign Language
National Technical University "Kharkiv Polytechnic Institute"

Zemliakova Olena

Candidate of Philological Sciences, Associate Professor,
Associate Professor at the Department of Cross-Cultural Communication and Foreign Languages
National Technical University "Kharkiv Polytechnic Institute"

Rubtsova Victoria

Associate Professor at the Department of Cross-Cultural Communication and Foreign Language
National Technical University "Kharkiv Polytechnic Institute"

**ANATOMY OF A LITERARY TEXT:
A PERSPECTIVE THROUGH THE LENS OF DIGITAL PHILOLOGY
(A CASE STUDY OF T. DREISER'S "THE FINANCIER")**

The article explores the potential of topic modeling as a tool for the quantitative analysis of literary texts in the context of contemporary digital philology. The aim of the study is to compare the effectiveness of two approaches to automatic topic extraction in literary discourse: the classical Latent Dirichlet Allocation (LDA) algorithm and the modern BERTopic model, which relies on contextual vector representations and is capable of capturing deeper semantic relationships within the text. The material of the study is Theodore Dreiser's novel "The Financier", on the basis of which a corpus for topic analysis was created. The research discusses text preprocessing strategies, including standard cleaning procedures, lemmatization, and part-of-speech filtering using NLP tools. The text was segmented, and optimal parameters were selected for each model. During modeling, the number of extracted topics, their coherence, semantic richness, interpretability, and correspondence to the literary content were analyzed. Special attention was given to model tuning methods (determining the number of topics, part-of-speech filtering, handling of proper nouns), visualization of results for improved interpretation, and the impact of preprocessing on topic quality. The comparative analysis revealed that BERTopic captures deeper semantic connections than LDA and, in most cases, generates topics that more accurately reflect the text's semantic dominants. It demonstrates a higher capacity for constructing semantically coherent thematic structures while preserving contextual relationships between words. The results of the study may be used for further exploration of the semantic structure of literary works and for developing methods of automated literary discourse analysis. The prospect of further research is seen in conducting a comparative topic analysis of multiple novels by the same author to trace the evolution of semantic dominants within the author's style.

Key words: topic modeling, Natural Language Processing, Latent Dirichlet Allocation, Bidirectional Encoder Representations from Transformers, quantitative analysis, semantic structure, literary text.

Постановка проблеми. Останні десятиліття гуманітарні науки дедалі частіше звертаються до цифрових методів аналізу текстів, і це не випадково. Завдяки стрімкому розвитку технологій оброблення природної мови (Natural Language Processing, NLP) та щораз більшій потребі в нових способах розуміння текстів дослідники отримали потужні інструменти для роботи з інформацією. Серед них особливе місце посідає тематичне моделювання – метод,

що допомагає виявляти приховані смисли у великих масивах текстів. Цей підхід давно став незамінним не лише в науці, а й у повсякденному житті: від аналізу соціальних мереж і маркетингу – до вивчення громадської думки та політичних тенденцій. Компанії використовують тематичне моделювання, щоб краще зрозуміти потреби своїх клієнтів, аналізуючи відгуки чи пости в соцмережах. Дослідники з його допомогою відстежують суспільні настрої, а аналітики – виявляють підтексти в новинних заголовках. Особливо цінним такий метод є тоді, коли потрібно зрозуміти, що турбує людей, які в них уподобання або як на них намагаються впливати за допомогою тонких маніпуляцій. Тематичне моделювання стало справжнім містком між «сухими» даними й глибоким розумінням людських емоцій та ідей.

Попри технічну зрілість підходу, його застосування щодо художніх текстів залишається обмеженим. Метою цього дослідження є осмислення потенціалу тематичного моделювання в контексті художнього тексту та оцінювання ефективності конкретних інструментів NLP у цій галузі. Це дозволяє не лише точніше визначити можливості кожного з підходів, а й окреслити перспективи глибшого осмислення художньої семантики засобами обчислювального аналізу.

Аналіз останніх досліджень та публікацій. Сучасні дослідження підкреслюють усе більшу важливість тематичного моделювання як інструменту цифрової філології, що дозволяє виявляти приховані смислові структури у великих корпусах текстів.

Первісно тематичне моделювання розглядалося як засіб виявлення латентних тем у збірках документів, що знайшло застосування в аналізі наукових текстів [3; 9], у визначенні тональності (сентимент-аналізі) коротких текстів, наприклад, для виявлення та систематизації тем у дописах у соціальних мережах [7; 12; 18], а також в аналізі настроїв онлайн-контенту [14], тематичному аналізі новинних повідомлень [11; 21; 24] або клієнтських відгуків. У роботі [2] розглядаються класичні підходи до тематичного моделювання: латентно-семантичний аналіз (LSA), імовірнісний латентно-семантичний аналіз (PLSA) та латентне розміщення Діріхле (LDA). Робота моделей тестується на корпусі відгуків користувачів інтернет-магазину цифрових товарів, результати оцінюються як за кількісними метриками (когерентність, перплексія), так і візуально – через хмари слів. У процесі систематичного порівняння класичних моделей демонструються відмінності в їх продуктивності та інтерпретованості.

Серед досліджень, присвячених порівнянню методів тематичного моделювання для аналізу клієнтських відгуків, заслуговує на увагу [11]. У контексті аналізу користувацького контенту автори порівняли шість популярних моделей тематичного аналізу: LSA, LDA, NMF, PAM, Top2Vec і BERTopic, застосовуючи їх до корпусів відгуків клієнтів. Особливу увагу приділено точності, інтерпретованості та когерентності отриманих тем. Згідно з результатами, BERTopic показала найвищі значення когерентності та смислової узгодженості, що підкреслює її ефективність для виявлення ключових тем у неструктурованих текстах. Це дослідження демонструє переваги нейромережових моделей, зокрема трансформерів, у семантичному аналізі користувацького контенту і є актуальним прикладом переходу від традиційних імовірнісних методів (LDA) до сучасних нейромережових підходів (BERTopic) [10] у практичних задачах аналізу тексту.

Вагомими перевагами застосування нейромереж до тематичного аналізу тексту описано в [16], де модель вперше була адаптована до морфологічно складної мови – сербської – на прикладі твітів, що виражали скепсис щодо вакцинації від COVID-19. У дослідженні порівнюються BERTopic, LDA та NMF у двох сценаріях попереднього оброблення: повному та частковому. Модель демонструє стійкість до зниження якості навіть за мінімального оброблення, тоді як традиційні моделі потребують лематизації та зменшення словника. За

показниками когерентності та тематичної варіативності BERTopic перевершує LDA та NMF, а також забезпечує більш інтерпретовані та змістовні теми. Ці результати особливо важливі для застосування BERTopic у мовах з обмеженими NLP-ресурсами та високою морфологічною варіативністю. BERTopic навіть за мінімального оброблення здатна формувати більш інформативні й різноманітні теми, ніж LDA чи NMF. На відміну від традиційних моделей, BERTopic працює на рівні контекстуальних векторних подань (ембеддінгів), що дозволяє їй краще «відчувати» сенс і тональність навіть у складних лінгвістичних умовах, як-от художній текст. Для аналізу літературного тексту це означає потенційну можливість зменшення втручання в початкову структуру тексту без втрати якості тематичного моделювання.

Що стосується художніх текстів, завдання застосування кількісного аналізу ускладнюється особливим характером художніх творів. Літературні твори, що вирізняються складною наративною структурою, стилістичною поліфонією та глибокою семантичною насиченістю, висувають особливі вимоги до методів тематичного аналізу. Попри окремі успішні приклади, систематичне використання NLP-інструментів у філологічному аналізі художньої прози поки що не набуло належного поширення. Особливо актуальним постає питання придатності та адаптивності різних моделей тематичного моделювання до специфіки художнього тексту, де традиційні методи можуть давати фрагментарні або неінтерпретовані результати.

У праці [13] проведено систематичний аналіз ключових моделей тематичного аналізу (LDA, NMF, нейромережових моделей) та продемонстровано як їхні переваги, так і методологічні обмеження під час роботи з літературними корпусами. Автори підкреслюють, що тематичне моделювання в літературознавстві потребує адаптації традиційного алгоритмічного підходу до особливостей художнього тексту. Інтерпретація тем становить складний аналітичний процес, що поєднує статистичні дані з гуманітарною експертизою. Контекстуальний і порівняльний аналіз допомагають пов'язати теми з літературним стилем, епохою та жанром твору. Міждисциплінарний підхід, що об'єднує соціологію, психологію та історію, дозволяє глибше зрозуміти художні ідеї, виявлені алгоритмами. Водночас важливо пам'ятати: статистика не вловлює метафор і риторики тексту, отже, необхідний якісний аналіз.

Практичні застосування тематичного моделювання в літературознавстві можуть охоплювати як аналіз окремих творів, так і масштабні порівняльні дослідження [6]. Моделі дозволяють виявляти стійкі тематичні структури, співвідносити їх з ідеологічним та жанровим контекстом [22], простежувати розвиток мотивів у межах одного тексту або між творами різних авторів, епох і культур [15; 20]. Зокрема, тематичне моделювання активно застосовується для виявлення сюжетів [17; 23], а також для інтерпретації емоційних та етичних домінант тексту [26].

Особливий інтерес з погляду пошуку нових технічних рішень для подолання суперечностей під час комп'ютерного оброблення саме художнього тексту становить робота [19], де запропоновано метод семантичної кластеризації, заснований на інтеграції вдосконаленого алгоритму латентного розподілу Діріхле (LDA) з можливостями глибинної мовної моделі BERT. У роботі також підкреслюються обмеження класичного LDA – зокрема, його нездатність точно вловлювати семантичні нюанси та контекстуальні зв'язки, що є особливо критичним під час роботи з великими й складними корпусами природної мови. Таким чином, одним із можливих шляхів розв'язання проблеми підвищення точності та інтерпретованості тематичного аналізу є поєднання модифікованих механізмів LDA з використанням контекстних ембеддингів BERT. За результатами порівняння запропонований

гібридний метод перевершує класичний LDA за якістю тематичної кластеризації, що підтверджує його високу придатність для лінгвістичних завдань, які потребують урахування тонких смислових відтінків. Таким чином, дослідження поєднує статистичні й трансформерні методи тематичного моделювання для глибшого аналізу текстів.

Низку досліджень присвячено обговоренню обмеженості застосування кількісних методів аналізу до художніх текстів [4; 5; 8]. У [25] автори досліджують обмеження тематичного моделювання (на прикладі LDA) під час аналізу літературних текстів. Вони зазначають, що LDA як статистичний інструмент радше виявляє елементи когезії (cohesion), ніж когерентності (coherence), що ускладнює інтерпретацію отриманих тем. У межах дослідження було побудовано моделі для кількох корпусів прози з різним рівнем очищення даних: повні тексти, сегментовані тексти та лематизовані тексти з обмеженням за частинами мови. Для підвищення якості тем застосовувалися методи попереднього оброблення, як-от видалення високо- й низькочастотних слів, стоп-слів, а також обмеження словникового складу. Результати показали, що якість тем суттєво залежить від ступеня структурної однорідності тексту та глибини очищення: за мінімального очищення моделі відтворюють елементи когезії, що не несуть смислового навантаження, тоді як сегментація та фільтрація покращують когерентність тем. Робота підкреслює важливість ретельної підготовки корпусу в разі застосування тематичного моделювання до літературних текстів.

Серед перспектив розвитку тематичного моделювання в гуманітарній науці можна виокремити вдосконалення алгоритмів з урахуванням особливостей художньої мови, інтеграцію тематичного аналізу з іншими цифровими методами (емоційний аналіз, аналіз персонажів, семантичні ембеддинги); розвиток інструментів візуалізації. Таким чином, тематичне моделювання може розглядатися не лише як допоміжний кількісний метод, а й як повноцінний інструмент літературного аналізу, що має високий евристичний потенціал за умови критично осмисленого застосування.

Метою дослідження є здійснення порівняльного аналізу ефективності алгоритмів LDA та BERTopic, оцінювання інтерпретованості та якості виявлених тем, а також обговорення параметрів налаштування та візуалізації, що сприяють точнішому семантичному відображенню тексту.

Методи та методики дослідження. У роботі застосовано кількісні та якісні методи аналізу, зокрема тематичне моделювання для виявлення смислових структур тексту, корпусна лінгвістика (попереднє оброблення даних: токенізація, лематизація, фільтрація стоп-слів), статистичні методи (оцінювання когерентності тем, частотний аналіз лексики), контекстуально-семантичний аналіз на основі векторних представлень (BERT), порівняльний аналіз ефективності алгоритмів (LDA та BERTopic), а також використано візуалізацію даних для інтерпретації результатів.

Виклад основного матеріалу дослідження. Гіпотеза цього дослідження полягає в тому, що сучасні контекстно-орієнтовані методи тематичного моделювання, як-от BERTopic, дозволяють глибше й точніше розкрити тематичну структуру художнього тексту порівняно з класичною моделлю LDA. Ми припускаємо, що завдяки здатності враховувати семантичні зв'язки та контекст, такі моделі мають вищий інтерпретаційний потенціал і тому є особливо цінними для завдань цифрової філології.

Об'єктом дослідження є текст роману Теодора Драйзера «Фінансист», що став основою його знаменитої «Трилогії бажання». У цьому творі, як і в усій трилогії, простежується складна внутрішня еволюція головного героя – Френка Каупервуда, де переплітаються теми влади, моралі, капіталу, самотності та життєвого вибору. Однак питання про те, як ці теми структуровані всередині самого тексту і які смислові домінанти визначають його

внутрішню організацію, залишається відкритим. Таким чином, завданням дослідження є зіставлення результатів, отриманих за допомогою двох різних підходів до тематичного моделювання, реалізованих у сучасних NLP-інструментах. Аналіз здійснюється на лексико-семантичному рівні та спрямований на реконструкцію прихованої тематики тексту.

Для отримання вихідної моделі LDA, з якою буде здійснено порівняння BERTopic, було проведено низку ітерацій з налаштування та оптимізації коду LDA (від версії 1 до 5), спрямованих на підвищення якості тематичного моделювання та покращення когерентності отриманих тем. На першому етапі (Код 1) було реалізовано базове передоброблення тексту з використанням SpaCy, фільтрацією частин мови та лематизацією. У таблиці нижче подано результати виконання коду з базовими параметрами та мінімальним очищенням тексту.

Отримані дані дозволяють зробити висновок, що теми мають сюжетні лінії, проте є розмитими – вони часто перетинаються. Видно наявність семантичних кластерів, але без чітких меж між темами. Це свідчить про слабку параметризацію моделі та недостатнє передоброблення тексту. Деякі теми (наприклад, 2 і 5; 3 і 4) змішують суміжні, але різні за контекстом елементи, що ускладнює їх інтерпретацію. Тема 7 важко піддається тлумаченню. З лінгвістичної точки зору це демонструє, що модель без оптимізації часто групує частотні слова за асоціативною ознакою, а не за семантичною когерентністю. Через це змістовний літературний аналіз ускладнюється.

У Коді 2 список стоп-слів було розширено шляхом включення найбільш частотної лексики корпусу – це дозволило прибрати нейтральну лексику та посилити тематичну специфіку. У Коді 3 було додано біграмну модель, що дало змогу виявити стійкі словосполучення, що підвищило семантичну цілісність тем (когерентність до 0.4528 при 7 темах). У четвертій версії (Код 4) список допустимих частин мови було звужено до найбільш семантично насичених (NOUN, VERB, ADJ), а також виключено власні іменники (PROPN), що дозволило уникнути концентрації тем навколо персонажів і підвищити тематичну абстракцію (когерентність 0.4552).

Таблиця 1

Інтерпретація тем LDA (Код 1)

№ теми	Ключові слова	Інтерпретація	Коментар
1	see, understand, way, try, find, possible, care, business, point, city, office, private, life, situation...	Абстрактні роздуми, мислення, сприйняття + частково бізнес	Тема досить розпорошена: частина слів абстрактно-когнітивні, частина – про справи й бізнес. Немає чіткої сфокусованості
2	dollar, money, loan, broker, pay, certificate, check, treasurer, bank, penitentiary...	Фінанси та юридичні наслідки	Досить стійка фінансова тема, але перетинається з темою в'язниці, покарання
3	good, old, mind, love, talk, mother, father, jury, judge, marry, thought...	Родина, емоції, суд, стосунки	Емоційно-особистісна сфера, пов'язана з родиною та соціальною мораллю. Присутні нерелевантні слова
4	reply, want, court, prison, law, face, black, white, dark, paper...	Судово-в'язнична сцена	Схожа на сцену у в'язниці або суді. Містить юридичну лексику. Загалом узгоджена
5	matter, sell, buy, street, stock, railway, financial, order, fail, fire, jail...	Бізнес, криза, крах, покарання	Суміш бізнес-контексту та невдач. Має сюжетну пов'язаність, але перетинається з темою 2
6	ask, hear, wife, gentleman, night, door, panic, feel, affair...	Домашня сцена, тривожні емоції	Оповідання, пов'язане із ситуацією вдома, вночі, з тривожними відчуттями. Нечітко виражена тема
7	get, leave, run, lie, wish, crime, woman, daughter, morning, deposit...	Тривога, порушення закону, родина	Динамічна тема з діями, тривогою, злочином і родинними фігурами. Важко інтерпретується

Фінальний Код 5 об'єднав усі попередні вдосконалення: розширене передоброблення, вилучення зайвої лексики, роботу з біграмами, налагоджувальні метрики, а також підбір оптимальної кількості тем із використанням когерентності c_v та перплексії. Результат – суттєве зростання когерентності моделі й семантично стійкі, тематично інтерпретовані кластери. П'ята ітерація LDA, яка має найвищу когерентність ($c_v = 0.4669$), виокремила вісім тем, що відображають ключові аспекти роману Теодора Драйзера «Фінансист». Оптимізована модель (Код 5) демонструє помітно вищу якість тематичної структури: теми набувають чіткіших меж, стають семантично когерентними, в більшості випадків добре розрізняються, лексично стійкі й охоплюють основні наративні лінії роману. У таблиці 2 подано ці теми включно з 30 ключовими словами для кожної, їх інтерпретацією та зв'язком із сюжетом.

Таким чином, Код 5 являє собою оптимізовану версію LDA-моделювання з найкращими показниками якості, що робить його релевантною основою для порівняння з BERTopic.

Таблиця 2

Інтерпретація тем LDA (Код 5)

№ теми	Ключові слова	Інтерпретація	Коментар
1	money, dollar, buy, wish, loan, sell, order, bank, check, offer, right, disaster, pay, interest, judge, certificate, city, lead, spread, condition, city_loan, explain, information, institution, purchase, small, trade, market, receive, give	Фінансові операції	Тема відображає фінансову діяльність Френка Каупервуда, центральну для «Фінансиста». Більшість слів описують його операції з активами, кредитами та банківською справою, вказують на фінансові інструменти. Але деякі слова стосуються суду та арешту
2	old, lose, work, prison, life, meet, deal, stay, begin, cell, banking_house, convict, dream, letter, catch, prosperity, trap, speculator, circumstance, action, suppose, private, trading, art, ruin, air, jailed, offense, keep, pardon	Юридичні наслідки	Тема описує ув'язнення Каупервуда та його фінансові втрати, крах. Проте тут багато шумових слів, спостерігається сильне перетинання з темою 3, часткове – з темами 4 і 6
3	power, understand, failure, world, night, land, able, fine, concern, young, significant, exclaim, point, jury, suggest, nature, idea, happen, attitude, rule, witness, depth, subtle, rise, connection, feeling, case, enter, appearance, month	Амбіції та суд	Тема не дуже чітка, змішана: розкриває амбіції Каупервуда і його судовий процес. Деякі слова, як уже зазначалося, належать до інших тем. Сильне перетинання з темою 2
4	love, help, house, wife, grant, brilliant, affair, people, understanding, ye, believe, divorce, review, thought, anxious, live, occasion, sit, low, story, center, justice, afternoon, daughter, strike, sister, rat, cause, visit, heart	Любовні стосунки	Тема описує особисте життя Каупервуда, включаючи романтичні та сімейні зв'язки; «rat» виглядає як шумове слово, можливо, артефакт оброблення тексту
5	hear, leave, panic, call, moment, fire, hope, tear, boy, return, smile, care, color, observe, news, strange, fight, child, light, water, gray, silence, end, start, big, trouble, brown, hair, hour, clerk	Кризові моменти	Тема відображає кризові події, як-от біржова паніка, пожежа, емоційні реакції на фінансові потрясіння
6	money, dollar, buy, wish, loan, sell, order, bank, check, offer, right, disaster, pay, interest, judge, certificate, city, lead, spread, condition, city_loan, explain, information, institution, purchase, small, trade, market, receive, give	Фінансові операції	Тема відображає фінансову діяльність Френка Каупервуда, центральну для «Фінансиста». Більшість слів описують його операції з активами, кредитами та банківською справою, вказують на фінансові інструменти. Деякі слова стосуються суду та арешту
7	old, lose, work, prison, life, meet, deal, stay, begin, cell, banking_house, convict, dream, letter, catch, prosperity, trap, speculator, circumstance, action, suppose, private, trading, art, ruin, air, jailed, offense, keep, pardon	Юридичні наслідки	Тема описує ув'язнення Каупервуда та його фінансові втрати, крах. Проте тут багато шумових слів, спостерігається сильне перетинання з темою 3, часткове – з темами 4 і 6.

Таблиця 3

Інтерпретація тем BERTopic

№ теми	Ключові слова	Інтерпретація	Коментар
1	Father, love, think, life, mother, wife, man, time, daughter, woman, want, family, marry, way, good	Родинні стосунки, почуття, цінності	Емоційно насичена, чітка і зрозуміла тема
2	Eye, look, hair, gray, black, blue, red, window, foot, face, cell, arm, house, door, room	Зовнішність, опис персонажів і оточення, інтер'єри	Описова тема, характерна для художнього дискурсу
3	City, loan, treasurer, dollar, money, city treasurer, city loan, political, certificate, cent, bank, matter, treasury, pay, fund	Фінанси в міському управлінні, муніципальна економіка	Чітка тема, можливий додатковий акцент на політичному компоненті розповіді
4	Jury, court, case, prison, judge, charge, jail, guilty, trial, lawyer, convict, prisoner, sentence, penitentiary, larceny	Судова система, злочин і покарання	Яскраво виражена, насичена тема, весь кластер складається з тематично значущих слів
5	Stock, street railway, line, share, buy, sell, dollar, business, great, street car, control, broker, money, local, want	Фінансові операції, біржа, вуличні залізниці	Бізнес й інвестиції, дуже чітка тематика

Для тематичного моделювання тексту роману була використана модель BERTopic, яка демонструє високу ефективність під час аналізу літературних даних. На першому етапі роман було розбито на послідовні блоки по три речення, що дозволило зберегти контекстуальну цілісність фрагментів і забезпечити більш усвідомлену кластеризацію.

Далі було проведено стандартне попереднє оброблення тексту, що включало токенизацію, видалення стоп-слів і пунктуації, лематизацію, а також очищення від неінформативних частин мови. Для підвищення тематичної чистоти з корпусу також були вилучені власні імена, оскільки їх висока частотність (особливо в художньому тексті) може зміщувати акценти у формуванні тем і ускладнювати інтерпретацію результатів.

Модель BERTopic була налаштована з використанням CountVectorizer, обмеженого уніграмами та біграмами. Значення параметрів $\text{min_df}=10$ і $\text{max_df}=0.9$ дозволили виключити відповідно рідкісні та надто частотні слова. Навчання моделі здійснювалося методом `fit_transform()`, після чого для кожної теми було виділено 15 найбільш значущих лексем.

Застосований підхід забезпечив отримання стійких і семантично інтерпретованих тем, релевантних як змісту, так і структурі тексту.

У результаті порівняння двох моделей тематичного моделювання стає очевидною перевага BERTopic у задачі аналізу художнього тексту. Наведені вище таблиці відображають розподіл тем, виокремлених із роману Теодора Драйзера «Фінансист», і дають підстави стверджувати, що BERTopic формує більш осмислені та інтерпретовані теми. Модель LDA продемонструвала певну інформативність: в її темах можна розпізнати ключові аспекти оповіді – фінансову діяльність головного героя, юридичні наслідки його вчинків, родинні стосунки та життєві кризи. Проте в темах LDA спостерігається високий ступінь тематичного перетину (наприклад, між темами 2 і 3, 1 і 8), а також наявність шумових або нерелевантних слів, що ускладнює інтерпретацію. Деякі теми виявилися надто розмитими або штучно об'єднували лексеми, пов'язані з різними сюжетними лініями, через що їх семантичне ядро ставало нестійким. BERTopic, навпаки, продемонстрував вищий рівень тематичної чистоти й зв'язності. Кожна з п'яти тем, виокремлених моделлю, вирізняється лексичною однорідністю та високою інтерпретованістю. Наприклад, тема 4 повністю охоплює судову систему й термінологію покарання, без домішок нерелевантної лексики; тема 5 чітко представляє біржові операції та вуличні залізниці, які відіграють ключову роль у бізнес-стратегіях

Каупервуда. Додатковим чинником, що вплинув на якість тем у BERTopic, стало використання гнучкішого підходу до структурування даних: модель навчалася на блоках по три речення, а також із фільтрацією власних назв. Це дало змогу знизити шум, пов'язаний із частими згадками персонажів, і зосередитися на змістовній лексиці. Також у BERTopic було застосовано представлення тексту у вигляді ембеддінгів (у нашому випадку – з використанням all-MiniLM-L6-v2), що дало змогу враховувати семантичні зв'язки між словами, а не лише їх частотне співіснування, як у LDA.

Таким чином, BERTopic продемонстрував значну перевагу над класичною LDA в контексті аналізу художнього тексту: він забезпечив точніше тематичне членування, кращу якість кластерів і вищу придатність для інтерпретації художнього матеріалу.

Висновки. Проведене дослідження показало, що сучасні цифрові методи дозволяють виявити приховані тематичні шари художнього тексту, відкриваючи нові горизонти для його лінгвістичної інтерпретації. Перспективним напрямом подальших досліджень вбачається порівняльний тематичний аналіз кількох творів одного автора – з метою простежити еволюцію смислових домінант і глибше зрозуміти особливості його художнього стилю.

Література:

1. Babalola, Olusola & Ojokoh, Bolanle & Boyinbode, Olutayo. Comprehensive Evaluation of LDA, NMF, and BERTopic's Performance on News Headline Topic Modeling. *Journal of Computing Theories and Applications*. 2024. 2. 268–89. DOI: <https://doi.org/10.62411/jcta.11635->
2. Бердник Д., Бойчук А. Порівняльний аналіз методів тематичного моделювання для аналізу відгуків в інтернет-магазині цифрових товарів. *Herald of Khmelnytskyi National University. Technical Sciences*. 2022. № 307(2). С. 37–41. <https://doi.org/10.31891/2307-5732-2022-307-2-37-41>.
3. Blei D. M., J. D. Lafferty. A correlated topic model of Science. *Ann. Appl. Stat.*, vol. 1, no. 1, Jun. 2007. doi: 10.1214/07-AOAS114.
4. Chen Y., Z. Peng, S.-H. Kim, C. W. Choi, "What We Can Do and Cannot Do with Topic Modeling: A Systematic Review," *Commun. Methods Meas.*, vol. 17, no. 2, pp. 111–130, Apr. 2023. doi: 10.1080/19312458.2023.2167965.
5. Da N. Z. The computational case against computational literary studies. *Critical Inquiry*. 2019. Vol. 45, № 3. P. 601–639. DOI:10.1086/702594.
6. Dahllöf, Mats & Berglund, Karl. Faces, Fights, and Families: Topic Modeling and Gendered Themes in Two Corpora of Swedish Prose Fiction. *Digital Humanities in the Nordic and Baltic Countries Publications*. 2. 92–111. 2019. DOI:10.5617/dhnpub.11084.
7. Egger R., J. Yu. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Front. Sociol.* vol. 7, May 2022. doi: 10.3389/fsoc.2022.886498.
8. Gius Evelyn, Janina Jacke. Are Computational Literary Studies Structuralist? *Journal of Cultural Analytics*, vol. 7, no. 4, Dec. 2022. <https://doi.org/10.22148/001c.46662>.
9. Griffiths T. L., M. Steyvers. Finding scientific topics. *Proc. Natl. Acad. Sci.*, vol. 101, no. suppl_1, pp. 5228–5235, Apr. 2004. doi: <https://doi.org/10.1073/pnas.0307752101>.
10. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv. Mar. 11, 2022. URL: <http://arxiv.org/abs/2203.05794>
11. Лаптев О., Юзва А. Дослідження використання технік моделювання для аналізу відгуків клієнтів. *Прикладні проблеми комп'ютерних наук, безпеки та математики*. 2024. № 2. С. 4–17. URL: <https://apcsm.vnu.edu.ua/index.php/Journalone/article/view/20/15>
12. Laureate C. D. P., W. Buntine, H. Linger. A systematic review of the use of topic models for short text social media analysis. *Artif. Intell. Rev.*, vol. 56, no. 12, pp. 14223–14255, Dec. 2023. doi: 10.1007/s10462-023-10471-x.
13. Li Defeng, Wu Kan and Lei Victoria L.C. Applying Topic Modeling to Literary Analysis: A Review. *Digital Studies in Language and Literature*. vol. 1, no. 1-2, 2024, pp. 113–141. <https://doi.org/10.1515/dsll-2024-0010>.

14. Lyu J. C., Han E. L., Luli G. K. COVID-19 Vaccine–Related Discussion on Twitter: Topic Modeling and Sentiment Analysis *J Med Internet Res* 2021. 23(6), e24435. doi: 10.2196/24435.
15. Martinelli G., P. Impicciché, E. Fersini, F. Mambrini, and M. Passarotti. Exploring Neural Topic Modeling on a Classical Latin Corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* Torino, Italy, 6929–34. 2024. Paris: ELRA and ICCL. URL: <https://aclanthology.org/2024.main-1.606>
16. Medvecki D., Bašaragin B., Ljajić A., Milošević N. Multilingual Transformer and BERTopic for Short Text Topic Modeling: The Case of Serbian. In: Trajanovic, M., Filipovic, N., Zdravkovic, M. (eds) *Disruptive Information Technologies for a Smart Society. ICIST 2023. Lecture Notes in Networks and Systems*, vol 872. Springer, Cham. 2024. https://doi.org/10.1007/978-3-031-50755-7_16.
17. Monika W., V. Amelia, Q. Aris, A. Nasution. Topic Modeling of Indonesian Children's Literature Using Latent Semantic Analysis. In *Proceedings of the 2nd International Conference on Environmental, Energy, and Earth Science, ICEEES 2023, 30 October 2023, Pekanbaru, Indonesia*. Pekanbaru: European Alliance for Innovation. <http://dx.doi.org/10.4108/eai.30-10-2023.2343063>.
18. Murshed, B. A. H., J. Abawajy, S. Mallappa, M. A. N. Saif, S. M. Al-Ghuribi, and F. A. Ghanem. Enhancing Big Social Media Data Quality for Use in Short-Text Topic Modeling. *IEEE Access*, vol. 10, pp. 105328–105351, 2022. doi: 10.1109/ACCESS.2022.3211396.
19. Narozhnyi V., Kharchenko V. Semantic clustering method using integration of advanced LDA algorithm and BERT algorithm. *Innovative technologies and scientific solutions for industries*. 2024. (1 (27), 140–153. <https://doi.org/10.30837/ITSSI.2024.27.140>.
20. Navarro-Colorado B. On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry. *Frontiers in Digital Humanities*. 2018. Vol. 5, № 15. <https://doi.org/10.3389/fdigh.2018.00015>.
21. Rajan S. D., T. Coombs, M. Jayabalan, N. A. Ismail. A Comparative Study of Methods for Topic Modelling in News Articles in Data Science and Emerging Technologies, 2024, pp. 269–277. doi: 10.1007/978-981-97-0293-0_20.
22. Schoch C. Topic modeling genre: an exploration of French classical and enlightenment drama. *Digital Humanities Quarterly*. 2017. Vol. 11. № 2. DOI:10.48550/arXiv.2103.13019
23. Schröter J., K. Du. Validating Topic Modeling as a Method of Analyzing Sujet and Theme. *Journal of Computational Literary Studies* 1 (1): 1–18. 2022. DOI: <https://doi.org/10.48694/jcls.91>.
24. Svensson K., Blad, J. Exploring NMF and LDA Topic Models of Swedish News Articles. Uppsala Universitet, 2020. URL: <https://uu.diva-portal.org/smash/get/diva2:1512130/FULLTEXT01.pdf>
25. Uglanova I., Gius E. The Order of Things. A Study on Topic Modelling of Literary Texts. Workshop on Computational Humanities Research. 2020. URL: <https://ceur-ws.org/Vol-2723/long7.pdf>
26. Völkl Y., Sarić S., Scholger, M. Topic Modeling for the Identification of Gender-specific Discourse. Virtues and Vices in French and Spanish 18th Century Periodicals. *Journal of Computational Literary Studies*. 1(1). 2022. doi: <https://doi.org/10.48694/jcls.108>.

References:

1. Babalola, O., Ojokoh, B., & Boyinbode, O. (2024). Comprehensive evaluation of LDA, NMF, and BERTopic's performance on news headline topic modeling. *Journal of Computing Theories and Applications*, 2, 268–289. <https://doi.org/10.62411/jcta.11635> [in English]
2. Berdnyk, D., & Boichuk, A. (2022). Porivnialnyi analiz metodiv tematychnogo modeliuvannia dlia analizu vidhukiv v internet-mahazyni tsyfrovyykh tovariv [Comparative analysis of topic modeling methods for analyzing reviews in an online store of digital goods]. *Herald of Khmelnytskyi National University. Technical Sciences*, 307(2), 37–41. <https://doi.org/10.31891/2307-5732-2022-307-2-37-41> [in Ukrainian]
3. Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *Annals of Applied Statistics*, 1(1), 17–35. <https://doi.org/10.1214/07-AOAS114> [in English]
4. Chen, Y., Peng, Z., Kim, S.-H., & Choi, C. W. (2023). What we can do and cannot do with topic modeling: A systematic review. *Communication Methods and Measures*, 17(2), 111–130. <https://doi.org/10.1080/19312458.2023.2167965> [in English]
5. Da, N. Z. (2019). The computational case against computational literary studies. *Critical Inquiry*, 45(3), 601–639. <https://doi.org/10.1086/702594> [in English]

6. Dahllöf, M., & Berglund, K. (2019). Faces, fights, and families: Topic modeling and gendered themes in two corpora of Swedish prose fiction. *Digital Humanities in the Nordic and Baltic Countries Publications*, 2, 92–111. <https://doi.org/10.5617/dhnbpub.11084> [in English]
7. 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> [in English]
8. Gius, E., & Jacke, J. (2022). Are computational literary studies structuralist? *Journal of Cultural Analytics*, 7(4). <https://doi.org/10.22148/001c.46662> [in English]
9. 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> [in English]
10. Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv*. Retrived from: <https://arxiv.org/abs/2203.05794> [in English]
11. Laptev, O., & Yuzva, A. (2024). Doslidzhennia vykorystannia tekhnik modeliuvannia dlia analizu vidhukiv kliientiv [Research on the use of modeling techniques for analyzing customer reviews]. *Prykladni problemy komp'uternykh nauk, bezpeky ta matematyky*, 2, 4–17. Retrived from: <https://apcssm.vnu.edu.ua/index.php/Journalone/article/view/20/15> [in Ukrainian]
12. Laureate, C. D. P., Buntine, W., & Linger, H. (2023). A systematic review of the use of topic models for short text social media analysis. *Artificial Intelligence Review*, 56(12), 14223–14255. <https://doi.org/10.1007/s10462-023-10471-x> [in English]
13. Li, D., Wu, K., & Lei, V. L. C. (2024). Applying topic modeling to literary analysis: A review. *Digital Studies in Language and Literature*, 1(1–2), 113–141. <https://doi.org/10.1515/dsll-2024-0010> [in English]
14. Lyu, J. C., Han, E. L., & Luli, G. K. (2021). COVID-19 vaccine-related discussion on Twitter: Topic modeling and sentiment analysis. *Journal of Medical Internet Research*, 23(6), e24435. <https://doi.org/10.2196/24435> [in English]
15. Martinelli, G., Impicciché, P., Fersini, E., Mambrini, F., & Passarotti, M. (2024). Exploring neural topic modeling on a classical Latin corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 6929–6934). ELRA and ICCL. Retrived from: <https://aclanthology.org/2024.main-1.606> [in English]
16. Medveckí, D., Bašaragin, B., Ljajić, A., & Milošević, N. (2024). Multilingual Transformer and BERTopic for short text topic modeling: The case of Serbian. In M. Trajanovic, N. Filipovic, & M. Zdravkovic (Eds.), *Disruptive information technologies for a smart society. ICIST 2023* (pp. 169–178). Springer. https://doi.org/10.1007/978-3-031-50755-7_16 [in English]
17. Monika, W., Amelia, V., Aris, Q., & Nasution, A. (2024). Topic modeling of Indonesian children's literature using Latent Semantic Analysis. In *Proceedings of the 2nd International Conference on Environmental, Energy, and Earth Science, ICEEES 2023. European Alliance for Innovation*. <https://doi.org/10.4108/eai.30-10-2023.2343063> [in English]
18. Murshed, B. A. H., Abawajy, J., Mallappa, S., Saif, M. A. N., Al-Ghuribi, S. M., & Ghanem, F. A. (2022). Enhancing big social media data quality for use in short-text topic modeling. *IEEE Access*, 10, 105328–105351. <https://doi.org/10.1109/ACCESS.2022.3211396> [in English]
19. Narozhnyi, V., & Kharchenko, V. (2024). Semantic clustering method using integration of advanced LDA algorithm and BERT algorithm. *Innovative Technologies and Scientific Solutions for Industries*, (1)27, 140–153. <https://doi.org/10.30837/ITSSI.2024.27.140> [in English]
20. Navarro-Colorado, B. (2018). On poetic topic modeling: Extracting themes and motifs from a corpus of Spanish poetry. *Frontiers in Digital Humanities*, 5, Article 15. <https://doi.org/10.3389/fdigh.2018.00015> [in English]
21. Rajan, S. D., Coombs, T., Jayabalan, M., & Ismail, N. A. (2024). A comparative study of methods for topic modelling in news articles. In *Data science and emerging technologies* (pp. 269–277). Springer. https://doi.org/10.1007/978-981-97-0293-0_20 [in English]
22. Schoch, C. (2017). Topic modeling genre: An exploration of French classical and enlightenment drama. *Digital Humanities Quarterly*, 11(2). <https://doi.org/10.48550/arXiv.2103.13019> [in English]
23. Schröter, J., & Du, K. (2022). Validating topic modeling as a method of analyzing sujet and theme. *Journal of Computational Literary Studies*, 1(1), 1–18. <https://doi.org/10.48694/jcls.91> [in English]

24. Uglanova, I., & Gius, E. (2020). The order of things. A study on topic modelling of literary texts. Workshop on Computational Humanities Research. Retrived from: <https://ceur-ws.org/Vol-2723/long7.pdf> [in English]
25. Völkl, Y., Sarić, S., & Scholger, M. (2022). Topic modeling for the identification of gender-specific discourse. Virtues and vices in French and Spanish 18th century periodicals. *Journal of Computational Literary Studies*, 1(1). <https://doi.org/10.48694/jcls.108> [in English]
26. Blad, J., & Svensson, K. (2020). Exploring NMF and LDA topic models of Swedish news articles. Uppsala Universitet. Retrived from: <https://uu.diva-portal.org/smash/get/diva2:1512130/FULLTEXT01.pdf> [in English]