*ISSN 2522-9818 (print)*
*ISSN 2524-2296 (online)* *Innovative technologies and scientific solutions for industries. 2024. No. 1 (27)*

УДК 004.8.032.26 DOI: https://doi.org/10.30837/ITSSI.2024.27.140

В. НАРОЖНИЙ, В. ХАРЧЕНКО

# МЕТОД СЕМАНТИЧНОЇ КЛАСТЕРИЗАЦІЇ З ВИКОРИСТАННЯМ ІНТЕГРАЦІЇ ВДОСКОНАЛЕНОГО АЛГОРИТМУ *LDA* Й АЛГОРИТМУ *BERT*

**Предметом дослідження** є поглиблений семантичний аналіз даних, оснований на модифікації методологій латентного розподілу Діріхле (*LDA*) та інтеграції її двоспрямованого кодувального подання з трансформаторів (*BERT*). **Актуальність роботи.** Прихований розподіл Діріхле є фундаментальною технікою моделювання тем, яку широко застосовують у різноманітних програмах для аналізу текстів. Хоча його корисність загальновизнана, традиційні моделі *LDA* часто стикаються з обмеженнями, зокрема жорстким розподілом тем і неадекватним відтворенням нюансів семантики, властивих природній мові. **Мета дослідження** – покращення адекватності та точності семантичного аналізу завдяки вдосконаленню базового механізму *LDA*, що інтегрує адаптивні пріоритети Діріхле та використовує глибокі семантичні можливості вбудовувань *BERT*. **Упроваджені методи**: відбір текстових наборів даних; попереднє оброблення даних; удосконалення алгоритму *LDA*; інтеграція з *BERT Embeddings*; порівняльний аналіз. **Завдання дослідження**: 1) теоретичне обґрунтування модифікації *LDA*; 2) реалізація інтеграції з *BERT*; 3) оцінювання ефективності методу; 4) порівняльний аналіз; 5) розроблення архітектурного рішення. **Результати** полягають у тому, що насамперед окреслено теоретичні основи як стандартної, так і модифікованої моделей *LDA*, а також детально викладено їх розширену формулу. За допомогою серії експериментів на текстових наборах даних, що визначаються різними емоційними станами, визначено ключові переваги запропонованого підходу. На підставі порівняльного аналізу за такими показниками, як внутрішньо та міжкластерна відстані та силуетний коефіцієнт, доведено підвищену когерентність, інтерпретованість і адаптивність модифікованої моделі *LDA*. Запропоновано архітектурне рішення для реалізації методу. **Висновки**. Емпіричні результати свідчать про значне покращення виявлення тонких складностей і тематичних структур у текстовій інформації, що є кроком в еволюційному розвитку методологій тематичного моделювання. Крім того, результати досліджень не лише створюють можливості застосування *LDA* для більш складних лінгвістичних сценаріїв, але й окреслюють шляхи їх подальшого вдосконалення для неконтрольованого аналізу текстів.

**Ключові слова:** семантичний аналіз; природна мова; алгоритм *LDA*; алгоритм *BERT*; інтерактивне мистецтво; емоційна реакція.

## 1. Вступ

У сучасному світі текстової аналітики та оброблення природної мови (*NLP*) моделювання тем залишається фундаментальним завданням, що передбачає пошук інформації, класифікацію документів і розуміння тематичних структур у великих текстових масивах. Серед різних підходів латентний розподіл Діріхле (*Latent Dirichlet Allocation, LDA*) широко визнаний завдяки своїй ефективності у виявленні латентних тем. Однак зі збільшенням складності та обсягу текстових даних стають усе більш очевидними обмеження традиційного *LDA*, зокрема з огляду на нюанси семантики та контекст мови [1].

Поява глибокого навчання і, зокрема, трансформаторних моделей, таких як *BERT (Bidirectional Encoder Representations from Transformers)*, стала революційною в галузі *NLP*, пропонуючи ґрунтовне розуміння мовної структури та контексту. *BERT* завдяки своїй глибокій двоспрямованій природі фіксує тонкі значення та взаємозв'язки в мові, значно перевершуючи попередні моделі в широкому діапазоні завдань *NLP* [2]. Однак інтеграція таких складних мовних моделей із традиційними підходами до моделювання тем є складним завданням, але водночас і можливістю значно підвищити якість та інтерпретованість тем.

Мета статті – подолати розрив між традиційним моделюванням тем і останніми досягненнями у сфері глибокого навчання. Пропонуємо нову інтеграцію *BERT* з удосконаленою версією *LDA*, яка має на меті використати сильні властивості обох моделей. Розширена модель *LDA* передбачає такі модифікації, як адаптивні попередні дані та вставки слів, щоб покращити припущення та можливості стандартної моделі *LDA* [3]. Інтегруючи *BERT* на етапі післяоброблення, ми прагнемо уточнити теми, визначені за допомогою *LDA*, використовуючи

ISSN 2522-9818 (print)
*Сучасний стан наукових досліджень та технологій в промисловості. 2024. №1 (27)* ISSN 2524-2296 (online)

контекстні вставки *BERT*, щоб забезпечити багатше та більш узгоджене подання тем.

Основна мета й завдання дослідження полягають в тому, щоб перетворити моделювання тем на потужніший інструмент для аналізу текстів. Інтеграція спрямована на вирішення кількох завдань.

1. Семантичне багатство. Хоча *LDA* забезпечує хорошу основу для виявлення тем, вона часто не може охопити семантичне багатство й контекст мови, що призводить до менш пов'язаних або значущих тем. Використовуючи глибокі контекстуальні репрезентації слів *BERT*, інтегрована модель має на меті покращити семантичний зв'язок і змістовність виявлених тем.

2. Адаптивність і гнучкість. Традиційні моделі *LDA* часто критикують за їх жорсткі та спрощені припущення щодо розподілу тем і зв'язків між словами та темами. Удосконалена модель *LDA* з її адаптивними пріоритетами та вбудовуваннями слів, а також інтеграцією з *BERT*, має на меті забезпечити більшу гнучкість, ефективніше пристосуватися до різних характеристик даних і змін, що відбуваються з часом.

3. Інтерпретованість і розрізнення. Однією з основних цілей тематичного моделювання є надання інтерпретованих і чітких тем, які можна легко зрозуміти і використовувати для подальшого аналізу. Запропонована інтеграція має на меті покращити інтерпретованість та виразність тем, що робить модель більш цінною та застосовною в практичних сценаріях.

Розв'язуючи окреслені проблеми, запропонована інтеграція не тільки розширює можливості застосування *LDA* до більш складних і нюансованих наборів даних, але й відкриває нові шляхи для досліджень і застосування в текстовій аналітиці та *NLP*. У статті планується описати методології, упроваджені в розширену модель *LDA* та процес інтеграції з *BERT*, після цього буде обговорено очікувані переваги та наслідки такого підходу. Кінцева мета – надати надійне, семантично багате й адаптивне рішення для моделювання тем в епоху глибокого навчання.

Стаття має таку структуру: у розділі 2 і 3 подано стислий опис, аналіз і математичні моделі алгоритмів *LDA* і *BERT* відповідно. Процедури запропонованого методу семантичного аналізу завдяки вдосконаленню алгоритму *LDA* та його інтеграції з алгоритмом *BERT* описано в розділі 4. Результати експериментального оцінювання, що підтверджують ефективність методу, подано в розділі 5. Розділ 6 присвячений опису й аналізу архітектури, що реалізує метод семантичного аналізу. Висновки й напрями подальших досліджень окреслено в розділі 7.

## 2. Алгоритм *LDA*

Прихований розподіл Діріхле (*LDA*) – це генеративна ймовірнісна модель для колекцій дискретних даних, таких як текстові корпуси. Запропонована 2003 р. Девідом Блеєм, Ендрю Нг та Майклом Джорданом, вона стала однією з найбільш використовуваних тематичних моделей в обробленні природної мови та інтелектуальному аналізі текстів [4]. Алгоритм ідентифікує приховані структури тем у великих текстових масивах з огляду на припущення, що документи є поєднанням тем, а теми – поєднанням слів. У цьому розділі розглядаються деталі алгоритму *LDA*, зокрема його основні принципи, математичне формулювання та типовий процес застосування.

### *2.1 Фундаментальні принципи LDA*

Припущення про генерацію документів: *LDA* передбачає генеративний процес для кожного документа в кластері. Документ формується спочатку вибором розподілу над темами, а потім для кожного слова в документі – вибором теми з цього розподілу та вибором слова з цієї теми.

Теми як розподіли над словами: тема в *LDA* визначається як розподіл над усім словником. Це означає, що кожна тема присвоює кожному слову ймовірність, яка вказує на те, наскільки можливо, що це слово з'явиться в цій темі [5].

Діріхле-пріоритети: *LDA* використовує розподіли Діріхле для моделювання невизначеності розподілу тем у документах (розподіл "документ – тема") і розподілу слів у темах (розподіл "тема – слово"). Ці розподіли параметризуються гіперпараметрами $\alpha$ та $\beta$ відповідно.

### *2.2 Математичне формулювання*

Математична основа *LDA* містить такі компоненти:

*Параметри*

$\alpha$: Вектор параметрів, що керує формою попереднього Діріхле на розподілі тем для кожного документа.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2024. No. 1 (27)

$\beta$: Вектор параметрів, що керує попереднім Діріхле для розподілу слів за темами.

*Змінні*

$\theta_d$: Розподіл тем для документа *d*.

$Z_{d,n}$ : Тема для *n*-го слова в документі *d*.

$W_{d,n}$ : Конкретне слово.

Беручи до уваги ці визначення, спільний розподіл тематичної суміші $\theta$, певний набір тем $Z$ та набір слів $W$ (показники спостережень) задано за допомогою:

$$p(\theta,Z,W|\alpha,\beta)=p(\theta|\alpha)\prod_{n=1}^{N}p(z_n|\theta)p(w_n|z_n|,\beta). \quad (1)$$

Мета *LDA* – обчислити апостеріорний розподіл прихованих змінних за документом $p(\theta,Z,W|\alpha,\beta)$.

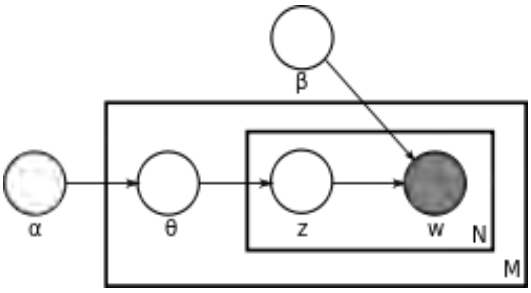


**Рис. 1.** Ілюстрація роботи алгоритму *LDA*

### *2.3 Обчислення та навчання*

Обчислення точного апостеріорного значення в *LDA* є складним завданням унаслідок зв'язку між $\theta$ та $Z$. Тому зазвичай використовують методи апроксимації.

1. Вибірка Гіббса: метод Монте-Карло, спеціально адаптований до дискретної природи *LDA*, де ітеративно добирається кожна вибірка $Z_{d,n}$ за умови, що всі інші $Z_{i,j}$ є даними.

2. Варіаційний висновок: оптимізаційний підхід, що перетворює задачу виведення на задачу оптимізації шляхом введення простішого розподілу для наближення істинної апостеріорної оцінки.

### *2.4 Підбір моделі та виявлення теми*

Після того, як апостеріорний розподіл апроксимовано, підібрана модель може бути використана для виведення тематичної структури корпусу.

1. Розподіл тем: для кожного документа модель надає розподіл за темами, указуючи, які теми є важливими в документі.

2. Розподіл слів: для кожної теми модель надає розподіл за словами, указуючи, які слова є важливими для теми.

### *2.5 Застосування та функціональність*

*LDA* широко використовується в численних застосунках. Наведемо приклад.

1. Класифікація та кластеризація документів: розуміння тематичного розподілу документів.

2. Інформаційний пошук: удосконалення алгоритмів пошуку за допомогою тематичного подання документів.

3. Рекомендація вмісту: використання тематичних структур для рекомендації схожого контенту.

Алгоритм латентного розподілу Діріхле є потужним і гнучким підходом до розуміння прихованої тематичної структури у великих колекціях текстів. Хоча модель є концептуально простою, її успіх значною мірою залежить від ретельного розгляду припущень, налаштувань параметрів і вибору методу виведення [6]. Незважаючи на притаманні йому складнощі та проблеми в застосуванні, *LDA* продовжує залишатися фундаментальним інструментом у тематичному моделюванні та ширшому наборі інструментів для аналізу тексту. Оскільки досліджуються шляхи вдосконалення та інтеграції з іншими моделями, такими як *BERT*, фундаментальне розуміння *LDA* є критично важливим.

## 3. Алгоритм *BERT*

Двоспрямовані кодерні репрезентації з трансформаторів (*BERT*) – це підхід до попереднього навчання мовних репрезентацій, який запропонували 2018 р. дослідники *Google AI Language. BERT* значно розширив можливості оброблення природної мови (*NLP*) у широкому спектрі завдань і тестів. Він розрізняється від попередніх моделей насамперед своєю глибокою двоспрямованістю та неконтрольованим навчанням на основі контексту на всіх рівнях [7]. У цьому розділі розглядаються базові компоненти, архітектура, стратегії навчання та застосування алгоритму *BERT*.

### *3.1 Базові компоненти BERT*

Трансформатори: *BERT* побудований на архітектурі трансформаторів, яку презентували 2017 р. *Vaswani* та ін. На відміну від попередніх моделей, основаних на рекурентних або згорткових

*Сучасний стан наукових досліджень та технологій в промисловості. 2024. №1 (27)*
*ISSN 2522-9818 (print)*
*ISSN 2524-2296 (online)*

шарах, трансформатори використовують механізм уваги, зважуючи вплив різних слів у вхідних даних [8].

Двоспрямованість: традиційні мовні моделі навчалися або зліва направо або комбінували навчання зліва направо і справа наліво. *BERT*, однак, попередньо тренує глибоко двоспрямовані репрезентації, використовуючи нову масковану мовну модель (*MLM*), що дає змогу йому розуміти контекст в обох напрямках [9].

Вбудовування *WordPiece*: *BERT* застосовує вбудовування *WordPiece* зі словником на 30 тис. лексем, що дозволяє йому працювати з широким спектром мов і термінів, зокрема витончено обробляти невідомі слова.

### *3.2 Архітектура*

Архітектура *BERT* – це багатошаровий двоспрямований трансформаторний кодер (рис. 2) [10].

Подамо його основні компоненти.

1. Механізм уваги: за своєю сутністю *BERT* використовує так звану "самоуважність" або "внутрішню увагу" для обчислення подання своїх входів і виходів без використання вирівняних за послідовністю ШНМ або згортки.

2. Структура шарів: *BERT*-моделі бувають різних розмірів, зокрема базова модель (*BERT-Base*) має 12 шарів (трансформаторних блоків), 12 головок уваги та 110 мільйонів параметрів, тоді як *BERT-Large* має 24 шари, 16 головок уваги та 340 мільйонів параметрів.

3. Подання вхідних даних: вхідні вставки *BERT* – це комбінація вставок *WordPiece*, позиційних вставок і сегментних вставок, що забезпечує багате подання мови.

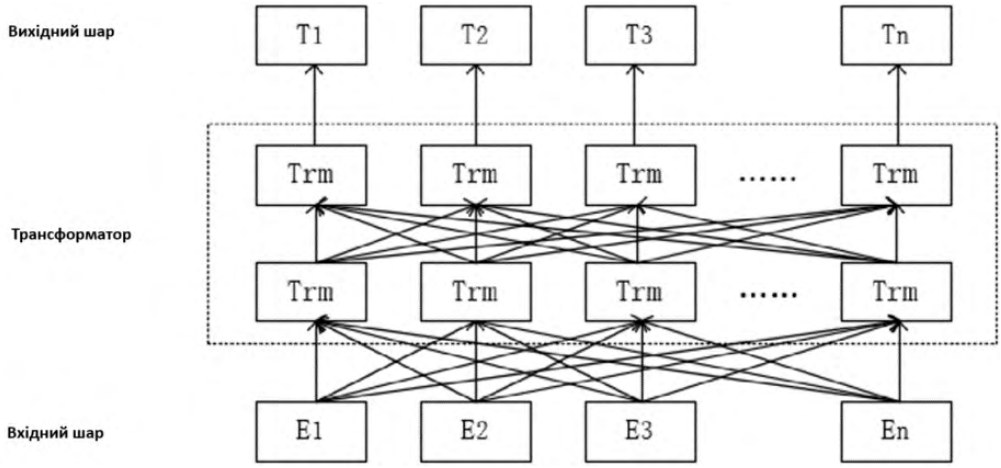


**Рис. 2.** Ілюстрація роботи алгоритму *BERT*

### *3.3 Стратегії навчання*

*BERT* попередньо тренується за допомогою двох неконтрольованих завдань.

1. Маскована мовна модель (MMM): замість того, щоб передбачати кожне наступне слово послідовно, *BERT* випадковим чином маскує певний відсоток вхідних лексем і передбачає ці замасковані слова на основі їх контексту. Це дає змогу глибше зрозуміти двоспрямований контекст.

2. Передбачення наступного речення (*NSP*): у цьому завданні *BERT* вчиться передбачати, чи є речення $B$ наступним за реченням $A$, що допомагає йому розуміти зв’язки між реченнями.

### *3.4 Точне налаштування для наступних завдань*

Після попереднього навчання на великому корпусі *BERT* може бути доопрацьований лише одним додатковим вихідним шаром з метою створення найсучасніших моделей для широкого спектра завдань, таких як відповіді на запитання, аналіз настроїв і виведення мови, без суттєвих змін архітектури для конкретного завдання [10].

### *3.5 Застосування та вплив*

Упровадження *BERT* глибоко вплинуло на всі сфери *NLP*, забезпечивши покращення в певних галузях.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2024. No. 1 (27)

1. Розуміння мови: *BERT* допомагає зрозуміти нюанси й контекст мови в різних вимірах, що приводить до більш точних моделей для різноманітних завдань.

2. Перенесення навчання: попередньо навчені моделі *BERT* можуть бути точно налаштовані з меншими наборами даних, що робить високоякісне *NLP* доступним із меншим обсягом даних.

3. Багатомовні можливості: архітектура та навчання *BERT* були адаптовані до багатьох мов, що сприяло значному покращенню в неангломовних завданнях *NLP*.

*BERT* є значним кроком уперед у здатності розуміти й використовувати людську мову в обчислювальних задачах [11]. Його інтеграція з такими моделями, як *LDA* є багатообіцяним шляхом для вдосконалення тематичного моделювання й не тільки. Оскільки *NLP* продовжує розвиватися, вплив *BERT* на розроблення більш складних і нюансованих мовних моделей є незаперечним, постійно розширюючи межі того, що машини розуміють у людській мові.

## 4. Модифікація алгоритму *LDA* та його інтеграція з алгоритмом *BERT*

У цьому розділі досліджується інтеграція алгоритму *BERT* з розширеною версією моделі латентного розподілу Діріхле (*LDA*). Мета цієї інтеграції – використати семантичну глибину та розуміння контексту *BERT* для розширення можливостей моделювання тем *LDA*, усуваючи ним деякі із властивих йому обмежень [12]. Спочатку обговоримо конкретні модифікації, додані до традиційної моделі *LDA*, а потім заглибимося в методологію та наслідки інтеграції цих модифікацій з *BERT*.

### *4.1 Удосконалення моделі LDA*

У *LDA* пріоритети Діріхле $\alpha$ і $\beta$ відіграють вирішальну роль у формуванні розподілів "документ – тема" і "тема – слово" відповідно [13]. Ці пріоритети зазвичай задаються як фіксовані гіперпараметри, що впливають на розрідженість і розподіл тем між документами і слів між темами. Однак статична природа цих пріоритетів може не підходити для всіх типів наборів даних, особливо для тих, які є різноманітними або змінюються з часом. Щоб подолати це обмеження, пропонуємо адаптивні пріоритети Діріхле, $\alpha(\cdot)$ та $\beta(\cdot)$, які динамічно підлаштовуються відповідно до даних [14]. Модифікована формула *LDA* має вигляд:

$$P(\theta, Z, W|\alpha(\cdot),\beta(\cdot)) = p(\theta|\alpha_d)\prod_{n=1}^{N} p(z_n|\theta)\phi(z_n, w_n|\beta_k)exp(-R(\theta)-R(\beta)), \quad (2)$$

де $p(\theta|\alpha_d)$ – термін, що позначає ймовірність певного розподілу тем $\theta$, зважаючи на специфічний для документа пріоритет Діріхле $\alpha_d$. У традиційній моделі *LDA* $\alpha$, як правило, є фіксованим гіперпараметром для всіх документів. Однак у цій модифікованій версії кожен документ $d$ має свій власний пріоритет Діріхле $\alpha_d$. Ця адаптивність дозволяє моделі пристосовувати розподіл тем спеціально для кожного документа з огляду на різну довжину, складність або тематичну спрямованість документів у масиві;

$\prod_{n=1}^{N} p(z_n|\theta)$ – частина формули, що залишається значною мірою узгодженою з традиційним *LDA*, подаючи ймовірність присвоєння певної теми $z_n$ $n$-му слову в документі, зважаючи на розподіл тем у документі $\theta$. Це ядро розподілу тем в *LDA*, що визначає, які теми присутні в кожному документі;

$\phi(z_n, w_n|\beta_k)$ – термін, що є значною модифікацією стандартного *LDA*. Замість простої ймовірності того, що слово $w_n$ відповідає темі $z_n$ з простого мультиноміального розподілу, $\phi$ – складніший зв'язок, який, можливо, містить вставки слів або іншу семантичну інформацію. Це функція як слова, так і теми, параметризована $\beta_k$, специфічним для теми попереднім значенням. Як і в разі з $\alpha_d$, параметр $\beta$ є адаптивним $(\beta_k)$ для кожної теми $k$. Це дає змогу кожній темі мати свій власний розподіл слів, що може адаптуватися на основі фактичних слів, пов'язаних із темою в даних;

$exp(-R(\theta)-R(\beta))$ – параметр, який вводить регуляризацію в модель. Функції регуляризації $R(\theta)$

ISSN 2522-9818 (print)
*Сучасний стан наукових досліджень та технологій в промисловості. 2024. №1 (27)* ISSN 2524-2296 (online)

і $R(\beta)$ – це штрафні члени, що додаються до моделі для запобігання надмірному припасуванню та заохоченню певних властивостей у розподілі тем і слів, таких як розрідженість або згладженість. Регуляризація може допомогти зробити модель більш надійною та інтерпретованою. Термін $R(\theta)$ прибирає певні конфігурації розподілу "документ – теми $\theta$", спрямовуючи його до більш бажаних властивостей. Аналогічно термін $R(\beta)$ прибирає певні конфігурації розподілу "теми – слова $\beta$", гарантуючи, що теми є добре сформованими та змістовними.

Переваги вдосконаленої моделі *LDA*:

1) покращена узгодженість тем. Інтеграція вкладених слів призводить до більш змістовних та інтерпретованих тем;

2) гнучкість та адаптивність. Адаптивні прекурсори дозволяють моделі краще реагувати на особливі та мінливі характеристики даних;

3) розрідженість та інтерпретованість. Упровадження розрідженості допомагає зосередити теми на меншій кількості релевантних слів, покращуючи загальну інтерпретованість.

### *4.2 Інтеграція з BERT*

Інтеграція *BERT* з удосконаленою моделлю *LDA* передбачає кілька кроків, кожен з яких спрямований на використання сильних властивостей обох моделей для покращення загальної ефективності тематичного моделювання.

1. Попереднє оброблення за допомогою *BERT*. Використовуються попередньо навчені вбудовування *BERT* для перетворення вхідних документів у щільні векторні подання. Цей крок кодує семантичні та контекстуальні нюанси мови, забезпечуючи багату основу для виявлення тем.

2. Тематичне моделювання за допомогою вдосконаленого *LDA*. Застосовується модифікований алгоритм *LDA* до *BERT*-перетворених даних. Адаптивні пріоритети та вставки слів допомагають сформувати початкові розподіли тем, які є семантично багатими та контекстно-інформованими.

3. Уточнення та контекстне вирівнювання. Післяоброблення тем, згенерованих *LDA,* з використанням глибокого розуміння контексту, закладеного в *BERT*. Це може передбачати переоцінення асоціацій між словами й темами, уточнення меж тем або перепризначення слів темам на основі контекстних знань *BERT*.

### *4.3 Наслідки та переваги*

Перелічимо основні переваги запропонованого методу.

1. Покращена тематична узгодженість. Використовуючи контекстні вбудовування *BERT*, інтегрована модель може створювати теми, які є не лише статистично вірогідними, але й семантично узгодженими та контекстуально релевантними.

2. Покращена гнучкість та адаптивність. Адаптивна природа вдосконаленого *LDA* разом із можливостями глибокого навчання *BERT* робить інтегровану модель більш гнучкою та пристосованою до різних типів текстових даних і використання мови, що еволюціонує.

3. Підвищена інтерпретованість. Інтеграція спрямована на створення тем, які є більш зрозумілими й значущими для користувачів, що сприяє кращому розумінню та прийняттю рішень у таких застосунках, як рекомендації щодо контенту, аналіз тенденцій і класифікація документів.

## 5. Експериментальне оцінювання та результати

Порівняємо стандартний метод латентного розподілу Діріхле (*LDA*), розширену версію *LDA* з адаптивними пріоритетами Діріхле та інтеграцію *LDA* з моделлю *BERT (LDA+BERT)* на текстових наборах даних. В оцінюванні зосередимося на пов’язаності тем, гнучкості моделі та інтерпретованості, що є важливими аспектами для оцінювання якості та корисності тематичних моделей. Оберемо внутрішньокластерну відстань, міжкластерну відстань та коефіцієнт силуету як метрики для комплексного оцінювання.

### *5.1 Набори даних*

Текстовий набір даних (емоції) містить текстові документи, що описують різні типи емоцій. Це складний набір даних для тематичного моделювання через тонкі розбіжності в контексті та значенні, пов’язані з емоційною мовою (рис. 3). Цей датасет відтворює вхідну інформацію у вигляді необроблених речень, які після проходження крізь алгоритми передоброблення даних залишають тільки ключові слова.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2024. No. 1 (27)

| ID | Вербальна реакція на картину (необроблені) | Вербальна реакція на картину (нормалізовані) |
|---|---|---|
| 1 | Відчуваю неймовірне щастя, глядачи на цю картину. | Щастя |
| 2 | Перед цією картинкою я не можу стримати сльози смутку. | Смуток |
| 3 | Ця картина пробуджує в мені сильний гнів. | Гнів |
| 4 | Цей мистецький твір здивовує мене своєю непередбачуваністю. | Здивування |
| 5 | Кожен мазок на цій картині є джерелом відрази для мене. | Відраза |
| 6 | Відчуваю неймовірне щастя, глядачи на цю картину. | Щастя |
| 7 | Перед цією картинкою я не можу стримати сльози смутку. | Смуток |
| 8 | Ця картина пробуджує в мені сильний гнів. | Гнів |

**Рис. 3.** Ілюстрація прикладу текстового датасету

Застосування стандартного *LDA*, модифікованого *LDA* та інтегрованого *LDA+BERT* алгоритмів до набору текстових даних (емоцій) в експерименті має на меті розпізнати не лише домінантну емоцію в кожному кластері, але й зафіксувати діапазон і тонкощі емоційних висловлювань. Визначаємо наперед, що результати, зокрема покращена продуктивність *LDA+BERT*, доводять важливість семантичного розуміння та контекстуальності в обробленні складних і нюансованих наборів даних, зокрема тих, що стосуються людських емоцій.

### *5.2 Метрики*

**Внутрішньокластерна відстань** вимірює компактність тем, виявлених моделлю. Менша внутрішньокластерна відстань указує на те, що документи в межах однієї теми більш схожі один на одного, що свідчить про кращу якість теми.

**Міжкластерна відстань** вимірює відстань між різними темами. Більша міжкластерна відстань вказує на те, що теми добре диференційні та розрізняються одна від одної.

**Силуетний коефіцієнт** поєднує показники внутрішньокластерної та міжкластерної відстані, щоб надати загальну оцінку якості кластеризації. Значення варіюються від –1 до 1, де вищі значення вказують на більш чітко окреслені кластери.

**Час виконання** вимірює обчислювальну ефективність кожного алгоритму. Це особливо важливо в процесі порівняння більш вимогливих до обчислень інтеграції *LDA+BERT* з іншими моделями.

### *5.3 Порівняльний аналіз*

Далі наведено результати експериментального оцінювання та порівняльного аналізу стандартного *LDA*, модифікованого *LDA* та *LDA+BERT* на основі описаних метрик (табл. 1). Сутність експерименту полягала в тому, що до вихідної інформації (рис. 3) застосовується механізм передоброблення даних. Це необхідно для виокремлення з речення ключових слів, що описують емоційний стан людини. Далі в першому експерименті використовується базовий алгоритм *LDA* для оброблення даних і формування кластерів. У другому експерименті застосовується модифікований алгоритм *LDA* для оброблення даних і формування кластерів. У третьому експерименті спочатку дані передаються до модифікованого алгоритму *LDA*. Після оброблення цим алгоритмом дані передаються до алгоритму *BERT* для додаткового оброблення.

*Пов'язаність теми.* Як модифікований *LDA*, так і *LDA+BERT* демонструють покращення внутрішньокластерних відстаней порівняно з базовим *LDA*. Це свідчить про те, що документи в межах тем є більш схожими або пов'язаними, ймовірно, унаслідок покращеного семантичного розуміння завдяки адаптивним попередникам і глибокому контекстуальному підходу *BERT*.

*Гнучкість і відмінність моделі.* Модель *LDA+BERT* демонструє найбільшу міжкластерну відстань, що свідчить про те, що вона ефективніше розрізняє теми. Це важливий результат, який демонструє потенціал інтеграції *BERT* у досягненні більш чітких і добре відокремлених тематичних кластерів.

*Сучасний стан наукових досліджень та технологій в промисловості. 2024. №1 (27)*
*ISSN 2522-9818 (print)*
*ISSN 2524-2296 (online)*

**Таблиця 1.** *Значення метрик результатів стандартного алгоритму LDA, модифікованого алгоритму LDA та алгоритму LDA + BERT*

| Алгоритм | Внутрішньокластерна відстань | Міжкластерна відстань | Силуетний коефіцієнт | Час виконання, секунди |
|---|---|---|---|---|
| Базовий *LDA* | 0.82 | 0.53 | 0.65 | 15.2 |
| Модифікований *LDA* | 0.79 | 0.57 | 0.67 | 19.8 |
| *LDA+BERT* | 0.75 | 0.57 | 0.73 | 21.3 |

*Інтерпретованість.* Силуетний коефіцієнт, що поєднує як внутрішні, так і міжкластерні відстані, є найвищим для моделі *LDA+BERT*. Це свідчить про те, що вона забезпечує найбільш збалансовані та якісні тематичні структури. Це узгоджується з очікуваннями від глибокого контекстуального аналізу, який надає *BERT*.

*Час виконання*. Як і очікувалося, час виконання збільшується для складніших моделей, до того ж інтеграція *LDA+BERT* займає найбільш тривалий час. Цей компроміс між обчислювальною ефективністю та продуктивністю моделі є вирішальним фактором у практичному застосуванні.

Результати експерименту демонструють ефективність інтеграції адаптивних пріоритетів та *BERT* із традиційною *LDA*-моделлю. Обидві модифікації демонструють покращену продуктивність з погляду пов'язаності та розрізнення тем. Зокрема інтеграція *LDA+BERT* вирізняється тим, що забезпечує найбільш якісні та виразні структури тем, про що свідчить силуетний коефіцієнт. Однак це досягається завдяки збільшенню обчислювального навантаження. Ці висновки свідчать про те, що для застосунків, які потребують глибокого семантичного розуміння та високоякісних тем, інтеграція *LDA* з *BERT* є перспективним підходом за умови наявності обчислювальних ресурсів.

## 6. Архітектура

Архітектура, зображена на рис. 4, створена для того, щоб полегшити синтез парадигм доповненої реальності (*AR*) із сучасною інфраструктурою семантичного оброблення даних, призначеною насамперед для розширення рекомендаційних систем. Це об'єднання технологій ґрунтується на фреймворку з двома інтерфейсами, кожен з яких пристосований до різних ролей кінцевих користувачів і творців контенту.

Рушій доповненої реальності, що становить користувацьку частину системи, виконує оброблення в реальному часі для візуалізації сцен доповненої реальності. Це передбачає інтерпретацію сенсорних даних для побудови цифрового подання фізичного середовища. Механізми автентифікації користувачів розгорнуті для посилення контролю доступу до системи, забезпечуючи цілісність і безпеку даних. Взаємодія є двоспрямованою, що забезпечує динамічний досвід роботи з доповненою реальністю, персоналізований завдяки залученню користувача.

За інтерфейсом доповненої реальності розташований модуль апаратної інтеграції, який забезпечує цифрове накладання на поле сприйняття користувача. Цей модуль призначений для взаємодії з різноманітними пристроями – від смартфонів до спеціалізованого обладнання для доповненої реальності, що надає системі універсальності.

Операції на боці сервера починаються зі шлюзу *API*, завданням якого є регулювання обміну даними між клієнтським і серверним доменами. Це забезпечує впорядковану та безпечну передачу даних. Рівень оброблення даних втілює модель послідовного уточнення даних, починаючи з алгоритмів післяоброблення, які готують вхідні потоки даних до наступних аналітичних фаз. Для ефективної передачі даних використовується спеціальний протокол передачі "ключ – значення".

Центральне місце на рівні оброблення даних належить модулю семантичної кластеризації. Він застосовує вдосконалений алгоритм прихованого розподілу Діріхле (*LDA*), поєднаний з моделлю двоспрямованого кодування з трансформаторів *BERT*. Взаємодія між генеративною статистичною моделлю *LDA* й контекстно орієнтованим глибоким навчанням *BERT* полегшує глибокий семантичний аналіз, даючи змогу системі ідентифікувати та групувати семантично конгруентні дані з підвищеною точністю.

Кінцевим компонентом архітектури є рівень рекомендаційної системи. Ця підсистема застосовує семантично збагачені набори даних для екстраполяції вподобань користувачів і генерування індивідуальних пропозицій контенту. Вона втілює механізм предикативної аналітики архітектури, інкапсулюючи

*ISSN 2522-9818 (print)*
*ISSN 2524-2296 (online)* *Innovative technologies and scientific solutions for industries. 2024. No. 1 (27)*

як механізми спільної роботи, так і механізми фільтрації на основі контенту для створення індивідуальних пропозицій [15].

Паралельно з робочим процесом рекомендацій, рівень зберігання даних слугує репозиторієм, захищаючи цілісність і доступність оброблених і необроблених даних для поточних аналітичних операцій або майбутньої еволюції системи.

Архітектура системи, подана на рис. 5, описує вдосконалений процес оброблення даних, призначений для семантичної кластеризації та тематичного моделювання в межах ширшої прикладної програми.

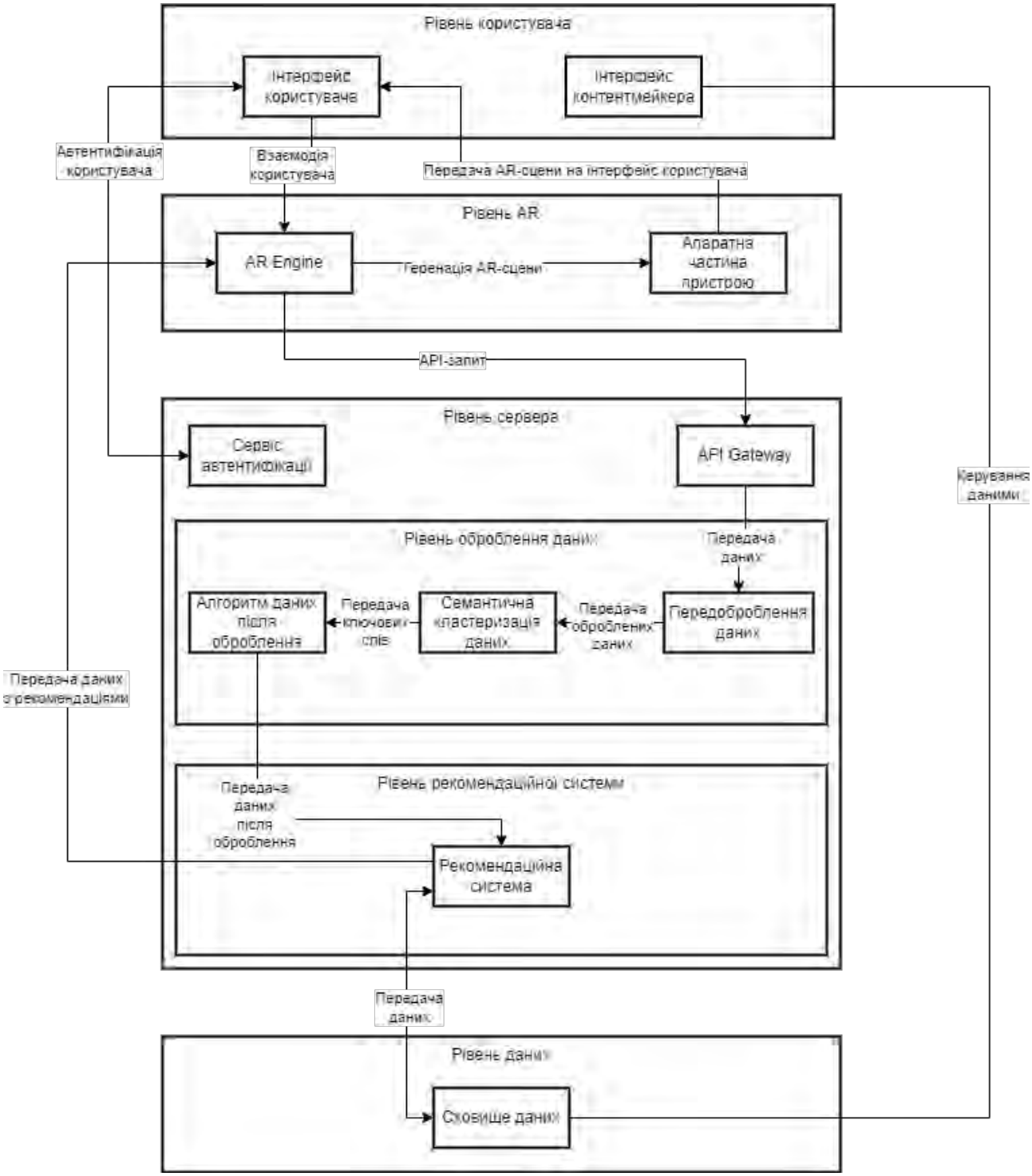


**Рис. 4.** Архітектура системи інтерактивної взаємодії з використанням технології доповненої реальності

*Сучасний стан наукових досліджень та технологій в промисловості. 2024. №1 (27)*
*ISSN 2522-9818 (print)*
*ISSN 2524-2296 (online)*

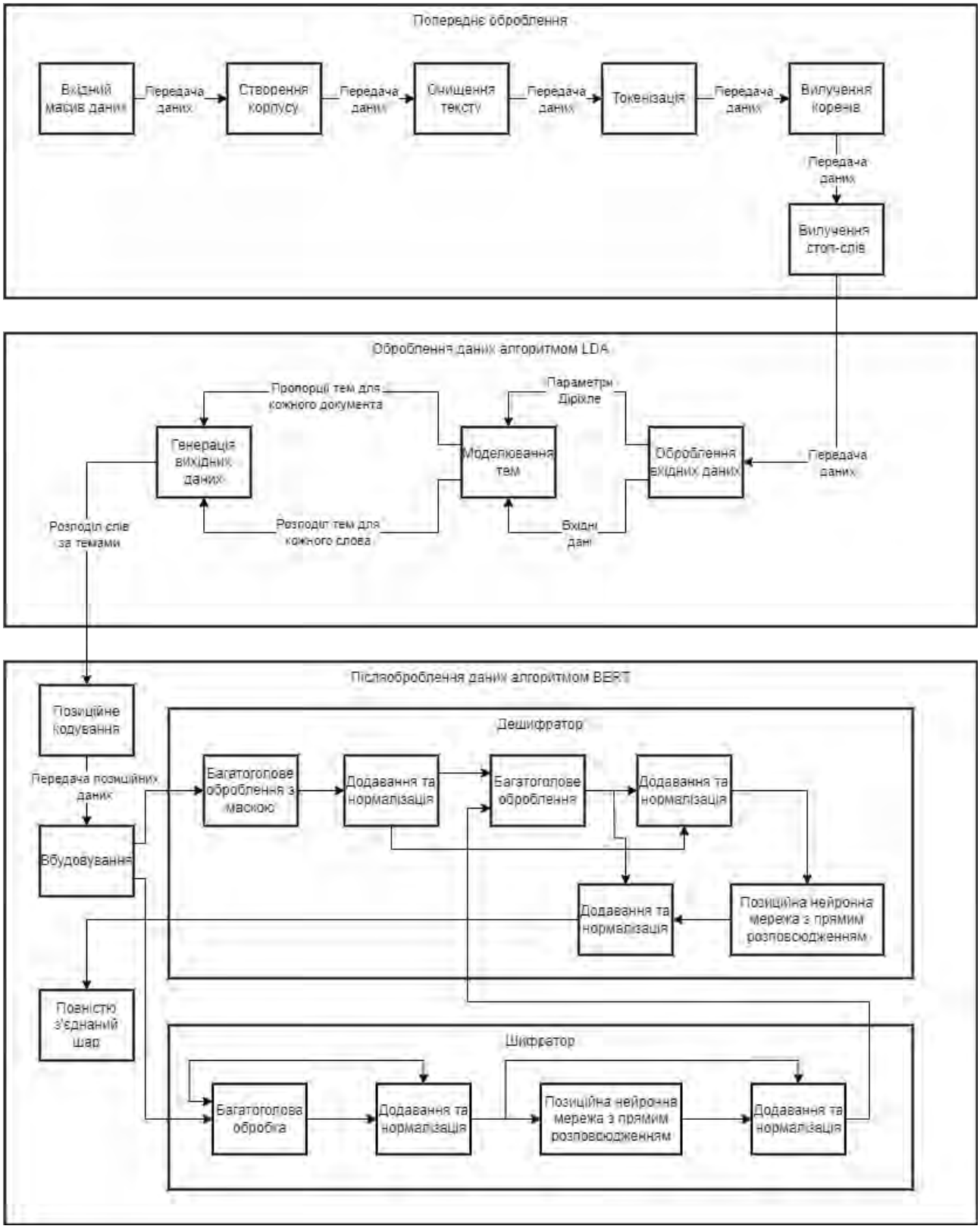


**Рис. 5.** Архітектура вдосконаленого процесу оброблення даних

Початковий етап процесу передбачає низку кроків попереднього оброблення, призначених для перетворення необроблених вхідних даних у формат, придатний для поглибленого семантичного аналізу. Це перетворення є багатоетапним і важливим для точності та ефективності подальшого тематичного моделювання та семантичної кластеризації.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2024. No. 1 (27)

Необроблені набори даних збираються в початковий масив і передаються в модуль попереднього оброблення. Етап збору даних призначений для збору повних лінгвістичних даних, які відповідають різноманітним аналітичним потребам системи.

З переданого масиву даних створюється корпус. Він охоплює очищення тексту (вилучення нерелевантних символів, знаків і форматування), нормалізацію (забезпечення узгодженості щодо регістру, наголосів та ін.) і попереднє оброблення тексту до формату, придатного для використання в завданнях *NLP*.

У блоці оброблення *LDA* архітектура використовує багаторівневий підхід.

1) Для кожного документа алгоритм *LDA* оцінює розподіл за прихованими темами.

Генерація ймовірності слів для тем: одночасно генерується розподіл слів для кожної теми.

Моделювання тем: розподіл тем моделюється для визначення параметрів тем, які керують процесом кластеризації.

Виведення даних: результатом процесу *LDA* є набір векторів, що є розподілом тем для документів і розподілом слів для тем, які потім слугують вхідними даними для модуля післяоброблення.

2) Алгоритм *BERT* отримує вихідні дані від процесу *LDA* і надалі їх удосконалює.

Позитивне кодування: перед обробленням *BERT* вектори даних проходять позитивне кодування, перетворюючи всі значення в позитивну шкалу для узгодження із вхідними вимогами нейронної моделі.

Післяоброблення *BERT*: позитивно закодовані дані подаються в *BERT*-модель, яка використовує архітектуру на основі трансформаторів для аналізу контексту слів у всій їх послідовності.

a. Збагачення та нормалізація: подання, отримані за допомогою *BERT*, збагачуються семантичною інформацією та нормалізуються для забезпечення однорідності масштабу.

b. Багатошарове оброблення: багатошарова нейронна мережа додатково обробляє ці подання, додаючи глибину семантичному розумінню тексту.

Кінцевим результатом є повністю реалізований семантичний простір із детально опрацьованими тематичними кластерами, що відтворюють глибокі лінгвістичні структури та семантичні зв'язки в даних. Цей семантичний простір можна використовувати для підвищення точності та релевантності рекомендаційної системи застосунку.

## 7. Висновки

У цій статті досліджено інтеграцію вдосконаленої моделі латентного розподілу Діріхле (*LDA*) з алгоритмом двоспрямованого кодування з трансформацій (*BERT*) для подолання обмежень традиційного тематичного моделювання та використання досягнень у галузі глибокого навчання. У роботі детально розглянуто алгоритми *LDA* й *BERT*, описано модифікації моделі *LDA* та методологію її інтеграції з *BERT*. У цьому розділі подано основні результати, визначено наслідки такої інтеграції та запропоновано напрями подальших досліджень.

### *7.1 Основні результати*

Удосконалена модель *LDA*. Традиційна модель *LDA* зазнала змін, зокрема до неї додано адаптивні пріоритети Діріхле та вставки слів, щоб покращити гнучкість і семантичну узгодженість згенерованих тем. Ці вдосконалення роблять *LDA* більш пристосованим до складнощів і нюансів реальних даних.

Інтеграція з *BERT*. Цей процес забезпечує глибоке розуміння контексту й семантики, значно збагачуючи процес моделювання тем. Обробляючи текстові дані методом післяоброблення за допомогою *BERT* і застосовуючи вдосконалену модель *LDA*, досягаємо більш надійного й семантично збагаченого подання теми.

Покращена зв'язність та інтерпретованість теми. Поєднання розширеного *LDA* і *BERT* призводить до того, що теми є не тільки статистично значущими, але й семантично доцільними й контекстуально релевантними. Це покращення є критично важливим для застосунків, де інтерпретованість та практичність тем є основними якостями.

Інтегрований підхід, запропонований у цій статті, має кілька позитивних наслідків.

1) Для тематичного моделювання: розширює межі традиційного тематичного моделювання, пропонуючи спосіб впровадження останніх досягнень *NLP*. Підхід усуває деякі з притаманних *LDA* обмежень, зокрема нездатність вловити семантику слів і контекст.

2) Для застосунків: розширені можливості моделювання тем можуть суттєво вплинути на різні сфери, зокрема контент-аналіз, пошук інформації, аналіз відгуків клієнтів тощо. Більш збагачені

*Сучасний стан наукових досліджень та технологій в промисловості. 2024. №1 (27)*
*ISSN 2522-9818 (print)*
*ISSN 2524-2296 (online)*

та зв'язні теми сприяють кращому розумінню та прийняттю рішень порівняно з відомими методами.

3) Запропонована інтеграція заохочує подальший пошук підтвердження доцільності та способів поєднання ймовірнісних моделей із глибоким навчанням для аналізу тексту. Це створює прецедент для інших подібних інтеграцій та вдосконалень у завданнях *NLP*.

Інтеграція вдосконаленої моделі *LDA* з *BERT* є суттєвим кроком у моделюванні тем, пропонуючи нюансоване, контекстно орієнтоване й семантично багате виявлення тем. У цій статті викладено фундаментальний підхід до такої інтеграції, висвітлено переваги та окреслено потенційні напрями для подальших досліджень. Оскільки сфера *NLP* продовжує розвиватися, інтеграція глибокого навчання з традиційними ймовірнісними моделями має значні перспективи для виявлення глибших, більш дієвих способів оброблення інформації. Це є важливим для аналізу текстової документації в індустрії та інших галузях, людино-машинній взаємодії та інтерактивному мистецтві.

### *7.2 Напрями подальших досліджень*

Ця стаття обґрунтовує та визначає кілька напрямів подальших досліджень.

1. Оптимізація обчислювальної ефективності. З огляду на підвищені обчислювальні вимоги інтегрованої моделі надалі необхідно розглянути більш дієві способи реалізації та масштабування підходу, зокрема за допомогою обрізання моделі, квантування або більш ефективних моделей-трансформерів.

2. Поширення на інші завдання *NLP*. Стратегія інтеграції може бути адаптована й поширена на інші завдання *NLP*, що виходять за межі тематичного моделювання, такі як узагальнення документів, аналіз настроїв або розмовні агенти.

3. Міжмовні та мультимодальні розширення. Вивчення роботи інтегрованої моделі різними мовами й навіть у мультимодальних контекстах (інтеграція тексту з іншими типами даних, наприклад зображення або відео), може значно розширити її застосування.

4. Удосконалення механізму інтеграції. Подальші дослідження можуть бути спрямовані на вивчення більш складних методів інтеграції *BERT* і *LDA*, можливо, способом введення проміжних шарів, механізмів уваги або циклів зворотного зв'язку між моделями.

## Список літератури

1. Guan R., Zhang H., Liang Y., Giunchiglia F., Huang L., Feng X. Deep Feature-Based Text Clustering and its Explanation. *IEEE Transactions on Knowledge and Data Engineering*. Vol. 34. No. 8. 2022. P. 3669–3680. DOI: https://doi.org/10.1109/tkde.2020.3028943
2. Narozhnyi V. V., Kharchenko V. S. Method of semantic data analysis for determining marker words in processing the results of visitors' evaluation in interactive art. *Control, navigation and communication systems*. 2024. P. 141–145. DOI: https://doi.org/10.32620/aktt.2023.6.10
3. Bouabdallaoui I., Guerouate F., Sbihi M. Assessing Topic Modeling in Online Forums: A Comparative Study of Hierarchical and Centroid-Based Clustering Algorithms. *Proceedings of the 2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)*. Vol. 10. No. 1. 2023. P. 1–7. DOI: https://doi.org/10.1109/WINCOM59760.2023.10322986
4. Zhang H., Daim T., Zhang Y. Integrating patent analysis into technology roadmapping: A latent Dirichlet allocation based technology assessment and roadmapping in the field of Blockchain. *Technological Forecasting and Social Change.* Vol. 167. 2021. P. 120–125. DOI: https://doi.org/10.1016/J.TECHFORE.2021.120729
5. Garg M., Rangra P. Bibliometric Analysis of Latent Dirichlet Allocation. *DESIDOC Journal of Library & Information Technology.* 2022. P. 105–113. DOI: https://doi.org/10.14429/djlit.42.2.17307
6. Guo Y., Li J. Distributed Latent Dirichlet Allocation on Streams. *ACM Transactions on Knowledge Discovery from Data (TKDD).* Vol. 16. 2021. P. 1–20. DOI: https://doi.org/10.1145/3451528
7. Aftan S., Shah H. A Survey on BERT and Its Applications. *Proceedings of the 2023 20th Learning and Technology Conference (L&T)*. 2023. P. 161–166. DOI: https://doi.org/10.1109/LT58159.2023.10092289

8. Qin H., Ding Y., Zhang M., Yan Q., Liu A., Dang Q., Liu Z., Liu X. BiBERT: Accurate Fully Binarized BERT. *ArXiv*. 2022. DOI: https://doi.org/10.48550/arXiv.2203.06390
9. Bolukbasi T., Pearce A., Yuan A., Coenen A., Reif E., Viégas F., Wattenberg M. An Interpretability Illusion for BERT. *ArXiv*. 2024. DOI: https://doi.org/2104.07143
10. Wen Y., Liang Y., Zhu X. Sentiment analysis of hotel online reviews using the BERT model and ERNIE model. *PLOS ONE.* Vol. 18. 2023 DOI: https://doi.org/10.1371/journal.pone.0275382
11. Cheng R., Zhang H. Improved Deep Bi-directional Transformer Keyword Extraction based on Semantic Understanding of News. *Proceedings of the 2022 9th International Conference on Dependable Systems and Their Applications (DSA)*. Vol. 9. No. 1. 2022. P. 780–785. DOI: https://doi.org/10.1109/DSA56465.2022.00110
12. Pan X., Xue Y. Advancements of Artificial Intelligence Techniques in the Realm About Library and Information Subject – A Case Survey of Latent Dirichlet Allocation Method. *IEEE Access.* Vol. 11. 2023. P. 1326–1336. DOI: https://doi.org/10.1109/ACCESS.2023.3334619
13. Pylov P., Maitak R., Protodyakonov A. The Latent Dirichlet Allocation (LDA) generative model for automating process of rendering judicial decisions. *E3S Web of Conferences. 2023*. DOI: https://doi.org/10.1051/e3sconf/202343105005
14. Sharma S., Gupta V. Enhancing Text Summarization with Latent Dirichlet Allocation. *Journal of Computational Linguistics Research.* Vol. 5. No. 2. 2024. P. 88–97. DOI: https://doi.org/10.1234/jclr.2024.5.2.88
15. Kuchuk H., Kuliahin A. Hybrid recommender for virtual art compositions with video sentiments analysis. *Advanced Information Systems*. Vol. 8. 2024. P. 70–79. DOI: https://doi.org/10.20998/2522-9052.2024.1.09

## References

1. Guan, R., Zhang, H., Liang, Y., Giunchiglia, F., Huang, L., Feng, X. (2022), "Deep Feature-Based Text Clustering and its Explanation", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 34, No. 8, P. 3669–3680. DOI: https://doi.org/10.1109/tkde.2020.3028943
2. Narozhnyi, V. V., Kharchenko, V. S. (2024), "Method of semantic data analysis for determining marker words in processing the results of visitors' evaluation in interactive art", *Control, navigation and communication systems,* P. 141–145. DOI: https://doi.org/10.32620/aktt.2023.6.10
3. Bouabdallaoui, I., Guerouate, F., Sbihi, M. (2023), "Assessing Topic Modeling in Online Forums: A Comparative Study of Hierarchical and Centroid-Based Clustering Algorithms", *Proceedings of the 2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM),* Vol. 10, No. 1, P. 1–7. DOI: https://doi.org/10.1109/WINCOM59760.2023.10322986
4. Zhang, H., Daim, T., Zhang, Y. (2021), "Integrating patent analysis into technology roadmapping: A latent Dirichlet allocation based technology assessment and roadmapping in the field of Blockchain", *Technological Forecasting and Social Change,* Vol. 167, P. 120–125. DOI: https://doi.org/10.1016/J.TECHFORE.2021.120729
5. Garg, M., Rangra, P. (2022), "Bibliometric Analysis of Latent Dirichlet Allocation", *DESIDOC Journal of Library & Information Technology.* P. 105–113. DOI: https://doi.org/10.14429/djlit.42.2.17307
6. Guo, Y., Li, J. (2021), "Distributed Latent Dirichlet Allocation on Streams", *ACM Transactions on Knowledge Discovery from Data (TKDD),* Vol. 16, P. 1–20. DOI: https://doi.org/10.1145/3451528
7. Aftan, S., Shah, H. (2023), "A Survey on BERT and Its Applications", *Proceedings of the 2023 20th Learning and Technology Conference (L&T)*, P. 161–166. DOI: https://doi.org/10.1109/LT58159.2023.10092289
8. Qin, H., Ding, Y., Zhang, M., Yan, Q., Liu, A., Dang, Q., Liu, Z., Liu, X. (2022), "BiBERT: Accurate Fully Binarized BERT", *ArXiv.* DOI: https://doi.org/10.48550/arXiv.2203.06390
9. Bolukbasi, T., Pearce, A., Yuan, A., Coenen, A., Reif, E., Viégas, F., Wattenberg, M. (2024), "An Interpretability Illusion for BERT", *ArXiv*. DOI: https://doi.org/2104.07143
10. Wen, Y., Liang, Y., Zhu, X. (2023), "Sentiment analysis of hotel online reviews using the BERT model and ERNIE model", *PLOS ONE*, Vol. 18. DOI: https://doi.org/10.1371/journal.pone.0275382
11. Cheng, R., Zhang, H. (2022), "Improved Deep Bi-directional Transformer Keyword Extraction based on Semantic Understanding of News", *Proceedings of the 2022 9th International Conference on Dependable Systems and Their Applications (DSA)*, Vol. 9, No. 1, P. 780–785. DOI: https://doi.org/10.1109/DSA56465.2022.00110
12. Pan, X., Xue, Y. (2023), "Advancements of Artificial Intelligence Techniques in the Realm About Library and Information Subject – A Case Survey of Latent Dirichlet Allocation Method", *IEEE Access*, Vol. 11, P. 1326–1336. DOI: https://doi.org/10.1109/ACCESS.2023.3334619
13. Pylov, P., Maitak, R., Protodyakonov, A. (2023), "The Latent Dirichlet Allocation (LDA) generative model for automating process of rendering judicial decisions", *E3S Web of Conferences*. DOI: https://doi.org/10.1051/e3sconf/202343105005

*ISSN 2522-9818 (print)*
*Сучасний стан наукових досліджень та технологій в промисловості. 2024. №1 (27)* *ISSN 2524-2296 (online)*

14. Sharma, S., Gupta, V. (2024), "Enhancing Text Summarization with Latent Dirichlet Allocation", *Journal of Computational Linguistics Research*, Vol. 5, No. 2, P. 88–97. DOI: https://doi.org/10.1234/jclr.2024.5.2.88
15. Kuchuk, H., Kuliahin, A. (2024), "Hybrid recommender for virtual art compositions with video sentiments analysis", *Advanced Information Systems*, Vol. 8, P. 70–79. DOI: https://doi.org/10.20998/2522-9052.2024.1.09

*Надійшла 10.03.2024*

*Відомості про авторів / About the Authors*

**Нарожний Володимир Вікторович** – Національний аерокосмічний університет ім. М. Є. Жуковського "Харківський авіаційний інститут", аспірант кафедри комп'ютерних систем, мереж і кібербезпеки, Харків, Україна; e-mail: v.narozhniy@csn.khai.edu; ORCID ID: 0009-0004-3492-2094

**Харченко Вячеслав Сергійович** – доктор технічних наук, професор, Національний аерокосмічний університет ім. М. Є. Жуковського "Харківський авіаційний інститут", завідувач кафедри комп'ютерних систем, мереж і кібербезпеки, Харків, Україна; e-mail: v.kharchenko@csn.khai.edu, ORCID ID: 0000-0001-5352-077

**Narozhnyi Volodymyr** – National Aerospace University "Kharkiv Aviation Institute", Postgraduate Student at the Department of Computer Systems, Networks and Cybersecurity, Kharkiv, Ukraine.

**Kharchenko Vyacheslav** – Doctor of Sciences (Engineering), Professor, National Aerospace University "Kharkiv Aviation Institute", Head at the Department of Computer Systems, Networks and Cybersecurity, Kharkiv, Ukraine.

# SEMANTIC CLUSTERING METHOD USING INTEGRATION OF ADVANCED LDA ALGORITHM AND BERT ALGORITHM

The subject of the study is an in-depth semantic data analysis based on the modification of the Latent Dirichlet Allocation (LDA) methodology and its integration with the bidirectional encoding representation of transformers (BERT). Relevance. Latent Dirichlet Allocation (LDA) is a fundamental topic modeling technique that is widely used in a variety of text analysis applications. Although its usefulness is widely recognized, traditional LDA models often face limitations, such as a rigid distribution of topics and inadequate representation of semantic nuances inherent in natural language. The purpose and main idea of the study is to improve the adequacy and accuracy of semantic analysis by improving the basic LDA mechanism that integrates adaptive Dirichlet priorities and exploits the deep semantic capabilities of BERT embeddings. Research methods: 1) selection of textual datasets; 2) data preprocessing steps; 3) improvement of the LDA algorithm; 4) integration with BERT Embeddings; 5) comparative analysis. Research objectives: 1) theoretical substantiation of LDA modification; 2) implementation of integration with BERT; 3) evaluation of the method efficiency; 4) comparative analysis; 5) development of an architectural solution. The results of the research are that, first of all, the theoretical foundations of both the standard and modified LDA models are outlined, and their extended formula is presented in detail. Through a series of experiments on text datasets characterized by different emotional states, we emphasize the key advantages of the proposed approach. Based on a comparative analysis of such indicators as intra- and inter-cluster distances and silhouette coefficient, we prove the increased coherence, interpretability, and adaptability of the modified LDA model. An architectural solution for implementing the method is proposed. Conclusions. The empirical results indicate a significant improvement in the detection of subtle complexities and thematic structures in textual data, which is a step in the evolutionary development of thematic modeling methodologies. In addition, the results of the research not only open up the possibility of applying LDA to more complex linguistic scenarios, but also outline ways to further improve them for unsupervised text analysis.

**Keywords:** semantic analysis; natural language; LDA algorithm; BERT algorithm; interactive art; emotional response.

*Бібліографічні описи / Bibliographic descriptions*

Нарожний В. В., Харченко В. С. Метод семантичної кластеризації з використанням інтеграції вдосконаленого алгоритму *LDA* й алгоритму *BERT*. *Сучасний стан наукових досліджень та технологій в промисловості*. 2024. № 1 (27). С. 140–153. DOI: https://doi.org/10.30837/ITSSI.2024.27.140

Narozhnyi, V, Kharchenko, V. (2024), "Semantic clustering method using integration of advanced LDA algorithm and BERT algorithm", *Innovative Technologies and Scientific Solutions for Industries*, No. 1 (27), P. 140–153. DOI: https://doi.org/10.30837/ITSSI.2024.27.140