УДК 004.8; 004.93

**А. Бистріцький, І. Толкунов,**
**О. Гавриленко, Н. Богданова**

# АНАЛІЗ СУЧАСНИХ МЕТОДІВ РЕФЕРУВАННЯ ТЕКСТІВ

*Анотація:* У статті досліджується можливості автоматизованого реферування текстів – процесу створення стислих, інформативних викладень великих обсягів текстової інформації. Завдяки поєднанню технологій обробки природної мови (NLP) та машинного навчання комп'ютери можуть аналізувати величезні обсяги текстової інформації, виділяти ключові моменти, відкидати несуттєві деталі та генерувати стислі, інформативні реферати, що зберігають основний зміст оригіналу [1]. Автоматизоване реферування текстів – це не просто технологічний тренд, а необхідність, що допоможе опанувати інформаційний простір та ефективно використовувати знання, закладені в текстах.

*Ключові слова:* автоматизоване реферування текстів, обробка природної мови (NLP), машинне навчання, екстрактивне реферування, анотаційне реферування, абревіативне реферування, TextRank, LSA, виділення ключових фраз, аналіз тексту, інформаційне перевантаження.

## Вступ

Сучасний світ потопає в морі інформації. Щодня створюються, обробляються та поширюються гігабайти текстових даних, і цей потік невпинно зростає. Наукові статті, новини, блоги, соціальні мережі – скрізь нас оточує лавина текстів, що вимагають нашої уваги. В умовах такого інформаційного перевантаження гостро постає проблема ефективної роботи з текстовими даними. Як швидко орієнтуватися в морі текстів, знаходити потрібну інформацію, виділяти головне і не потонути в потоці деталей?

Традиційні методи аналізу текстів, що базуються на ручному перегляді та обробці, стають все менш ефективними в умовах зростаючих обсягів інформації. На допомогу приходять інноваційні технології, засновані на штучному інтелекті. Одним з таких перспективних напрямків є автоматизоване реферування текстів, що дозволяє створювати стислі, інформативні викладення основного змісту оригіналу, не втрачаючи ключових моментів.

Автоматизоване реферування текстів має значний потенціал для вирішення ряду актуальних проблем:

- Економія часу та ресурсів.
- Підвищення ефективності роботи з інформацією.
- Покращення доступності інформації [2].



 ISSN 1560-8956

## Матеріали та методи

Створення ефективної системи автоматизованого реферування текстів – це комплексне завдання, що вимагає поєднання різних методів та технологій. В основі таких систем лежить розуміння людської мови та вміння виділяти ключові моменти в тексті. Для цього використовуються дві основні галузі штучного інтелекту: обробка природної мови (NLP) та машинне навчання.

## Обробка природної мови (NLP)

Обробка природної мови (NLP, Natural Language Processing) – це захоплююча та багатогранна галузь штучного інтелекту, що фокусується на взаємодії між комп'ютерами та людською мовою. Головна мета NLP – наділити комп'ютери здатністю "розуміти", інтерпретувати та маніпулювати людською мовою так само, як це роблять люди. Ця технологія відкриває безліч можливостей для аналізу, розуміння та генерації текстів, що робить її незамінним інструментом в сучасному світі, де обсяги текстової інформації зростають з кожним днем.

NLP охоплює широкий спектр задач. Розпізнавання мови перетворює усне мовлення на текст, відкриваючи двері для голосового управління, транскрипції аудіо та відео, а також розробки інтелектуальних голосових асистентів. Сегментація тексту розділяє текст на речення, слова, пунктуаційні знаки та інші елементи, що допомагає структурувати текст та підготувати його для подальшого аналізу [1]. Морфологічний аналіз визначає частини мови, граматичні форми, роди, числа та інші характеристики слів, що дозволяє комп'ютеру "розуміти" граматичну структуру тексту. Синтаксичний аналіз займається аналізом структури речень, визначенням зв'язків між словами та фразами, побудовою синтаксичних дерев, що допомагає комп'ютеру "розуміти" граматичну правильність та змістовну зв'язність тексту. Семантичний аналіз визначає значення слів, фраз та речень, з'ясовує контекст, виявляє взаємозв'язки між різними частинами тексту, що дозволяє комп'ютеру "зрозуміти" зміст тексту та його головні ідеї.

Давайте розглянемо алгоритм обробки інформації:

Перший етап – це отримання текстової інформації, що слугує вхідними даними для алгоритму. Далі відбувається сегментація та токенізація, де текст розбивається на окремі слова або фрази (токени). На третьому етапі проводиться очищення тексту, де видаляються непотрібні елементи, такі як пунктуаційні знаки, спеціальні символи та зайві пробіли. Очищений текст потім проходить векторизацію та інженерію ознак (четвертий етап), де слова та фрази перетворюються у числові вектори, що відображають їхні характеристики. П'ятий етап – лематизація та стемінг, де слова зводяться до свого основного стовбура (леми) або кореня, щоб уніфікувати їхнє представлення. Отримані векторизовані дані потім передаються алгоритмам

машинного навчання (шостий етап) для аналізу, класифікації, кластеризації чи інших завдань. На заключному, сьомому етапі відбувається інтерпретація результатів, отриманих від алгоритмів машинного навчання [3].

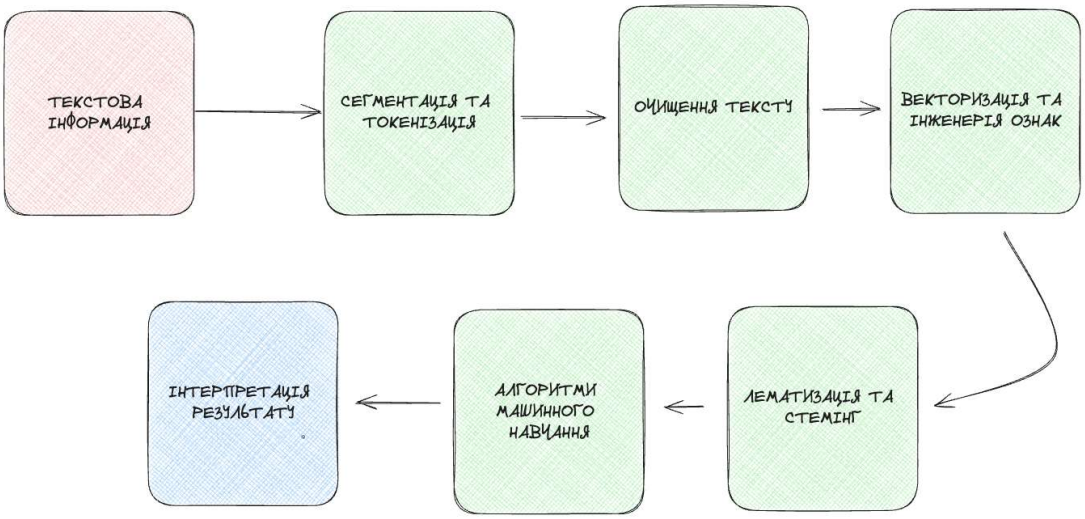


*Рисунок 1.* Алгоритм обробки інформації

В контексті автоматизованого реферування текстів NLP відіграє ключову роль, забезпечуючи фундамент для розуміння та аналізу тексту. NLP допомагає виділити ключові слова та фрази, що найкраще відображають основний зміст тексту, визначити важливість речень та оцінити вагу кожного речення в контексті всього тексту, що дозволяє відфільтрувати менш важливу інформацію. Крім того, NLP допомагає зберегти логічні зв'язки та структуру речень, що робить реферат зрозумілим та легким для сприйняття.

NLP використовує різноманітні методи та алгоритми для обробки текстів. Статистичні методи аналізують частоту появи слів, фраз та інших елементів тексту для визначення їх важливості. Методи машинного навчання використовують алгоритми, що навчаються на основі великих масивів даних, для розпізнавання образів та прогнозування. Глибоке навчання використовує складні нейронні мережі для аналізу тексту та виявлення складних залежностей між словами та фразами [1].

Розвиток NLP відкриває безліч можливостей для створення інтелектуальних систем, здатних ефективно працювати з текстовою інформацією. Це стосується автоматичного реферування текстів – створення стислих та інформативних викладень основного змісту великих текстів, машинного перекладу – автоматичного перекладу текстів з однієї мови на іншу, що полегшує міжнародну комунікацію та доступ до інформації з різних країн світу, чат-ботів – створення віртуальних співрозмовників, здатних розуміти людську мову та відповідати на запитання, надавати інформацію та

 ISSN 1560-8956

виконувати різноманітні завдання, аналізу тональності тексту – визначення емоційного забарвлення тексту, що дозволяє аналізувати відгуки, коментарі, публікації в соціальних мережах та отримувати важливу інформацію про думку аудиторії.

**Машинне навчання**

Машинне навчання являє собою потужний інструмент штучного інтелекту, який наділяє комп'ютери здатністю навчатися на основі даних без необхідності явного програмування. Замість того, щоб розробляти жорсткі правила для кожного можливого сценарію, алгоритми машинного навчання аналізують великі набори даних, виявляючи приховані закономірності та самостійно формуючи моделі. Ці моделі надають комп'ютерам здатність робити прогнози, класифікувати інформацію та вирішувати інші комплексні завдання.

У сфері обробки природної мови машинне навчання відіграє ключову роль, допомагаючи вирішувати широкий спектр задач. Алгоритми машинного навчання здатні аналізувати тексти та класифікувати їх за тематикою, визначаючи, наприклад, чи є текст новиною, науковою статтею, блогом тощо. Крім того, ці алгоритми здатні аналізувати емоційне забарвлення тексту, визначаючи, чи є він позитивним, негативним або нейтральним [4]. Важливою функцією є виділення ключових слів та фраз шляхом аналізу частоти появи слів, їх взаємозв'язків та контексту, що дозволяє виділити ключові елементи, які найкраще відображають зміст тексту. Машинне навчання також знаходить своє застосування в автоматичному реферуванні текстів, де алгоритми здатні створювати стислі та інформативні викладення основного змісту великих текстів, визначаючи важливість речень, виділяючи ключові фрази та генеруючи зв'язний реферат. Ще однією важливою сферою застосування є машинний переклад, де алгоритми, навчаючись на величезних масивах паралельних текстів, здатні автоматично перекладати тексти з однієї мови на іншу, постійно вдосконалюючи якість перекладу.

Існує велика різноманітність алгоритмів машинного навчання, кожен з яких має свої особливості та переваги. Лінійна регресія використовується для прогнозування неперервних значень, таких як ціна акцій або температура повітря. Логістична регресія застосовується для класифікації даних, наприклад, для визначення спаму. Дерева рішень створюють моделі у вигляді дерева, що дозволяє класифікувати дані на основі ряду умов. Метод опорних векторів використовується для класифікації даних шляхом побудови гіперплощини, що розділяє дані різних класів. Нейронні мережі моделюють роботу людського мозку, що дозволяє вирішувати складні задачі, такі як розпізнавання образів та обробка природної мови [5].

Розглянемо два популярні алгоритми, які застосовуються для реферування текстів: PageRank та Latent Semantic Analysis (LSA).

PageRank – це алгоритм, заснований на графах, який використовується для ранжування речень в тексті за їх важливістю. Спочатку будується граф, де речення тексту представляються вершинами, а ребра між вершинами відображають подібність між реченнями, яка може вимірюватися, наприклад, за кількістю спільних слів. Далі алгоритм PageRank за допомогою ітеративного процесу обчислює вагу кожної вершини, враховуючи вагу сусідніх вершин. Вага вершини відображає важливість відповідного речення в контексті всього тексту [5]. Після завершення ітеративного процесу речення ранжуються за їх вагою. Речення з найбільшою вагою вважаються найбільш важливими і включаються до реферату. Повністю алгоритм виглядає наступним чином:

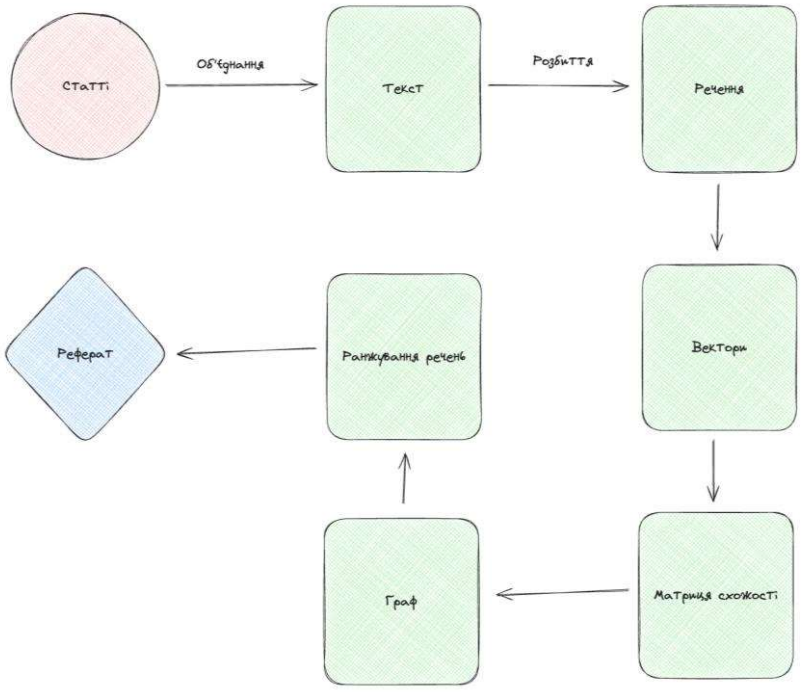


*Рисунок 2.* Схема роботи алгоритму PageRank

Для вираховування ваги кожного речення в контексті всього тексту, тобто вершини графу використовується наступна формула:

$$S(V_i) = (1-d) + d \times \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_{k \in Out(V_j)}} w_{jk}} \times S(V_j), \qquad (1)$$

де $S(V_i)$ – вага вершини; $V_i$ – важливість речення $i$; $d$ – коефіцієнт демпфірування (зазвичай встановлюється на 0,85); $In(V_i)$ – множина вершин, що мають ребро, спрямоване до вершини $V_i$; $Out(V_j)$ – множина вершин, до яких спрямовані ребра від вершини $V_j$; $w_{ji}$ – вага ребра між вершинами $V_i$ та $V_j$ (подібність між реченнями $j$ та $i$).

 ISSN 1560-8956

LSA (Latent Semantic Analysis) – це метод, який використовує сингулярний розклад матриці (SVD) для аналізу семантичного зв'язку між словами в тексті. LSA дозволяє зменшити розмірність текстових даних, виділяючи основні теми та концепції. Алгоритм LSA для реферування текстів починається з побудови матриці term-document, де рядки представляють слова, а стовпці - документи (речення або абзаци). Значення в матриці відображають частоту появи слів в документах. Далі матриця term-document розкладається на три матриці: $U$, $\Sigma$ та $V$. Матриця $\Sigma$ містить сингулярні числа, які відображають важливість латентних семантичних факторів. Наступний крок – зменшення розмірності, де вибирається певна кількість найбільших сингулярних чисел, які відображають основні теми тексту. Матриці $U$, $\Sigma$ та $V$ обрізаються до вибраної розмірності. Використовуючи зменшену матрицю term-document, обчислюється вага кожного речення, враховуючи його семантичний зв'язок з основними темами тексту. Нарешті, речення ранжуються за їх вагою. Речення з найбільшою вагою вважаються найбільш важливими і включаються до реферату [6]. LSA дозволяє зберегти основні семантичні зв'язки в тексті, створюючи більш змістовні та інформативні реферати. Формула сингулярного розкладу матриці використовується для розкладання матриці term-document на три матриці, які розкривають латентні семантичні зв'язки в тексті. Формула виглядає наступним чином:

$$A = U\Sigma V^T, \tag{2}$$

де $A$ – матриця term-document; $U$ – ортогональна матриця, що містить ліві сингулярні вектори; $\Sigma$ – діагональна матриця, що містить сингулярні числа; $V^T$ – транспонована ортогональна матриця, що містить праві сингулярні вектори.

Алгоритм у графічному зображенні:

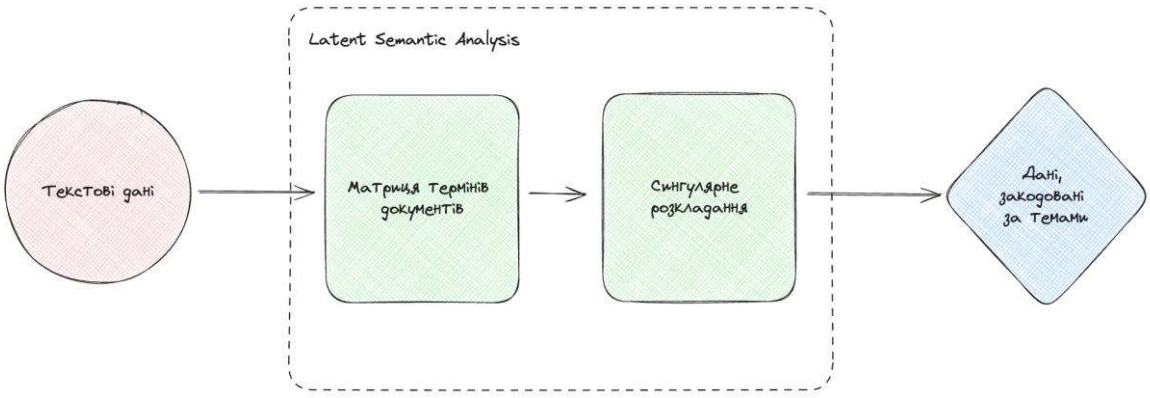


*Рисунок 3.* Схема роботи алгоритму Latent Semantic Analysis

## Типи реферування

У світі автоматичного реферування текстів існує три основні підходи, кожен з яких надає унікальний спосіб стиснення інформації. Екстрактивне реферування працює як майстерний редактор, вибираючи найважливіші речення з оригінального тексту та компонуючи їх у стислий виклад [7]. Цей метод відрізняється простотою та швидкістю, але іноді може створювати незв'язні та не завжди логічні реферати, особливо при обробці багатогранних текстів. Анотаційне реферування нагадує письменника, який переказує сутність тексту своїми словами, використовуючи методи NLP для визначення ключових концепцій та їх взаємозв'язків [8]. Результатом є більш зв'язний та логічний реферат, проте його створення вимагає більше обчислювальних ресурсів. Нарешті, абревіативне реферування працює як скульптор, який обережно видаляє зайві фрагменти з тексту, зберігаючи його основну форму та зміст [9]. Алгоритми абревіативного реферування аналізують текст, виділяють найбільш значущі фрагменти та видаляють несуттєві деталі, забезпечуючи стислий та змістовний виклад. Вибір оптимального підходу залежить від конкретного завдання та бажаного рівня деталізації.

## Виклики та перспективи

Сфера автоматизованого реферування текстів демонструє значний прогрес, проте шлях до створення ідеальних систем, здатних повністю замінити людину в цьому завданні, все ще залишається довгим та складним. Сучасні алгоритми успішно справляються з багатьма аспектами реферування, але певні виклики все ще потребують уваги дослідників та розробників.

Одним з головних каменів спотикання є невичерпна складність людської мови. Її багатозначність, контекстуальна залежність та гнучкість ускладнюють створення алгоритмів, здатних повноцінно осягнути та адекватно передати всі нюанси змісту оригінального тексту. Наприклад, значення одного і того ж слова може суттєво змінюватися залежно від контексту, що ставить перед алгоритмами завдання правильної інтерпретації семантики. Більше того, мова постійно розвивається, з'являються нові слова та вирази, змінюються норми вживання мовних конструкцій, що вимагає постійного адаптування алгоритмів до нових реалій.

Ще однією проблемою є оцінка якості згенерованих рефератів. Відсутність однозначних критеріїв якості та суб'єктивність сприйняття ускладнюють створення універсальних метрик оцінювання. Те, що один читач сприйме як якісний реферат, інший може вважати неповним або неточним. Різноманітність потреб користувачів та контекстів використання рефератів створюють додаткові виклики для розробки систем оцінювання, які були б одночасно об'єктивними та відповідними до заданого контексту.

Незважаючи на ці складності, сфера автоматизованого реферування текстів має величезний потенціал для подальшого розвитку. Вдосконалення алгоритмів обробки

 ISSN 1560-8956

природної мови та машинного навчання, розробка нових архітектур нейронних мереж, здатних глибше аналізувати тексти та враховувати контекст, – все це відкриває нові перспективи для створення більш ефективних та точних систем реферування. Важливим напрямком є розробка алгоритмів, здатних генерувати не просто екстрактивні реферати, що складаються з фрагментів оригінального тексту, а анотаційні реферати, які описують зміст тексту більш абстрактно та змістовно[10].

Розвиток методів оцінювання якості рефератів та створення об'єктивних метрик, адаптованих до різних типів текстів та потреб користувачів, є необхідним кроком для побудови дійсно ефективних систем реферування. З розвитком технологій штучного інтелекту та зростанням обсягів доступної інформації автоматизоване реферування текстів стає все більш затребуваним інструментом для ефективної роботи з текстовими даними. Воно допомагає людям швидко та ефективно опрацьовувати величезні обсяги інформації, виділяти головне та приймати обгрунтовані рішення.

## Висновки

В епоху інформаційного буму, де потік текстових даних зростає з кожним днем, автоматизоване реферування текстів постає не просто технологічним трендом, а необхідністю. Це інструмент, що допомагає опанувати лавину інформації, не втрачаючи головного. Хоча створення ідеальної системи реферування, яка б повністю відповідала на всі запити та враховувала всі нюанси людської мови, залишається складним завданням, досягнуті успіхи в цій галузі заслуговують на увагу.

Сучасні алгоритми реферування, поєднуючи методи обробки природної мови та машинного навчання, демонструють вражаючі результати. Вони здатні аналізувати великі обсяги тексту, виділяти ключову інформацію, враховувати семантичні зв'язки між словами та реченнями, створюючи стислі та змістовні реферати. Проте подальший розвиток цієї галузі вимагає постійної роботи над удосконаленням алгоритмів, розширенням їхніх можливостей та адаптацією до нових реалій мовного середовища. Важливим напрямком є розробка більш гнучких та чутливих до контексту алгоритмів, здатних генерувати не просто екстрактивні реферати, а анотаційні реферати, що відображають зміст тексту на більш високому рівні абстракції.

Розвиток методів оцінювання якості рефератів, створення об'єктивних метрик, адаптованих до різних типів текстів та запитів користувачів, є важливим завданням для побудови ефективних систем реферування. З кожним роком кількість інформації зростає, і вміння швидко та ефективно виділяти головне стає все більш затребуваним навичкою. Автоматизоване реферування текстів – це не просто технологічний тренд, а необхідність, яка допоможе опанувати інформаційний простір та ефективно використовувати знання, закладені в текстах.

## СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Natural Language Processing (NLP): What Is It & How Does it Work?: website URL: https://monkeylearn.com/natural-language-processing/ (application date: 08.07.2024).

2. GPT-4 Can’t Reason: website URL: https://medium.com/@konstantine_45825/gpt-4-cant-reason-2eab795e2523 (application date: 29.07.2024).

3. Natural Language Processing: website URL: https://www.deeplearning.ai/resources/natural-language-processing/ (application date: 10.07.2024).

4. What is Machine Learning (ML)?: website URL: https://cloud.google.com/learn/what-is-machine-learning (application date: 10.07.2024).

5. Understanding Page Rank: website URL: https://medium.com/@sarthakanand/page-rank-b7072c61dd85 (application date: 18.07.2024).

6. What is Latent Semantic Analysis?: website URL: https://ahrefs.com/seo/glossary/latent-semantic-analysis-lsa (application date: 18.07.2024).

7. A knowledge-based approach to citation extraction: website URL: https://ieeexplore.ieee.org/document/1506448 (application date: 20.07.2024).

8. How to format a citation in the abstract: website URL: https://academia.stackexchange.com/questions/98484/how-to-format-a-citation-in-the-abstract (application date: 27.07.2024).

9. Definition of Abbreviations: website URL: https://apastyle.apa.org/style-grammar-guidelines/abbreviations/definition (application date: 31.07.2024).

10. Overcoming the Challenges of Unstructured Data in Multi-site, Electronic Medical Record-based Abstraction: website URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5024721/ (application date: 31.07.2024).

 ISSN 1560-8956