

DOI: <https://doi.org/10.15276/ict.03.2026.12>

УДК 004.85

Багатокритеріальний вибір альтернатив кластеризації текстових даних за Silhouette та bootstrap-стабільністю

Коваль Ігор Михайлович

ORCID: <https://orcid.org/0009-0001-2083-1747>; aspirant_kovalja@lntu.edu.ua

Луцький національний технічний університет, вул. Львівська, 75. Луцьк, 43018, Україна

АНОТАЦІЯ

У роботі досліджено проблему вибору альтернатив кластеризації текстових даних за умов, коли оцінювання результату лише за одним внутрішнім критерієм не дає повного уявлення про властивості сформованого розбиття. Значення Silhouette характеризує компактність і відокремленість кластерів, однак не відображає відтворюваність отриманого рішення за зміни складу вибірки. У зв'язку з цим метою дослідження є розроблення та експериментальна перевірка процедури багатокритеріального вибору альтернатив кластеризації текстових даних на основі спільного врахування Silhouette та bootstrap-стабільності. Стабільність визначається за допомогою повторних bootstrap-вибірок і порівняння сформованих розбиттів із базовим результатом за Adjusted Rand Index. Для узгодження двох критеріїв використано Pareto-аналіз, що дає змогу виділити множину компромісних альтернатив без введення довільних вагових коефіцієнтів. Експериментальну перевірку виконано на тематичному корпусі, сформованому на основі відкритого набору новинних текстів British Broadcasting Corporation, із використанням алгоритмів KMeans та агломеративної кластеризації Ward. Встановлено, що Silhouette та bootstrap-стабільність формують різні ранжування кластерних альтернатив. Для KMeans найбільше значення Silhouette отримано за десяти кластерів, тоді як максимальна bootstrap-стабільність спостерігалася за п'яти кластерів. Для Ward конфігурації з максимальними значеннями Silhouette, стабільності та зовнішньої якості також не збігалися. Це показує, що bootstrap-стабільність не може розглядатися як універсальний критерій правильності кластеризації, але надає додаткову інформацію про відтворюваність кластерного рішення. Застосування Pareto-домінування за Silhouette та bootstrap-стабільністю дало змогу скоротити початкову множину з вісімнадцяти альтернатив до трьох недовінованих конфігурацій, тобто більш ніж на вісімдесят три відсотки. Еталонні тематичні мітки не використовувалися під час формування Pareto-множини та не входили до критеріїв багатокритеріального відбору; вони застосовувалися лише для зовнішнього оцінювання отриманих конфігурацій. Отримані результати підтверджують доцільність спільного аналізу внутрішньої геометричної якості та відтворюваності під час вибору альтернатив кластеризації текстових даних.

Ключові слова: машинне навчання; кластеризація текстових даних; обробка даних; якість; оцінювання; кластерний аналіз; навчання без учителя; інтелектуальний аналіз даних

Для цитування: Коваль І. М. “Багатокритеріальний вибір альтернатив кластеризації текстових даних за Silhouette та bootstrap-стабільністю”. *Інформатика. Культура. Техніка*. 2026; Том 3 № 1 (3): 144–154. DOI: <https://doi.org/10.15276/ict.03.2026.12>

Актуальність. Кластеризація текстових даних є одним із поширених методів навчання без учителя, що використовується для автоматичного групування документів за їх подібністю, виявлення тематичної структури та впорядкування великих масивів неструктурованої інформації. На відміну від задач класифікації, під час кластеризації заздалегідь відомі мітки класів, як правило, відсутні, тому оцінювання отриманого результату є окремою та складною задачею. Якість сформованого розбиття залежить від способу подання текстових документів, обраного алгоритму та його параметрів, зокрема кількості кластерів. У результаті для одного набору даних може бути сформовано декілька альтернатив кластеризації, між якими необхідно здійснити обґрунтований вибір. Для оцінювання таких альтернатив широко використовуються внутрішні метрики, які не потребують наявності еталонних міток. Однією з найбільш відомих є Silhouette, що характеризує співвідношення компактності об'єктів усередині кластерів та їх відокремленості від інших кластерів. Водночас вибір конфігурації лише за максимальним значенням однієї внутрішньої метрики враховує передусім геометричні властивості сформованого розбиття. Такий підхід не показує, наскільки отримана кластерна структура зберігатиметься за зміни складу вхідної вибірки та чи буде результат відтворюваним за повторного формування даних.

Окремою характеристикою кластерного рішення є його стабільність. У задачах навчання без учителя стабільність дає змогу оцінити, наскільки подібні розбиття формуються після контрольованого збурення вихідних даних. Одним зі способів такого оцінювання є

bootstrap-перевибірка, за якої багаторазово формуються вибірки з початкового корпусу, виконується повторна кластеризація та визначається ступінь узгодженості отриманих результатів із базовим розбиттям. При цьому висока стабільність сама по собі не означає, що сформовані кластери є правильними або найкращими за внутрішньою геометричною структурою. Вона характеризує іншу властивість результату — його відтворюваність за зміни складу вибірки.

У зв'язку з цим актуальним є завдання спільного врахування внутрішньої якості та стабільності під час вибору альтернатив кластеризації текстових даних. Використання цих характеристик як окремих критеріїв дає змогу аналізувати не лише компактність і розділеність сформованих кластерів, а й стійкість отриманого рішення до bootstrap-збурень вибірки. Водночас необхідно уникнути введення суб'єктивних вагових коефіцієнтів між критеріями та забезпечити можливість відбору компромісних конфігурацій. Для цього доцільним є застосування багатокритеріального підходу, за якого альтернативи оцінюються одночасно за внутрішньою якістю та відтворюваністю, а подальший вибір здійснюється серед рішень, які не поступаються іншим альтернативам за сукупністю визначених критеріїв.

Мета дослідження полягає у розробленні та експериментальній перевірці процедури багатокритеріального вибору альтернатив кластеризації текстових даних на основі спільного врахування внутрішньої якості кластерного розбиття та його відтворюваності. Для цього значення Silhouette використовується як критерій внутрішньої геометричної якості, bootstrap-стабільність – як характеристика відтворюваності кластерного рішення, а формування множини компромісних альтернатив виконується за допомогою Pareto-аналізу без використання еталонних тематичних міток на етапі вибору.

Аналіз останніх досліджень і публікацій. Оцінювання результатів кластеризації є окремою проблемою навчання без учителя, оскільки під час формування кластерів еталонна структура даних у більшості практичних випадків заздалегідь невідома. Для порівняння отриманих розбиттів широко використовуються внутрішні критерії, які дають змогу оцінювати властивості кластерної структури без залучення зовнішніх міток. Одним із найбільш відомих показників цього типу є Silhouette, значення якого формується на основі співвідношення внутрішньокластерної компактності та відокремленості об'єкта від найближчого альтернативного кластера [1].

Silhouette залишається актуальним критерієм і в сучасних дослідженнях кластерного аналізу. Зокрема, запропоновано його ймовірніше розширення для задач глибокої кластеризації, у якому також зберігається орієнтація на формування компактних і добре відокремлених кластерів [2]. Таким чином, Silhouette безпосередньо характеризує геометричні властивості сформованого розбиття, що робить метрику зручною для порівняння декількох кластерних конфігурацій. Водночас отримане значення відображає якість конкретного розбиття для конкретної вибірки та саме по собі не характеризує те, наскільки це рішення буде відтворюватися після зміни складу вхідних даних.

Для текстових даних питання оцінювання додатково ускладнюється особливостями їх числового представлення. Результат кластеризації може змінюватися залежно від способу векторизації документів, зменшення розмірності, вибраного алгоритму та його параметрів [3]. Сучасні експериментальні дослідження текстової кластеризації показують, що різні способи векторного представлення та алгоритми формують різні за якістю кластерні структури, а для їх порівняння доцільно використовувати декілька показників [4]. Це підтверджує, що оцінювання текстової кластеризації не повинно зводитися лише до механічного пошуку максимального значення одного внутрішнього критерію.

Окремий напрям досліджень пов'язаний зі стабільністю кластерних рішень. У межах такого підходу оцінюється не лише структура одного отриманого розбиття, а й здатність алгоритму формувати подібні результати після контрольованого збурення початкових даних. Стабільнісні методи використовуються як додатковий інструмент валідації кластерного рішення та можуть застосовуватися під час визначення параметрів кластеризації [5]. Такий

підхід дає змогу розглядати відтворюваність як окрему характеристику результату, що не є тотожною внутрішньокластерній компактності або міжкластерній відокремленості.

Одним зі способів оцінювання стабільності є використання bootstrap-перевибірок. Для цього з початкового набору багаторазово формуються нові вибірки, до яких повторно застосовується алгоритм кластеризації, після чого отримані розбиття порівнюються між собою або з базовим результатом. У дослідженні [6] запропоновано bootstrap-схему оцінювання нестабільності кластеризації та показано можливість її використання для вибору кількості кластерів. Перевагою такого підходу є те, що стабільність визначається без залучення еталонних класів і відображає чутливість кластерного рішення до зміни складу вибірки.

Методи оцінювання стабільності на основі повторного формування даних охоплюють bootstrap-перевибірки, формування підвибірок та інші способи контрольованого збурення вибірки. Відтворюваність при цьому може оцінюватися на рівні окремих об'єктів, кластерів або кластерного розбиття загалом [5]. У такому випадку внутрішній критерій і стабільність виконують різні функції: перший використовується для оцінювання структури кластерів, а другий дозволяє перевірити, наскільки отримане рішення зберігається за розумних збурень даних. Водночас висока відтворюваність не є прямим підтвердженням правильності кластеризації, оскільки стабільним може бути і рішення, яке не є найкращим за іншими характеристиками якості.

Необхідність одночасного врахування декількох властивостей кластерного рішення пов'язана з розвитком багатокритеріальних підходів. У сучасних роботах запропоновано оцінювати кластеризацію одночасно за декількома внутрішньокластерними та міжкластерними показниками, використовуючи невідомі сортування для формування множини компромісних рішень [7]. У такому випадку покращення одного критерію може супроводжуватися погіршенням іншого, тому результатом є множина невідомі альтернатив, а не обов'язково одна конфігурація. Це дає змогу уникнути безпосереднього об'єднання декількох показників в одну функцію з наперед заданими ваговими коефіцієнтами.

Багатокритеріальний підхід використовується безпосередньо і в задачах кластеризації текстових даних. У дослідженні [8] задачу кластеризації текстових результатів пошуку сформульовано як багатоцільову оптимізацію з одночасним урахуванням компактності та відокремленості кластерів. Експериментальна перевірка на текстових наборах даних підтвердила доцільність використання декількох критеріїв у випадках, коли оцінювання за однією цільовою характеристикою є недостатнім.

Таким чином, проведений аналіз показує наявність трьох пов'язаних напрямів оцінювання кластеризації: використання внутрішніх критеріїв якості, оцінювання стабільності за повторного формування вибірок та багатокритеріальний відбір рішень. Водночас стабільність у розглянутих роботах переважно використовується для визначення кількості кластерів або подальшої перевірки вже вибраного розбиття, тоді як багатокритеріальні підходи здебільшого поєднують декілька геометричних або оптимізаційних характеристик. Порівняно менше уваги приділяється спільному використанню внутрішньої геометричної якості та bootstrap-відтворюваності як двох нееквівалентних критеріїв саме для відбору вже сформованих альтернатив кластеризації текстових даних. Це обґрунтовує доцільність підходу, у якому Silhouette використовується для оцінювання внутрішньої якості, bootstrap-стабільність характеризує відтворюваність кластерного рішення, а Pareto-аналіз застосовується для виділення множини невідомі альтернатив без попереднього встановлення ваг між цими критеріями та без використання еталонних тематичних міток на етапі відбору.

Матеріали та умови експерименту. Експериментальне дослідження проведено на текстовому наборі даних, сформованому на основі відкритого BBC dataset, розміщеного на платформі Kaggle [9]. Вихідний набір містив 2410 записів із текстовим полем desc та набором тематичних тегів tags. На першому етапі було перевірено наявність точних

дублікатів за текстовим полем. У результаті виявлено та вилучено 282 повторні записи, після чого обсяг корпусу становив 2128 документів.

Вихідний набір мав багатоміткову структуру, а поле tags одночасно містило тематичні категорії, географічні назви, організації, події та інші сутності. Для формування контрольованого тематичного корпусу було визначено п'ять основних категорій: business, entertainment, politics, sports та technology. Документ включався до кінцевої вибірки лише за умови наявності рівно однієї з визначених тематичних категорій. Записи, які не містили жодної з п'яти категорій або одночасно належали до двох і більше з них, вилучалися. Після застосування цього правила з 2128 документів було відібрано 1702, що становило 79,98 % очищеного корпусу. Розподіл документів за тематичними категоріями становив: business – 375, entertainment – 338, politics – 322, sports – 482 та technology – 185.

Тематичні теги використовувалися виключно на етапі формування однозначної контрольної вибірки. Вони не включалися до текстового представлення, не використовувалися як ознаки під час кластеризації та не передавалися алгоритмам під час bootstrap-оцінювання. Таким чином, KMeans та агломеративна кластеризація працювали в режимі навчання без учителя. Сформовані тематичні мітки було збережено лише для зовнішньої перевірки результатів після завершення основної процедури багатокритеріального відбору.

Попередню обробку текстових документів виконано за єдиною відтворюваною процедурою для всього корпусу. Вона включала приведення тексту до нижнього регістру, вилучення HTML-елементів і вебпосилань, очищення від пунктуації, цифр та інших нетекстових символів, токенизацію, вилучення англійських стоп-слів і нормалізацію пробілів. Заперечення no, not та nor зберігалися, оскільки їх вилучення може змінювати зміст текстових конструкцій. Після попередньої обробки документи перетворювалися на числове представлення за допомогою TF-IDF [3]. Для всіх експериментальних конфігурацій використовувалися уніграми, $\min_df = 2$, $\max_df = 0,95$, без штучного обмеження максимальної кількості ознак та з L2-нормалізацією.

Висока розмірність і розрідженість TF-IDF-подання зумовили необхідність попереднього діагностичного порівняння варіантів представлення текстових даних. Перевірялися зміни параметрів TF-IDF, обмеження словника, використання біграм, а також латентні представлення зі 100, 200 і 300 компонентами. За результатами діагностичного етапу для основного експерименту було зафіксовано схему $TF-IDF \rightarrow LSA \rightarrow L2$, у якій метод Latent Semantic Analysis [10] реалізовано за допомогою TruncatedSVD зі 100 компонентами та $\text{random_state} = 42$. Вибір 100 компонент забезпечував суттєве скорочення початкового ознакового простору та давав змогу зафіксувати компактне латентне представлення без подальшого налаштування його розмірності за результатами основного експерименту. Для обраного представлення частка поясненої дисперсії становила 0,250812. Після зменшення розмірності отримані вектори повторно нормалізувалися за L2-нормою, а вибране представлення залишалося незмінним для всіх основних експериментів.

Для формування альтернатив використано два алгоритми, що реалізують різні принципи побудови кластерів: KMeans та агломеративну ієрархічну кластеризацію Ward. Для KMeans застосовано ініціалізацію k-means++, $n_init = 50$, $\max_iter = 300$, поріг збіжності $\text{tol} = 10^{-4}$, $\text{random_state} = 42$ та алгоритм Lloyd. Для агломеративної кластеризації використано метод Ward з евклідовою відстанню та побудовою повного дерева кластеризації ($\text{compute_full_tree} = \text{True}$). Вибір Ward здійснено після попереднього порівняння декількох схем агломеративної кластеризації. Варіант average із косинусною відстанню формувал один домінуючий кластер і декілька малочисельних груп, тоді як complete характеризувався слабшими показниками внутрішньої якості. Ward забезпечував більш збалансовану структуру кластерів, тому був зафіксований для основного експерименту.

Кількість кластерів для обох алгоритмів змінювалася від ($k=2$) до ($k=10$) з кроком 1. Отже, для кожного алгоритму було сформовано дев'ять конфігурацій, а загальна множина складалася з 18 альтернатив. У межах дослідження альтернативою вважається конкретна

комбінація алгоритму кластеризації та значення (k). Усі альтернативи оцінювалися на одному і тому самому підготовленому представленні даних, що виключало вплив зміни попередньої обробки або векторизації під час їх порівняння.

Внутрішню геометричну якість кожної альтернативи оцінювали за Silhouette із використанням косинусної відстані. Відтворюваність кластерних рішень визначали окремо за допомогою bootstrap-процедури з 500 повторними вибірками. Для зовнішньої перевірки використовували Adjusted Rand Index, Normalized Mutual Information та V-measure. При цьому зовнішні метрики не брали участі у формуванні множини альтернатив та не використовувалися під час багатокритеріального відбору. Основними критеріями подальшого Pareto-аналізу були лише Silhouette та bootstrap-стабільність.

Методика багатокритеріального оцінювання альтернатив кластеризації
Запропонована методика передбачає оцінювання кожної сформованої альтернативи за двома критеріями, що характеризують різні властивості кластерного рішення. Під альтернативою (A) розуміється конкретна комбінація алгоритму кластеризації та кількості кластерів (k). Першим критерієм є внутрішня геометрична якість розбиття, яка визначається за Silhouette, другим – bootstrap-стабільність, що характеризує відтворюваність кластерного рішення за зміни складу вибірки. Обидва критерії підлягають максимізації, однак не об'єднуються в одну зважену функцію, а подальший відбір виконується за принципом Pareto-домінування.

Для кожної альтернативи спочатку формувалося базове кластерне розбиття на повному підготовленому корпусі.

Внутрішню якість отриманого розбиття оцінювали за Silhouette [1]. Для окремого документа значення метрики визначається як

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad (1)$$

де $a(i)$ – середня відстань від документа (i) до інших документів його кластера; $b(i)$ – найменша середня відстань від документа (i) до документів іншого кластера. Для оцінювання альтернативи використовувалося середнє значення $s(i)$ для всіх документів корпусу. Більше значення Silhouette відповідає кращому співвідношенню внутрішньокластерної компактності та міжкластерної відокремленості.

Для всіх 18 альтернатив Silhouette обчислювався з використанням косинусної відстані. Це забезпечило застосування єдиного внутрішнього критерію під час порівняння результатів двох алгоритмів. Водночас необхідно розрізняти метрику, яка використовується безпосередньо алгоритмом, і метрику подальшого оцінювання отриманого розбиття. Зокрема, агломеративна кластеризація Ward виконувалася з евклідовою метрикою, оскільки вона є необхідною для використаного методу зв'язування, тоді як якість уже сформованого розбиття, як і для KMeans, оцінювалася за Silhouette з косинусною відстанню. Вхідні LSA-вектори перед кластеризацією були нормалізовані за L2-нормою, а однакова схема оцінювання використовувалася для всіх альтернатив.

Другим критерієм було визначено відтворюваність кластерного рішення за bootstrap-збуреннями вибірки. Для кожної альтернативи (A) спочатку зберігалося базове розбиття (y^A), отримане на повному корпусі з ($N=1702$) документів. Після цього виконувалося ($B=500$) bootstrap-ітерацій. На кожній ітерації (b) формувалася вибірка індексів (I_b) розміром (N) шляхом випадкового відбору документів із поверненням. Отже, окремий документ міг бути представлений у bootstrap-вибірці декілька разів, тоді як інші документи могли до неї не потрапити.

Для всіх значень (k), а також для KMeans і Ward використовувалася однакова послідовність із 500 bootstrap-вбірок. Така організація експерименту забезпечила порівнюваність альтернатив на однакових збуреннях даних та можливість їх подальшого попарного аналізу. Генератор bootstrap-вбірок було зафіксовано значенням 20260811.

Ознакове представлення всередині bootstrap-ітерацій повторно не формувалося. TF-IDF та LSA були навчені один раз на підготовленому корпусі, після чого bootstrap-перевибірка

здійснювалася над рядками зафіксованої LSA-матриці. Таким чином, у межах дослідження оцінювалася стабільність саме етапу кластеризації за фіксованого ознакового представлення, а не стабільність усього конвеєра попередньої обробки та векторизації. Це дозволило відокремити зміни кластерного рішення, спричинені зміною складу вибірки, від варіативності, яка могла б додатково виникати внаслідок повторного навчання TF-IDF або LSA.

На кожній bootstrap-ітерації з базового розбиття (y^A) вибиралися мітки документів відповідно до індексів (I_b), а алгоритм кластеризації повторно запускався на відповідній bootstrap-вибірці з незмінним значенням (k) та іншими зафіксованими параметрами. Подібність базового та повторно сформованого розбиттів оцінювалася за Adjusted Rand Index [11].

Для уникнення неоднозначності з подальшою зовнішньою перевіркою bootstrap-значення ARI позначимо як $ARI_b^{boot}(A)$:

$$ARI_b^{boot}(A) = ARI(y^A[I_b], \widehat{y}_b^A), \quad (2)$$

де ($y^A[I_b]$) – базові кластерні мітки документів, включених до (b)-ї bootstrap-вибірки; ($\widehat{y}_b^A$) – мітки, отримані після повторної кластеризації цієї вибірки. Використання ARI є доцільним для такого порівняння, оскільки показник не залежить від числового позначення кластерів та коригує оцінку узгодженості з урахуванням випадкового збігу.

Після виконання всіх bootstrap-ітерацій для кожної альтернативи отримувалося 500 значень $ARI_b^{boot}(A)$.

Підсумкову bootstrap-стабільність визначали як їх середнє значення:

$$S_{boot}(A) = \frac{1}{B} \sum_{b=1}^B ARI_b^{boot}(A), \quad B = 500. \quad (3)$$

Більше значення $S_{boot}(A)$ відповідає вищій відтворюваності кластерного рішення за використаного типу bootstrap-збурення. Додатково для кожної альтернативи фіксувалися стандартне відхилення, медіана, 2,5-й і 97,5-й процентилі розподілу bootstrap-значень та Monte Carlo standard error. Процентильні межі використовувалися для опису варіативності отриманих значень, а не трактувалися як самостійний критерій вибору альтернативи.

Після обчислення двох основних показників кожну альтернативу (A) представляли парою ($S_{sil}(A), S_{boot}(A)$), де ($S_{sil}(A)$) – середнє значення Silhouette. Для обох критеріїв кращим вважається більше значення. Внутрішня якість та стабільність навмисно не об'єднувалися шляхом додавання або введення вагових коефіцієнтів, оскільки це вимагало б попереднього встановлення відносної важливості двох нееквівалентних характеристик.

Для формування множини компромісних рішень використано принцип Pareto-домінування [12].

Альтернатива (A_i) вважалася такою, що домінує альтернативу (A_j), якщо вона була не гіршою за обома критеріями та строго кращою щонайменше за одним із них:

$$A_i > A_j \Leftrightarrow [S_{sil}(A_i) \geq S_{sil}(A_j)] \wedge [S_{boot}(A_i) \geq S_{boot}(A_j)] \wedge [S_{sil}(A_i) > S_{sil}(A_j) \vee S_{boot}(A_i) > S_{boot}(A_j)]. \quad (4)$$

Альтернативи, для яких у загальній множині не існувало іншої конфігурації, що домінувала б їх відповідно до умови (4), включалися до глобальної Pareto-множини. Таким чином, процедура не передбачає примусового визначення одного «найкращого» рішення, а формує скорочену множину недомінованих альтернатив, кожна з яких представляє певний компроміс між внутрішньою геометричною якістю та bootstrap-відтворюваністю.

Принциповою умовою багатокритеріального відбору була відсутність еталонних тематичних міток у процедурі оцінювання. Вони не використовувалися під час розрахунку

Silhouette, визначення $S_{boot}(A)$ або формування Pareto-множини. Лише після завершення відбору результати порівнювалися з відомою тематичною структурою корпусу за зовнішніми метриками Adjusted Rand Index, Normalized Mutual Information та V-measure. Для відокремлення цієї оцінки від ARI, використаного всередині bootstrap-процедури, зовнішній показник надалі позначається як (ARI^{ext}) . Такий порядок дозволяє виконати перевірку отриманих альтернатив, не використовуючи зовнішню інформацію безпосередньо під час їх багатокритеріального відбору.

Експериментальні результати та зовнішня валідація. Проведене оцінювання 18 альтернатив показало, що максимізація Silhouette та bootstrap-стабільності приводить до різних варіантів кластерного розбиття. Найбільш виражена відмінність спостерігалася для KMeans. Максимальне значення Silhouette становило 0,169021 і було отримано при $(k=10)$, тоді як найбільша bootstrap-стабільність ($S_{boot}=0,968937$) відповідала конфігурації $(k=5)$. Для $(k=5)$ значення Silhouette становило 0,129452, тобто за внутрішньою геометричною якістю ця конфігурація займала лише шосте місце серед дев'яти варіантів KMeans. Отже, вибір лише за максимальним Silhouette надавав перевагу іншій альтернативі, ніж оцінювання відтворюваності кластерного рішення.

Для Ward відмінність між критеріями мала інший характер. Максимальне значення Silhouette 0,146764 отримано при $(k=9)$, тоді як максимальна bootstrap-стабільність ($S_{boot}=0,875052$) спостерігалася при $(k=3)$. Конфігурація $(k=5)$ займала друге місце за стабільністю зі значенням 0,807324. Таким чином, для обох алгоритмів максимуми двох основних критеріїв не збігалися, однак характер цього розходження залежав від алгоритму.

Рангове порівняння також показало виражену неузгодженість між Silhouette та bootstrap-стабільністю. Коефіцієнт рангової кореляції Спірмена для KMeans становив $(\rho=-0,8167)$, а для Ward $(\rho=-0,8333)$. Отриманий негативний зв'язок показує, що в дослідженому діапазоні збільшення внутрішньої геометричної якості часто супроводжувалося зменшенням відтворюваності. Це підтверджує недоцільність трактування Silhouette та (S_{boot}) як взаємозамінних критеріїв і обґрунтовує їх спільний аналіз.

Співвідношення Silhouette та bootstrap-стабільності для всіх досліджених альтернатив наведено на Рис. 1. Точки KMeans і Ward займають різні області двовимірного простору критеріїв. Для KMeans чітко простежується компроміс між конфігурацією $(k=5)$, яка характеризується максимальною відтворюваністю, та $(k=10)$, яка має максимальне значення Silhouette. Проміжне положення займає $(k=7)$. Саме ці три конфігурації утворили глобальну Pareto-множину.

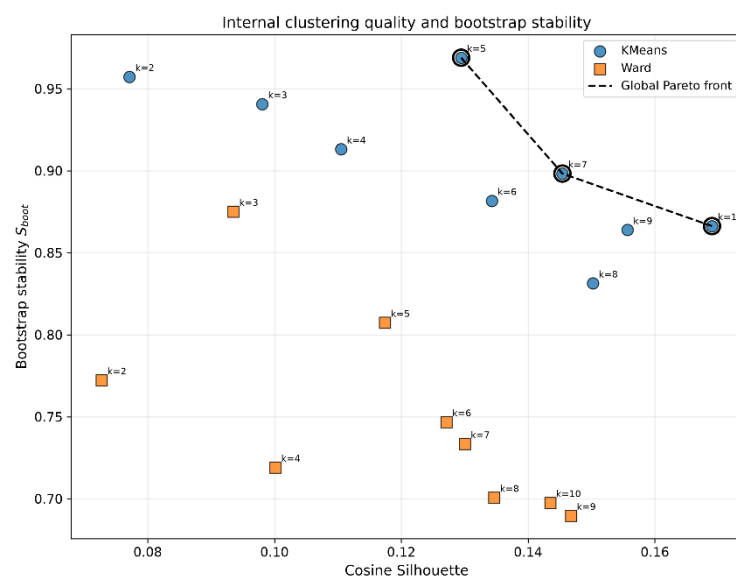


Рис. 1. Внутрішня якість кластеризації, bootstrap-стабільність та глобальна Pareto-множина

Джерело: розроблено авторами

Застосування Pareto-домінування до загальної множини з 18 альтернатив дало змогу вилучити 15 домінованих конфігурацій. У результаті кількість альтернатив скоротилася до трьох, тобто на 83,33 %. До глобальної Pareto-множини увійшли лише KMeans ($k=5$), ($k=7$) та ($k=10$). Їх значення Silhouette послідовно збільшувалися від 0,129452 до 0,169021, тоді як (S_{boot}) зменшувався від 0,968937 до 0,866269. Таким чином, Pareto-аналіз не визначає одну конфігурацію як безумовно найкращу, а залишає три альтернативи, що представляють різні співвідношення внутрішньої якості та відтворюваності.

Для додаткового аналізу відмінностей стабільності використано парне порівняння на однакових bootstrap-вибірках. Для KMeans значення (ARI_b^{boot}) для конфігурації ($k=5$) перевищувало відповідне значення для ($k=10$) у 93,8 % із 500 парних bootstrap-реплікацій. Для Ward значення (ARI_b^{boot}) для ($k=3$) перевищувало відповідне значення для ($k=5$) у 93,6 % реплікацій. Водночас значення (ARI_b^{boot}) для ($k=5$) перевищувало відповідне значення для ($k=9$), конфігурації з максимальним Silhouette, у 95,2 % випадків. Отримані результати показують, що відмінності середніх значень (S_{boot}) підтримуються результатами попарного порівняння на однакових bootstrap-вибірках.

Розподіли 500 bootstrap-значень для ключових альтернатив наведено на рис. 2. Для KMeans ($k=5$) основна частина значень зосереджена поблизу верхньої межі ARI, тоді як для ($k=7$) та ($k=10$) спостерігається поступове зниження рівня стабільності та більша варіативність. Для Ward розподіли характеризуються нижчими значеннями та більшою мінливістю, особливо для конфігурації ($k=9$). Таким чином, використання лише середнього (S_{boot}) доповнюється аналізом фактичного розподілу результатів повторної кластеризації.

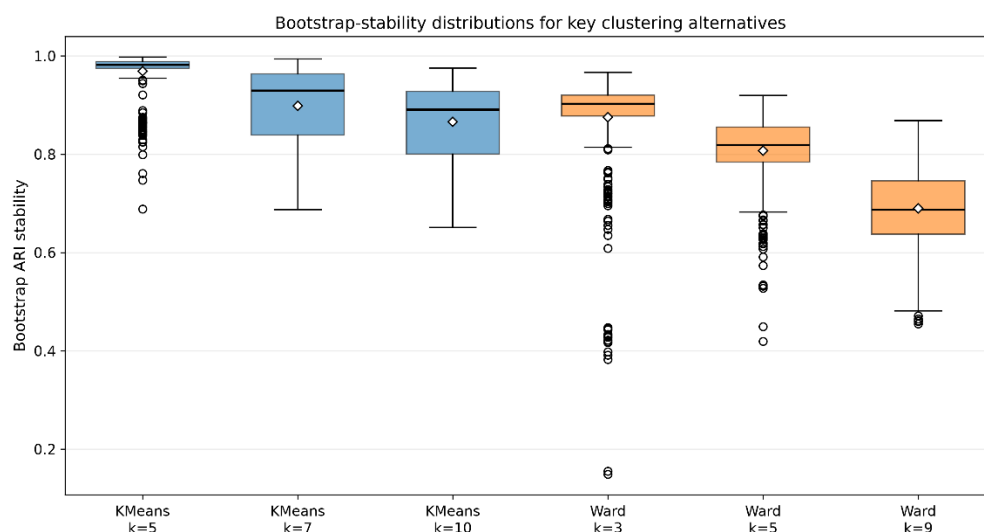


Рис. 2. Розподіли bootstrap-стабільності для ключових альтернатив кластеризації

Джерело: розроблено авторами

Після завершення багатокритеріального відбору проведено зовнішню валідацію з використанням еталонних тематичних міток. Значення (ARI^{ext}), NMI та V-measure не використовувалися під час формування Pareto-множини та застосовувалися лише для перевірки того, як отримані конфігурації співвідносяться з відомою тематичною структурою корпусу. Ключові результати наведено в Таблиці 1.

Для KMeans зовнішня валідація показала, що конфігурація ($k=5$), яка мала максимальну bootstrap-стабільність, одночасно забезпечила найвище значення $ARI^{ext}=0,929564$, NMI = 0,901863 та V-measure = 0,901863. При цьому альтернативі ($k=10$), обраній за максимальним Silhouette, відповідало значно нижче значення $ARI^{ext}=0,587885$. Важливим є те, що конфігурація ($k=5$) не була втрачена під час багатокритеріального відбору та увійшла до глобальної Pareto-множини без використання зовнішніх міток.

Таблиця 1. Ключові результати внутрішнього оцінювання, bootstrap-стабільності та зовнішньої валідації

Алгоритм	k	Silhouette	S_{boot}	ARI_{ext}	NMI	V-measure
KMeans	5	0,129	0,969	0,930	0,902	0,902
KMeans	7	0,145	0,898	0,740	0,820	0,820
KMeans	10	0,169	0,866	0,588	0,754	0,754
Ward	3	0,094	0,875	0,529	0,668	0,668
Ward	5	0,117	0,807	0,833	0,804	0,804
Ward	9	0,147	0,690	0,581	0,696	0,696

Джерело: розроблено авторами

Результати Ward демонструють іншу ситуацію. Максимальна стабільність була отримана при (k=3), однак найбільшу відповідність еталонній тематичній структурі забезпечила конфігурація (k=5), для якої ($ARI_{ext}=0,833038$), NMI = 0,803631 та V-measure = 0,803631. Конфігурація (k=9), яка мала максимальний Silhouette, також не була найкращою за зовнішньою якістю. Отже, ні максимальне значення Silhouette, ні максимальна bootstrap-стабільність для Ward не відтворили зовнішньо найкраще значення (k).

Отримані результати не дають підстав розглядати bootstrap-стабільність як заміну Silhouette або як універсальний засіб визначення правильної кількості кластерів. Для KMeans стабільність надала корисну додаткову інформацію та вказала на конфігурацію, яка надалі виявилася найкращою за зовнішніми показниками. Для Ward такого прямого збігу не спостерігалось. Тому основним результатом є не перевага одного критерію над іншим, а можливість спільного використання внутрішньої геометричної якості та bootstrap-відтворюваності для скорочення множини альтернатив і виділення компромісних кластерних рішень без залучення еталонних міток на етапі відбору.

Висновки. У межах роботи нами розроблено та експериментально перевірено процедуру багатокритеріального вибору альтернатив кластеризації текстових даних, яка поєднує оцінювання внутрішньої геометричної якості за Silhouette та відтворюваності кластерного рішення за bootstrap-стабільністю. Для формування множини компромісних рішень використано Pareto-домінування без попереднього встановлення вагових коефіцієнтів між критеріями та без використання еталонних тематичних міток на етапі відбору.

Експериментальне дослідження охопило 18 альтернатив, сформованих алгоритмами KMeans та Ward для різної кількості кластерів. Встановлено, що максимальні значення Silhouette та bootstrap-стабільності не збігаються, а ранжування альтернатив за цими критеріями суттєво відрізняється. Застосування Pareto-аналізу дало змогу скоротити початкову множину з 18 до трьох недомінованих конфігурацій, тобто на 83,33 %. Для KMeans конфігурація (k=5), яка характеризувалася найбільшою bootstrap-стабільністю ($S_{boot}=0,968937$), одночасно мала найвище значення зовнішнього ($ARI_{ext}=0,929564$). Водночас для Ward максимальна стабільність при (k=3) не збіглася з найкращим результатом зовнішньої валідації, який було отримано при (k=5). Це показує, що bootstrap-стабільність не може використовуватися як універсальний критерій правильності кластеризації або як заміна Silhouette.

Отримані результати підтверджують доцільність спільного використання внутрішньої якості та відтворюваності як двох нееквівалентних характеристик кластерного рішення. Практичне значення запропонованого підходу полягає у можливості скорочення множини альтернатив до набору компромісних конфігурацій без залучення зовнішньої інформації під час основного етапу відбору. Обмеженням дослідження є використання одного англomовного тематичного корпусу та фіксованого представлення TF-IDF $\rightarrow$ LSA-100 $\rightarrow$ L2, тому узагальнення отриманих результатів на інші корпуси, мови та способи векторизації потребує

окремої перевірки. Bootstrap-процедура виконувалася над зафіксованою LSA-матрицею, тому оцінена стабільність характеризує передусім етап кластеризації, а не весь конвеєр формування ознак; подальші дослідження мають охопити інші корпуси та альтернативні способи представлення документів.

Конфлікт інтересів: Автор декларує, що не має конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті

Фінансування: Дослідження проводилося без фінансової підтримки

Доступність даних: Код, параметри експерименту, результати кластеризації, bootstrap-оцінювання, Pareto-аналізу та графічні матеріали, використані для одержання наведених результатів, доступні у відкритому репозиторії Zenodo. DOI: <https://doi.org/10.5281/zenodo.22048111>.

Використання засобів штучного інтелекту (ШІ): Автор підтверджує, що не застосовував інструментальні засоби ШІ для написання цієї роботи

СПИСОК ЛІТЕРАТУРИ

1. Rousseeuw, P. J. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. *Journal of Computational and Applied Mathematics*. 1987; 20: 53–65. DOI: [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
2. Vardakas, G., Papakostas, I. & Likas, A. “Deep clustering using the soft silhouette score: Towards compact and well-separated clusters”. *Machine Learning*. 2026; 115: 81. DOI: <https://doi.org/10.1007/s10994-026-07026-w>.
3. Cozzolino, I. & Ferraro, M. B. “Document clustering”. *WIREs Computational Statistics*. 2022; 14 (6): e1588. DOI: <https://doi.org/10.1002/wics.1588>.
4. Petukhova, A., Matos-Carvalho, J. P. & Fachada, N. “Text clustering with large language model embeddings”. *International Journal of Cognitive Computing in Engineering*. 2025; 6: 100–108, <https://www.scopus.com/pages/publications/85210765539>. DOI: <https://doi.org/10.1016/j.ijcce.2024.11.004>.
5. Liu, T., Yu, H. & Blair, R. H. “Stability estimation for unsupervised clustering: A review”. *WIREs Computational Statistics*. 2022; 14 (6): e1575. DOI: <https://doi.org/10.1002/wics.1575>.
6. Fang, Y. & Wang, J. “Selection of the number of clusters via the bootstrap method”. *Computational Statistics & Data Analysis*. 2012; 56 (3): 468–477, <https://www.scopus.com/pages/publications/80455177001>. DOI: <https://doi.org/10.1016/j.csda.2011.09.003>.
7. Azevedo, B. F., Rocha, A. M. A. C. & Pereira, A. I. “A multi-objective clustering approach based on different clustering measures combinations”. *Computational and Applied Mathematics*. 2025; 44: 59. DOI: <https://doi.org/10.1007/s40314-024-03004-x>.
8. Acharya, S., Saha, S., Moreno, J. G. & Dias, G. “Multi-objective search results clustering”. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics. 2014. p. 99–108. – Available from: <https://aclanthology.org/C14-1011>. – [Accessed: 15.05.2026].
9. Kirpekar, S. “BBC-News Dataset”. *Kaggle*. – Available from: <https://www.kaggle.com/datasets/sahilkirpekar/bbcnews-dataset>. – [Accessed: 13.05.2026].
10. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. “Indexing by latent semantic analysis”. *Journal of the American Society for Information Science*. 1990; 41 (6): 391–407. DOI: [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6%3C391::AID-ASI1%3E3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9).
11. Hubert, L. & Arabie, P. “Comparing partitions”. *Journal of Classification*. 1985; 2 (1): 193–218. DOI: <https://doi.org/10.1007/BF01908075>.
12. Handl, J. & Knowles, J. “An evolutionary approach to multiobjective clustering”. *IEEE Transactions on Evolutionary Computation*. 2007; 11 (1): 56–76. DOI: <https://doi.org/10.1109/TEVC.2006.877146>.

Отримано 06.07.2026

Отримано після перегляду 28.08.2026

Прийнято 04.09.2026

DOI: <https://doi.org/10.15276/ict.03.2026.12>

UDC 004.85

Multi-criteria selection of text clustering alternatives based on Silhouette and Bootstrap stability

Ihor M. Koval

ORCID: <https://orcid.org/0009-0001-2083-1747>; aspirant_kovalja@lntu.edu.ua
Lutsk National Technical University, 75 Lvivska Street, Lutsk, Volyn Region, 43018, Ukraine

ABSTRACT

The paper investigates the problem of selecting text clustering alternatives under conditions where evaluating the result using only a single internal criterion does not provide a complete understanding of the properties of the resulting partition. The Silhouette score characterizes cluster compactness and separation but does not reflect the reproducibility of the obtained solution when the sample composition changes. Therefore, the aim of the study is to develop and experimentally validate a procedure for multi-criteria selection of text clustering alternatives based on the joint consideration of Silhouette and bootstrap stability. Stability is assessed using repeated bootstrap samples and by comparing the resulting partitions with the baseline clustering solution using the Adjusted Rand Index. Pareto analysis is applied to reconcile the two criteria, making it possible to identify a set of compromise alternatives without introducing arbitrary weighting coefficients. The experimental evaluation was conducted on a thematic corpus derived from an open news text dataset of the British Broadcasting Corporation using KMeans and Ward agglomerative clustering. The results show that Silhouette and bootstrap stability produce different rankings of clustering alternatives. For KMeans, the highest Silhouette value was obtained with ten clusters, whereas the maximum bootstrap stability was observed with five clusters. For Ward clustering, the configurations corresponding to the maximum Silhouette, maximum stability, and highest external quality also differed. This indicates that bootstrap stability cannot be considered a universal criterion of clustering correctness, but it provides additional information about the reproducibility of a clustering solution. Applying Pareto dominance based on Silhouette and bootstrap stability reduced the initial set of eighteen alternatives to three non-dominated configurations, corresponding to a reduction of more than eighty-three percent. Reference thematic labels were not used to construct the Pareto set and were not included among the criteria of the multi-criteria selection; they were used only for external evaluation of the obtained configurations. The findings confirm the feasibility of jointly analyzing internal geometric quality and reproducibility when selecting text clustering alternatives.

Keywords: machine learning; text data clustering; data processing; quality; evaluation; cluster analysis; unsupervised learning; data mining

ІНФОРМАЦІЯ ПРО АВТОРА



Коваль Ігор Михайлович - аспірант, каф. Інженерії програмного забезпечення. Луцький національний технічний університет, вул. Львівська, 75. Луцьк, 43018, Україна

Галузь досліджень: машинне навчання та інтелектуальний аналіз даних

Ihor Koval - PhD student, Department of Software Engineering, Lutsk National Technical University, 75, Lvivska Street, Lutsk, 43018, Ukraine

ORCID: <https://orcid.org/0009-0001-2083-1747>; aspirant_kovalja@lntu.edu.ua

Research field: machine learning and data mining