DOI: https://doi.org/10.36910/6775-2524-0560-2025-58-18
УДК 004.912
Приходченко Сергій Дмитрович, к.т.н.
https://orcid.org/0000-0002-6562-0601
Приходченко Оксана Юріївна, к.е.н.
https://orcid.org/0000-0001-5080-737X
Шевцова Ольга Сергіївна, PhD
https://orcid.org/0000-0002-0148-5877
Національний технічний університет «Дніпровська політехніка», м. Дніпро, Україна

ПРОГРАМНЕ ПОРІВНЯННЯ АНОТАЦІЙ НАУКОВИХ СТАТЕЙ ЗА ДОПОМОГОЮ СТАТИСТИЧНИХ МЕТОДИК ОБРОБКИ ПРИРОДНІХ МОВ

Приходченко С.Д., Приходченко О.Ю., Шевцова О.С. Програмне порівняння анотацій наукових статей за допомогою статистичних методик обробки природніх мов. Стаття присвячена розробці та аналізу методик автоматизованого порівняння анотацій наукових статей за допомогою статистичних методів обробки природної мови (NLP). Авторами розглянуто сучасні підходи до текстової аналітики, включаючи алгоритми Cosine Similarity, Jaccard Similarity та TF-IDF, а також використання кластеризації та моделей машинного навчання, зокрема SciBERT та BioBERT. Запропонована дослідницька модель програмного забезпечення дозволяє автоматично визначати схожість між текстами, що сприяє ефективному аналізу наукових публікацій, зменшенню дублювання досліджень і підвищенню якості бібліографічного аналізу. Робота має значний потенціал для застосування в науковій комунікації, академічному плагіат-менеджменті та автоматизованих системах оцінки літератури. Подальші дослідження спрямовані на вдосконалення алгоритмів та інтеграцію запропонованої системи з науковими базами даних.
Ключові слова: обробка природної мови (NLP), автоматизований аналіз тексту, наукові анотації, порівняння текстів, автоматизований літературний огляд.

Prykhodchenko S., Prykhodchenko O., Shevtsova O. Software comparison of scientific article annotations using statistical natural language processing methods. This paper presents an advanced approach to the automated comparison of scientific article abstracts using statistical natural language processing (NLP) techniques. The authors analyze state-of-the-art methodologies, including Cosine Similarity, Jaccard Similarity, and TF-IDF, alongside clustering methods and machine learning models such as SciBERT and BioBERT. A research-based software model is proposed to enhance text similarity assessment, facilitating efficient scientific literature analysis, reducing research duplication, and improving bibliographic accuracy. The study highlights the practical application of NLP techniques in academic publishing, plagiarism detection, and automated literature review systems. The proposed system integrates various computational approaches to refine text analysis and classification, making it a valuable tool for researchers and journal editors. Future research directions include optimizing NLP algorithms, incorporating deep learning methods, and integrating the system with major scientific databases to enhance further its applicability and performance in academic and industrial contexts.
Keywords: natural language processing (NLP), automated text analysis, scientific annotations, text comparison, automated literature review.

Вступ. Порівняння наукових текстів – це не лише спосіб аналізу подібностей і відмінностей між дослідженнями, але й ключ до розуміння того, як розвивається наука. У сучасному світі, де обсяг наукових публікацій зростає експоненційно, стає критично важливим володіти інструментами й методами, які дозволяють ефективно зіставляти та аналізувати тексти. Особливо це стосується таких галузей, як Natural Language Processing (NLP), де текст є як об'єктом, так і інструментом дослідження.

Таке порівняння вимагає глибокого розуміння структури та логіки академічного письма. Ми маємо справу з різноманітними жанрами текстів – від статей у рецензованих журналах до тез конференцій, технічних звітів і дисертацій. Кожен із цих жанрів має свої особливості викладу, що впливають на інтерпретацію їхнього змісту. До того ж, тексти можуть бути написані різними мовами, використовувати відмінну термінологію або ж відображати різні підходи до розв'язання однієї проблеми. Вміння аналізувати ці нюанси – це навичка, яка стоїть на перетині лінгвістики, когнітивної науки та комп'ютерних технологій.

Важливим викликом для NLP у цьому контексті є автоматизація процесів, які раніше виконувалися вручну. Виділення ключових ідей з тексту; встановлення чи є між двома роботами схожість в ідеях, чи вони пропонують кардинально різні рішення; виявлення потенційної плагіатність або надмірну залежність від попередніх джерел? Це лише частина питань, які постають перед редакторами сучасних наукових журналів, а також – вченими, що намагаються осягнути нові горизонти знання. Основна мета цієї роботи – створення програмного застсунка, побудованого на

статистичних методиках порівняння текстових масивів, для порівняння інформаційних складових в анотаціях наукових статей.

Постановка задачі. Аналіз подібності допомагає уникнути включення однакових або дуже схожих джерел у список літератури, що підвищує об'єктивність та достовірність дослідження. Врахування подібності наповнення допомагає відфільтрувати дублікати та вибрати найбільш важливі та релевантні джерела для дослідження.

Автоматичне порівняння анотацій статей є актуальним і важливим завданням в галузі комп'ютерних наук, особливо в контексті сучасного наукового дослідження. Так, кількість опублікованих наукових статей зростає експоненційно, і дослідники стикаються з проблемою обробки та аналізу величезних обсягів інформації. Автоматичне порівняння анотацій допомагає швидко виокремити ідентичні або подібні дослідження серед цього потоку даних.

Автоматичне порівняння текстів статей є важливим завданням у галузі обробки природної мови (Natural Language Processing, NLP) та інформаційного пошуку.

Метою статті є аналіз існуючих підходів до програмного автоматизованого порівняння анотацій наукових статей та визначення напрямків подальших досліджень у галузі статистичних методик обробки природної мови для покращення якості та ефективності аналізу наукових текстів.

Аналіз попередніх досліджень і публікацій. Порівняння програмного забезпечення для аналізу тез наукових статей з використанням методів статистичної обробки природної мови (Natural language processing - NLP) виявляє різноманітні методології та результати. Останні дослідження підкреслюють ефективність різних моделей NLP та статистичних особливостей у класифікації та категоризації наукових рефератів, демонструючи досягнення в автоматизованому скринінгу літератури.

Попередньо навчені мовні моделі Scibert та BioBert були використані для вилучення значущих уявлень та класифікації рефератів. SciBERT продемонстрував покращену категоризацію порівняно з традиційними методами [1], тоді як BioBert показав прийнятну ефективність у медичних реферативних оглядів [2].

Підходи до машинного навчання: Випадкові лісові моделі в поєднанні з вбудовуваннями Word2Vec досягли оцінки F1 0,775 при класифікації медичних рефератів, що вказує на потенціал статистичних ознак у класифікації тексту [3].

Різні дослідження повідомляли про високі бали F1, причому одне досягло 0,92 для класифікації документів з використанням комбінації методів NLP [4]. Це підкреслює надійність цих методів при точній обробці наукових текстів. Також, дослідження показують, що такі моделі, як випадкові ліси та BioBERT, досягають високої точності класифікації, з оцінками F1 близько 0,775 та 0,854 відповідно [2,3]. Класифікація спирається на статистичні властивості тексту, такі як ключові слова та вбудовування, які мають вирішальне значення для розрізнення різних наукових областей [3,5]. Використання оцінки силуету для визначення оптимальної кластеризації в категоризації тексту ще більше підкреслює статистичну суворість, застосовану в цих аналізах [2]. Хоча досягнення в НЛП для абстрактної класифікації є перспективними [6,7], залишаються проблеми [8] в забезпеченні того, щоб моделі добре узагальнювались у різних наукових областях та підтримували точність у галузях, що швидко розвиваються. Порівняння програмного забезпечення для аналізу наукових рефератів з використанням статистичних методів в обробці природної мови (NLP) виявляє різноманітні підходи та інструменти, що покращують скринінг літератури. Кілька досліджень підкреслюють ефективність моделей машинного навчання та методів НЛП у класифікації рефератів, демонструючи їх потенціал для впорядкування процесу огляду. Такі інструменти, як RStudio, Python та IBM SPSS, часто використовуються для статистичного аналізу, кожен з яких пропонує унікальні функціональні можливості, придатні для різних аналітичних завдань [6]. Оцінка цих інструментів часто включає такі показники, як точність, точність та відкликання, які є важливими для оцінки ефективності алгоритмів NLP [8].

Хоча досягнення в НЛП та машинному навчанні дають значні переваги для абстрактної класифікації, залишаються проблеми у забезпеченні надійності та інтерпретації цих автоматизованих систем, особливо у складних областях досліджень.

Дослідження.

Існує декілька алгоритмів та методів для автоматичного порівняння текстів статей. Опишемо застосування основних:

Алгоритм Cosine Similarity визначає схожість між двома текстами на основі кутової відстані між їхніми векторами. Він використовує терм-документну матрицю, де кожен рядок представляє

текст статті, а кожний стовпець - слово (термін). Кутова відстань обчислюється між векторами, які представляють текстові документи. Використовується для порівняння текстів у багатьох галузях, включаючи пошук і ранжування наукових статей за схожістю.

Алгоритм Jaccard Similarity визначає схожість між двома текстами на основі кількості спільних слів у двох текстах, поділених на загальну кількість унікальних слів в обох текстах. Використовується для порівняння текстових документів та пошуку схожих статей.

TF-IDF (Term Frequency-Inverse Document Frequency) обчислює вагу кожного терміну (слова) в документі, враховуючи і частоту цього терміну в документі (Term Frequency) і зворотну частоту цього терміну в усіх документах колекції (Inverse Document Frequency). Схожість текстів обчислюється на основі цих ваг. TF-IDF використовується для порівняння текстів і ранжування статей в пошуку наукових інформаційних ресурсів.

TF-IDF - це статистичний показник, який використовується для визначення важливості кожного слова в документі в контексті всього корпусу документів. Після обчислення ваги кожного слова у кожному документі, можемо використовувати ці значення для порівняння текстів. Чим вище значення TF-IDF для конкретного слова у документі, тим важливіше це слово для розуміння змісту документа.

Кластеризація документів – це метод аналізу текстової інформації, що дозволяє групувати схожі документи разом у класи або кластери на підставі їхнього вмісту чи інших властивостей. Цей підхід знаходить широке застосування в областях, де обробка та розуміння великої кількості текстової інформації має ключове значення, таких як інформаційний пошук, аналіз соціальних мереж, або наукові дослідження. Для визначення схожості документів можна використати також методики кластерізації. Першим кроком є визначення мети кластеризації. Наприклад, це може бути групування новинних статей за темами, класифікація наукових статей за галузями чи сферами дослідження, або категоризація користувачських відгуків за темами. Вибір ознак – це значення, які ознаки документів будуть використані для оцінки їхньої схожості. Це може включати в себе використання словникового запасу, синтаксичних структур, ключових слів, або інших лінгвістичних та статистичних ознак. Перетворення текстових документів у векторну форму, що може бути використана алгоритмами кластеризації. Це може бути векторна модель простору слів (word embeddings), TF-IDF вектори, чи інші представлення.

Щоб створити візуалізацію з тексту, необхідна обробка тексту [9] або обробка природної мови для створення якісних або кількісних характеристик тексту [10].

Актуальна візуалізація це здебільшого - аналіз тексту. У більшості випадків аналізується цілий текстовий корпус, а не частина текстів. До текста можна застосувати різні НЛП і методи статистичного аналізу.

Сумка слів і N-грами – це мовні моделі, які використовуються для побудови обчислювального представлення для корпуса.

Інтелектуальний аналіз тексту: методи, онтології та інструменти

- Відображення тематичних (текстуальних) даних
- Діаграми: хмара слів, накладення тексту
- Таблиці: GRIDL, Періодична система
- Графіки: кругова візуалізація, перехресні карти, гліфи
- Геопросторові карти: самоорганізуюча карта (SOM)
- Мережні графіки: мережі спільного використання слів, концептуальні карти, накладання наукових карт, візуалізація дерев

Хмара слів також відома як "тег-хмара" або "словесна хмара" (рис. 1.), є графічним зображенням слів, де частота вживання слова в тексті відображається його розміром чи кольором. Це ефективний інструмент для візуалізації та аналізу частотності вживання слів у тексті. Хмара слів дозволяє швидко отримати уявлення про ключові теми та терміни у тексті. Слова, що найчастіше зустрічаються, будуть більшими і більш помітними. Завдяки хмарі слів можна визначити основні теми, які покриває текст чи набір текстів. Вона дозволяє швидко виявити ключові концепції та ідеї. Великі слова у хмарі можуть вказувати на сильний фокус або емоційне навантаження у тексті. Використання хмари слів в лінгвістиці дозволяє швидко отримувати важливу інформацію про текст та виявляти мовні особливості, полегшуючи аналіз та розуміння лінгвістичних явищ.

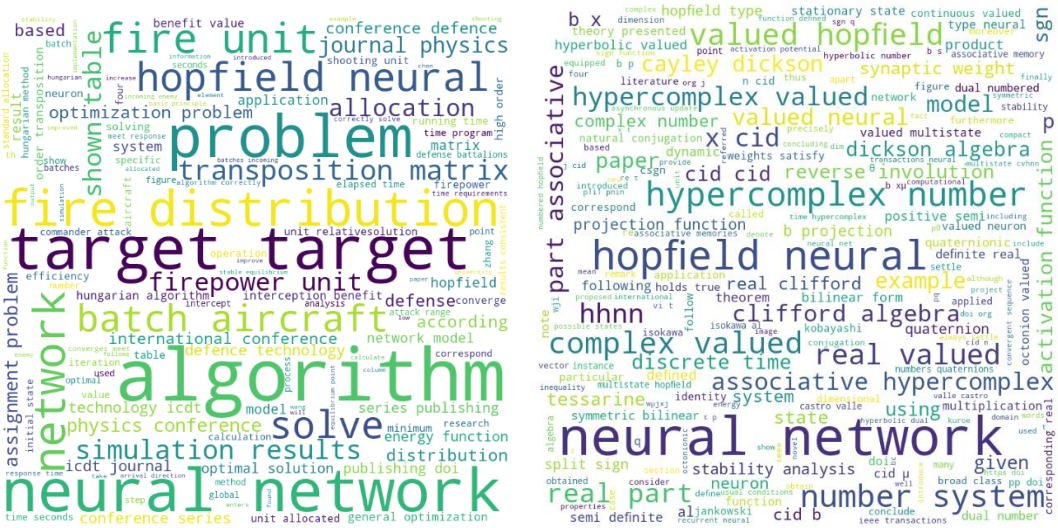


Рис. 1. Приклади хмар слів для цілком оброблених статей

Розробка моделі програмного забезпечення (рис. 2) для аналізу схожості наукових статей може бути важливою для дослідників, академічних установ, а також компаній, що займаються інтелектуальним власництвом. Ось загальна структура такої моделі:

Модуль збору та попередньої обробки даних може містити у собі компонент по збору статей, а саме визначення джерел [11,12], з яких будуть зібрані наукові статті (наприклад, бази даних, журнали, конференції). Крім того може містити компонент попередньої обробка - токенізації, лематизації, видалення стоп-слів, нормалізації регістру.

Також необхіден модуль векторизації тексту, який буде слугувати основним елементом перетворення текстових даних в числові вектори, враховуючи важливість кожного слова в контексті статті та корпусу статей загалом.

Модуль порівняння схожості може бути побудований на багатьох алгоритмах, деякі з которих розглянуті вище. Можливе використання косинусного відстані між векторами TF-IDF для визначення схожості між статтями. Окрім того можна запропонувати використання нейронних мереж, таких як Siamese Networks або Transformer-based моделі для визначення схожості.

Модуль візуалізації результатів може включати в себе використання хмари слів для візуалізації ключових термінів у схожих статтях, а також візуалізацію за допомогою створення графіків для відображення ступеня схожості між статтями у вигляді мережі або кластерів.

Модуль оптимізації , що містить у собі застосування зворотного зв'язку, що має враховувати отримані відгукі для постійного вдосконалення моделі. Бажано також мати тюнінг параметрів, тобто оптимізацію параметрів для досягнення кращої продуктивності.

Модуль інтеграції з іншими інструментами та бібліотеками повинен мати API для доступу, тобто забезпечення API для взаємодії з іншими програмами та інструментами. Крім того важлива інтеграція з платформами зберігання статей, а саме взаємодія з платформами, такими як PubMed, Google Scholar, Scopus, Web of Science для автоматичного оновлення бази даних.

Модуль забезпечення безпеки та конфіденційності, який містить у собі такі компоненти, як аутентифікація та авторизація – захист доступу до системи, а також компонент шифрування даних задля захисту конфіденційності зібраних даних.

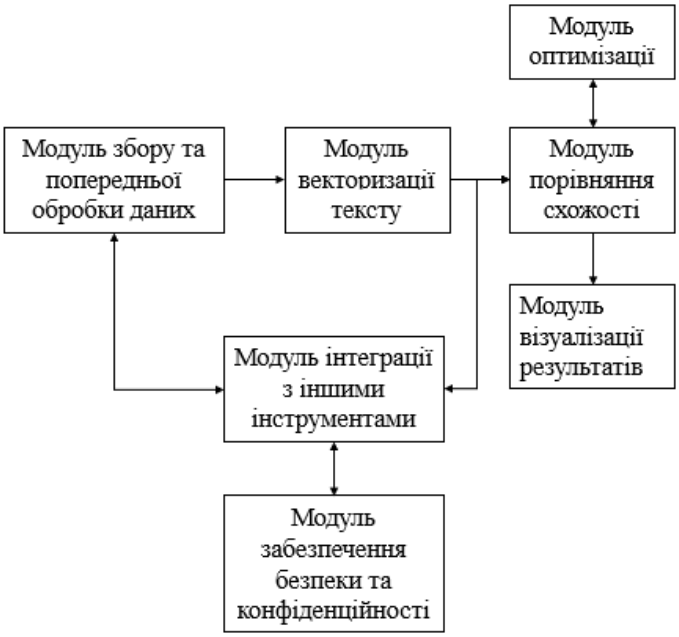


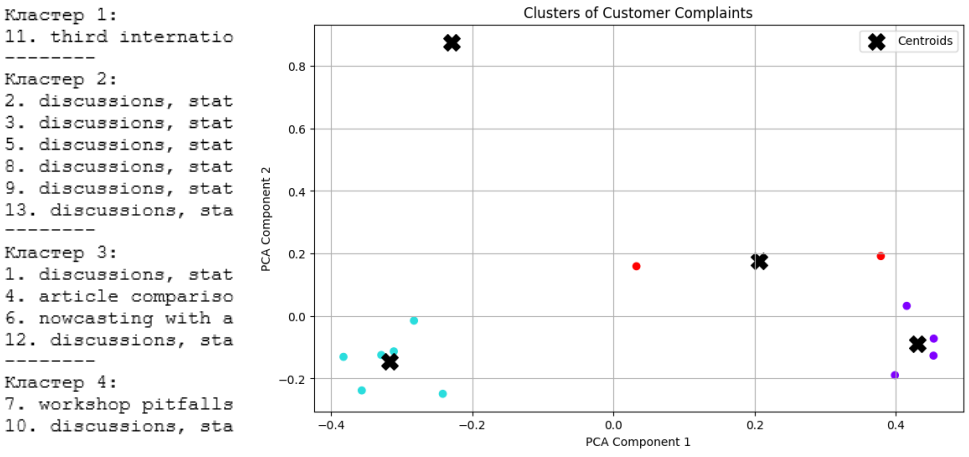
Рис. 2. Дослідницька модель програмного забезпечення для аналізу схожості наукових статей

Практичне застосування алгоритмів аналізу схожості текстів є широким і важливим в різних сферах, включаючи лінгвістику, інформаційний пошук, комп'ютерні науки, аналіз даних і багато інших. Ось кілька сценаріїв застосування:

Аналіз схожості текстів може бути використаний для виявлення плагіату чи копіювання контенту в академічних або веб-текстах. Алгоритми порівнюють структуру та слова у текстах, щоб виявити схожі патерни.

В результаті роботи по кожній із статей можна отримати хмару слів, яка візуально відображає найуживаніші ключові слова, та візуально оцінити їх вагу та кількість у статті, що аналізується (рис. 1,).

Також, у якості результата, розроблений програмний продукт видає дослідникові інформацію, щодо кластерів схожості всіх статей, що приймали участь у дослідженні у двох виглядах: у текстовому вигляді (рис. 3а), та у вигляді двовимірного графіка із означенням центрів кластерів, що були виявлені (рис. 3б.).



а) б)

Рис. 3. Розподіл статей по кластерах. а) текстовий б) графічний двовимірний

Висновки та перспективи подальшого дослідження. У роботі було досліджено та проаналізовано сучасні методи автоматизованого порівняння анотацій наукових статей із

використанням статистичних підходів до обробки природної мови. Запропонована модель програмного забезпечення дозволяє здійснювати ефективний аналіз текстової подібності, що є важливим для наукової комунікації, автоматизованого літературного огляду та виявлення плагіату. Використання таких алгоритмів, як Cosine Similarity, Jaccard Similarity та TF-IDF, у поєднанні з сучасними нейромережевими моделями, зокрема SciBERT та BioBERT, забезпечує високу точність при обробці та класифікації наукових текстів. Отримані результати демонструють, що автоматизовані системи аналізу наукових публікацій можуть значно підвищити швидкість і об'єктивність бібліографічного аналізу, зменшуючи ризик дублювання досліджень та полегшуючи пошук релевантних джерел.

Подальші дослідження у цьому напрямі мають зосередитися на оптимізації існуючих алгоритмів, зокрема на вдосконаленні методів текстового порівняння шляхом поєднання статистичних моделей із сучасними глибинними нейронними мережами. Важливим напрямом є також розширення функціональних можливостей програмного забезпечення, зокрема інтеграція методів семантичного аналізу, які дозволять не лише оцінювати поверхневу подібність текстів, а й глибше розуміти їхній зміст. Додаткової уваги потребує адаптація алгоритмів до міждисциплінарних досліджень, що дасть змогу підвищити якість текстової аналітики в різних наукових сферах.

Перспективним напрямом є розробка механізмів інтеграції з глобальними науковими базами даних, такими як Scopus, Web of Science та Google Scholar, що забезпечить автоматичне оновлення аналізованого корпусу наукових статей. Дослідження візуалізаційних технологій також відкривають нові можливості для представлення результатів аналізу у вигляді інтуїтивно зрозумілих графічних моделей, що сприятиме більш ефективному сприйняттю інформації. Подальше вдосконалення методів автоматизованого аналізу наукових текстів сприятиме розвитку інформаційної екосистеми науки та підвищенню ефективності наукових досліджень.

Список бібліографічного опису:

1. Turrisi, R. (2023). *Beyond original Research Articles Categorization via NLP*. Retrieved from https://arxiv.org/abs/2309.07020.
2. Masoumi, S., Amirkhani, H., Sadeghian, N. et al. Natural language processing (NLP) to facilitate abstract review in medical research: the application of BioBERT to exploring the 20-year use of NLP in medical research. Syst Rev 13, 107 (2024). https://doi.org/10.1186/s13643-024-02470-y
3. Timur, Ishankulov., Gleb, Danilov., Konstantin, Kotik., Yu., N., Orlov., Mikhail, Shifrin., Alexander, Potapov. (2022). The Classification of Scientific Abstracts Using Text Statistical Features. MedInfo, 290:263-267. doi: 10.3233/SHTI220075
4. Starukhin, Yaroslav & Diukarev, Vladimir. (2024). Automation of text data processing using NLP. The American Journal of Engineering and Technology. 6. 24-39. 10.37547/tajet/Volume06Issue07-04.
5. (2022). The Classification of Scientific Abstracts Using Text Statistical Features. doi: 10.3233/shti220075
6. V, G, Dubrovin., Larysa, Deineha., Anastasiya, Yatsenko. (2023). Statistical analysis software. Electrical Engineering and Power Engineering, doi: 10.15588/1607-6761-2023-3-3
7. Safoora, Masoumi., Hossein, Amirkhani., Najmeh, Sadeghian., Saeid, Shahraz. (2024). Natural language processing (NLP) to facilitate abstract review in medical research: the application of BioBERT to exploring the 20-year use of NLP in medical research. Systematic Reviews, 13 doi: 10.1186/s13643-024-02470-y
8. Ayhan, Arısoy. (2024). Natural language processing algorithms and performance comparison. Yalvaç akademi dergisi, doi: 10.57120/yalvac.1536202
9. Marbilia, Possagnolo, Sergio., Talita, de, Souza, Costa., Marcelo, S., de, Paula, Pessoa., Paulo, Sérgio, Martins, Pedro. (2019). A Semantic Approach to Support the Analysis of Abstracts in a Bibliographical Review. 259-264. doi: 10.1109/WETICE.2019.00062
10. Atanassova, Iana & Bertin, Marc & Lariviere, Vincent. (2016). On the Composition of Scientific Abstracts. Journal of Documentation. 72. 10.1108/JDOC-09-2015-0111.
11. Sérgio, Eduardo, de, Paiva, Gonçalves., Paulo, Cortez., Sérgio, Moro. (2018). A deep learning approach for sentence classification of scientific abstracts. 479-488. doi: 10.1007/978-3-030-01424-7_47
12. Muhammad, Rizqi, Nur., Gandhi, Surya, Buana., Nur, Aini, Rakhmawati. (2023). Comparative Analysis of Research Article Matching using SIF, RNN, Attention, and Hybrid Methods. 170-175. doi: 10.1109/icts58770.2023.10330854

References:

1. Turrisi, R. (2023). *Beyond original Research Articles Categorization via NLP*. Retrieved from https://arxiv.org/abs/2309.07020.
2. Masoumi, S., Amirkhani, H., Sadeghian, N. et al. Natural language processing (NLP) to facilitate abstract review in medical research: the application of BioBERT to exploring the 20-year use of NLP in medical research. Syst Rev 13, 107 (2024). https://doi.org/10.1186/s13643-024-02470-y
3. Timur, Ishankulov., Gleb, Danilov., Konstantin, Kotik., Yu., N., Orlov., Mikhail, Shifrin., Alexander, Potapov. (2022). The Classification of Scientific Abstracts Using Text Statistical Features. MedInfo, 290:263-267. doi: 10.3233/SHTI220075

4. Starukhin, Yaroslav & Diukarev, Vladimir. (2024). Automation of text data processing using NLP. The American Journal of Engineering and Technology. 6. 24-39. 10.37547/tajet/Volume06Issue07-04.
5. (2022). The Classification of Scientific Abstracts Using Text Statistical Features. doi: 10.3233/shti220075
6. V, G, Dubrovin., Larysa, Deineha., Anastasiya, Yatsenko. (2023). Statistical analysis software. Electrical Engineering and Power Engineering, doi: 10.15588/1607-6761-2023-3-3
7. Safoora, Masoumi., Hossein, Amirkhani., Najmeh, Sadeghian., Saeid, Shahraz. (2024). Natural language processing (NLP) to facilitate abstract review in medical research: the application of BioBERT to exploring the 20-year use of NLP in medical research. Systematic Reviews, 13 doi: 10.1186/s13643-024-02470-y
8. Ayhan, Arısoy. (2024). Natural language processing algorithms and performance comparison. Yalvaç akademi dergisi, doi: 10.57120/yalvac.1536202
9. Marbilia, Possagnolo, Sergio., Talita, de, Souza, Costa., Marcelo, S., de, Paula, Pessoa., Paulo, Sérgio, Martins, Pedro. (2019). A Semantic Approach to Support the Analysis of Abstracts in a Bibliographical Review. 259-264. doi: 10.1109/WETICE.2019.00062
10. Atanassova, Iana & Bertin, Marc & Lariviere, Vincent. (2016). On the Composition of Scientific Abstracts. Journal of Documentation. 72. 10.1108/JDOC-09-2015-0111.
11. Sérgio, Eduardo, de, Paiva, Gonçalves., Paulo, Cortez., Sérgio, Moro. (2018). A deep learning approach for sentence classification of scientific abstracts. 479-488. doi: 10.1007/978-3-030-01424-7_47
12. Muhammad, Rizqi, Nur., Gandhi, Surya, Buana., Nur, Aini, Rakhmawati. (2023). Comparative Analysis of Research Article Matching using SIF, RNN, Attention, and Hybrid Methods. 170-175. doi: 10.1109/icts58770.2023.10330854