https://doi.org/10.31891/2307-5732-2025-355-59
УДК 004.9

ПАДУЧАК ОЛЕГ
Національний університет «Львівська політехніка»
https://orcid.org/0009-0001-6292-6277
e-mail: oleh.v.paduchak@lpnu.ua
САМОТИЙ ВОЛОДИМИР
Національний університет «Львівська політехніка»
https://orcid.org/0000-0003-2344-2576
e-mail: volodymyr.v.samotyi@lpnu.ua

АНАЛІЗ НАУКОВИХ ТЕКСТІВ ДЛЯ ВИЗНАЧЕННЯ ПОТЕНЦІЙНИХ СПІВАВТОРІВ

У статті проаналізовано сучасні підходи до пошуку наукових співавторів, зокрема засоби, що базуються на аналізі текстових даних. Розглянуто ключові недоліки традиційних методів, таких як використання наукових баз даних і академічних соціальних мереж, що обмежують точність і повноту пошуку потенційних партнерів. Зазначено, що такі методи часто фокусуються лише на явних зв'язках між дослідниками, ігноруючи приховані спільноти або менш очевидні точки перетину наукових інтересів. Наголошено на перспективності використання інтелектуального аналізу тексту (text mining) та великих мовних моделей (LLM) для виявлення дослідників із релевантною експертизою. Показано, що ці методи можуть автоматично витягати ключові концепції з текстів наукових статей, визначати спільні дослідницькі інтереси та виявляти потенційно продуктивні наукові тандеми. Зазначено переваги цих методів, зокрема здатність ідентифікувати приховані зв'язки між дослідниками та забезпечувати глибший контекстуальний аналіз. Крім того, підкреслено можливість застосування таких методів для пошуку нових міждисциплінарних напрямків досліджень. Окремо розглянуто питання оптимізації LLM для зменшення витрат на обчислювальні ресурси та екологічного навантаження. Наголошено на необхідності використання менш ресурсомістких архітектур та алгоритмів з підвищеною ефективністю обчислень. Підкреслено важливість комбінування класичних статистичних підходів із сучасними мовними моделями для підвищення ефективності рекомендацій. Розглянуто методи адаптації LLM під специфіку наукових текстів для поліпшення якості результатів. Отримані результати свідчать про значний потенціал таких методів у розв'язанні проблеми ідентифікації наукових співавторів у сучасних системах відкритої науки. Окрім цього, окреслено перспективи подальших досліджень у напрямку поєднання різних підходів до текстового аналізу для створення комплексних рішень у галузі наукометрії.

Ключові слова: Аналіз тексту, Наукові публікації, Мережа дослідників, Рекомендація співавторів, Великі мовні моделі.

PADUCHAK OLEH
SAMOTYY VOLODYMYR
Lviv Polytechnic National University

ANALYSIS OF SCIENTIFIC TEXTS FOR IDENTIFYING POTENTIAL CO-AUTHORS

The article analyzes modern approaches to finding scientific co-authors, particularly methods based on text data analysis. It discusses the key drawbacks of traditional methods, such as the use of scientific databases and academic social networks, which limit the accuracy and completeness of the search for potential partners. It is noted that these methods often focus only on explicit links between researchers, overlooking hidden communities or less obvious intersections of scientific interests. The potential of using text mining and large language models (LLM) to identify researchers with relevant expertise is emphasized. These methods can automatically extract key concepts from scientific papers, identify shared research interests, and uncover potentially productive scientific pairings. The advantages of these methods include the ability to identify hidden connections between researchers and provide deeper contextual analysis. Furthermore, the possibility of applying these methods to discover new interdisciplinary research directions is highlighted. The article also addresses the optimization of LLM to reduce computational resource costs and environmental impact. The necessity of using less resource-intensive architectures and algorithms with increased computational efficiency is emphasized. The importance of combining traditional statistical approaches with modern language models to enhance the effectiveness of recommendations is underlined. Methods of adapting LLM to the specifics of scientific texts to improve the quality of results are also discussed. The results suggest a significant potential of these methods in solving the problem of identifying scientific co-authors in modern open science systems. Additionally, the article outlines prospects for further research in combining various text analysis approaches to create comprehensive solutions in the field of scientometrics.

Keywords: Text Mining, Scientific Publications, Researcher Network, Co-authorship Recommendation, Large Language Models.

Стаття надійшла до редакції / Received 26.05.2025
Прийнята до друку / Accepted 26.06.2025

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

Сучасний науковий процес значною мірою залежить від ефективної комунікації між дослідниками та формування продуктивних наукових колаборацій. Однак зі зростанням обсягів наукової інформації та збільшенням кількості дослідницьких проєктів і публікацій постає проблема ідентифікації потенційних співавторів. Ускладнюється не лише пошук фахівців із суміжних галузей, а й виявлення дослідників, чиї наукові інтереси та методологічний підхід можуть бути релевантними для спільної роботи.

Одним із перспективних напрямів розв'язання цієї проблеми є використання автоматизованих методів аналізу текстових даних. Завдяки цим методам можливо здійснювати масштабний аналіз наукових публікацій,

виявляючи потенційно продуктивні наукові зв'язки на основі спільних тем, ключових слів або методологічних підходів. Дослідження в цій галузі активно розвиваються, пропонуючи нові алгоритми та підходи для поліпшення точності рекомендацій.

Пошук відповідного співавтора є важливим етапом у науковій діяльності, оскільки правильно обраний партнер може значно покращити якість дослідження, розширити його міждисциплінарний характер та сприяти збільшенню цитованості результатів. Співавтор з відповідною експертизою може запропонувати нові ідеї, надати доступ до нових методологій або сприяти залученню додаткових ресурсів. Ефективна співпраця також підвищує ймовірність публікації у високорейтингових журналах і сприяє зміцненню наукових зв'язків.

У сучасних умовах пошук співавторів здебільшого відбувається через наукові бази даних, такі як Scopus, Web of Science або Google Scholar. Дослідники здійснюють пошук за ключовими словами, назвами статей або тематичними напрямками, щоб знайти потенційних партнерів. Після виявлення релевантних авторів користувачі аналізують їхні профілі, публікації та афіліації, щоб оцінити відповідність наукових інтересів.

Також важливим інструментом є академічні соціальні мережі, такі як ResearchGate чи Academia.edu, надають додаткові можливості для пошуку співавторів шляхом встановлення зв'язків між дослідниками на основі їхніх профілів та активності на платформі. Крім того, активно використовуються особисті контакти та рекомендації колег, особливо на конференціях, семінарах або в межах професійних спільнот. Проте цей підхід залежить від соціальних зв'язків і не завжди охоплює всіх потенційних партнерів.

Існуючі інструменти для пошуку співавторів мають кілька важливих недоліків, які обмежують їх ефективність. Одним із таких недоліків є неповнота та застарілість даних. Науковці не завжди регулярно оновлюють свої профілі, що призводить до того, що інформація про їхні дослідження може бути неповною або неактуальною. Це ускладнює пошук релевантних партнерів, оскільки науковці можуть бути представлені лише частково, і не всі їхні публікації можуть бути відображені в системах.

Ще одним обмеженням є недостатня специфічність пошукових алгоритмів. Більшість платформ дозволяє проводити пошук за загальними критеріями, такими як ключові слова або автори, що може призвести до дуже широких або нечітких результатів. Це ускладнює точний підбір співавторів, особливо в контексті складних, міждисциплінарних досліджень, де важливо враховувати конкретні підходи, методи чи підгалузі [1] [2].

Існує також проблема обмеженого охоплення дослідників. Платформи, такі як академічні соціальні мережі, часто фокусуються лише на певних колах дослідників, що обмежує доступ до нових або менш відомих науковців, які можуть мати релевантні навички та інтереси, але не є активними в цих мережах. Окрім того, якість анотацій до існуючих досліджень залишає бажати кращого [3]. Автори можуть опустити частину критично важливої інформації через особисті переконання, людський фактор та інші причини. Під кінець, традиційні інструменти часто не взаємодіють між собою, що вимагає від дослідників використання кількох платформ для збору інформації. Це ускладнює процес і знижує його ефективність, оскільки науковець може не побачити повної картини потенційних партнерів, а також втратити час на пошук на різних ресурсах.

Аналіз досліджень та публікацій

Як і говорилося раніше поточні засоби пошуку наукових співавторів не завжди дають бажаного результату, та вимагаю докладання великої кількості зусиль. Враховуючи поточні обмеження наявних засобів, логічним висновком постає розгляд інших способів для вирішення цієї проблеми, і найбільш очевидним із них постають наукові тексти, оскільки вони містять детальні дані про тематику досліджень, методологію та наукові підходи. На відміну від профілів у наукових базах, вони дозволяють точніше оцінити фаховість дослідника та його спеціалізацію.

Крім того, аналіз текстів дає змогу виявити структуру наукової співпраці через бібліографічні посилання та спільні публікації, що сприяє ідентифікації активних дослідницьких спільнот. Таким чином, наукові тексти є не лише джерелом явної інформації про дослідників, а й основою для глибшого аналізу зв’язків між їхніми науковими інтересами, що робить їх ефективним інструментом у вирішенні проблеми пошуку співавторів.

Відповідно, для аналізу текстів, в тому числі й наукових використовується процес інтелектуального аналізу тексту (англ. text mining). Цей підхід відкриває нові можливості для ефективного аналізу наукових текстів, що робить його цінним інструментом у процесі пошуку співавторів. Завдяки здатності обробляти великі обсяги текстових даних, цей підхід дозволяє досліджувати значно ширший спектр наукових публікацій, ніж традиційні методи [4] [5]. Автоматизоване виявлення ключових слів, тем та концепцій забезпечує глибший аналіз змісту статей і сприяє ідентифікації дослідників із релевантною експертизою, навіть якщо їхні праці не мають явних спільних ознак [6].

Особливо важливою є здатність text mining виявляти приховані зв’язки між дослідженнями. Аналіз бібліографічних посилань, спільних тем чи згадок науковців дозволяє виявити потенційних партнерів, які працюють у суміжних напрямах, що може залишитися непомітним при стандартному пошуку. Крім того, автоматизація цього процесу значно скорочує час, необхідний для аналізу наукових текстів, а постійне оновлення баз даних забезпечує актуальність результатів. Поєднання цих переваг робить text mining не лише корисним, а й необхідним інструментом для розширення наукової співпраці в умовах постійного зростання кількості наукових публікацій.

Попри потенціал text mining у розв'язанні проблеми пошуку співавторів, його застосування у контексті наукових текстів супроводжується низкою суттєвих обмежень. Однією з основних проблем є неоднорідність текстових даних, зумовлена значними відмінностями у структурі, стилі викладу та термінології між різними науковими дисциплінами. Наукові статті часто містять спеціалізовану лексику, яка вимагає ретельної побудови словників та моделей обробки природної мови (NLP) для забезпечення коректного аналізу. Чітким проявом цієї проблеми є поширеність досліджень щодо використання індивідуального аналізу тексту в межах вузьких галузей [2] [7].

Також великим недоліком цього підходу є те що інтелектуальний аналіз тексту є енергомістким процесом, що становить суттєву проблему при аналізі великих обсягів наукових текстів. Навчання моделей типу трансформерів вимагає значних обчислювальних ресурсів, а їх подальше використання для кластеризації або аналізу наукових зв'язків залишається енерговитратним [8] [9]. Це не лише збільшує фінансові витрати, а й створює екологічне навантаження, оскільки тренування таких моделей може генерувати значні обсяги викидів CO_2. Оптимізація обчислювальних процесів стає критично важливою для мінімізації цих негативних наслідків.

Використання великих мовних моделей для аналізу наукових праць

Великі мовні моделі (LLM) набули широкої популярності завдяки здатності обробляти складні мовні конструкції та генерувати зв'язний текст із високою точністю. Їхнє активне використання охоплює різні сфери, включно з обробкою наукових текстів, де необхідно ідентифікувати приховані патерни та виявляти тематичні зв'язки [10].

Завдяки глибокому навчанні на масштабних масивах текстів, LLM демонструють високу здатність до узагальнення інформації, що дозволяє виявляти приховані патерни у наукових статтях, корелювати тематичні напрями та ідентифікувати дослідників із суміжними інтересами [11] [12]. Прикладом цього є застосування великих мовних моделей із використанням генерації доповненої пошуком (RAG) для виділення онтологій та створення актуальних анотацій у біології та медицині [13][14].

Окрім цього, LLM добре адаптуються до різноманітних форматів наукових текстів, що важливо в умовах неоднорідності стилю викладу у різних дисциплінах. Їхня здатність до контекстуального аналізу дозволяє точно визначати ключові поняття, методологічні підходи та тематику досліджень навіть у складних мовних конструкціях [10].

Як і говорилося в попередньому розділі, великі мовні моделі споживають велику кількість електроенергії під час обробки великих текстів, що становить велику перепону для їхнього використання. Відповідно, постає необхідність їхньої оптимізації, і одним із ефективних підходів для досягнення ефективності є їхня дистиляція великих мовних моделей у менші більш специфічні моделі [15]. Успішне застосування цього методу було застосовано в дослідженні, яке показало збільшену ефективність моделі у десятки разів, що також дозволило зменшити витрати у 3-6 разів [8]. Це доводить те, що використання сирих великих мовних моделей для опрацювання великих масивів даних є недоцільним.

Попри те що в минулому розділі було описано переваги використання великих мовних моделей, варто розглянути й інші методи індивідуального аналізу тексту, оскільки попри високу їхню точність, та відносну простоту застосування, вони вимагають велику кількість ресурсів для ефективного застосування. Найпоширенішим серед них є статистичні моделі. Наприклад, латентний розподіл Діріхле можна використовувати для тематичного моделювання наукових текстів [16], яке у свою чергу може бути використаним для визначення потенційних співавторів.

Окрім цього, такі моделі як моделювання тема-шум можуть бути також успішно використані із цими ж цілями[17]. Також, попри наявність досліджень про застосування окремих підходів та моделей для вирішення цієї проблеми, поширеним є використання кількох підходів водночас для збільшення продуктивності, фільтрації надлишкових даних та отримання точніших результатів [18]. Відповідно, не варто відкидати можливість поєднаднання класичних підходів із застосуванням великих мовних моделей.

Висновки з даного дослідження

Отже, пошук наукових співавторів на основі текстових даних є перспективним напрямом, що може значно підвищити ефективність і точність виявлення потенційних колаборацій. Аналіз літератури свідчить, що методи інтелектуального аналізу тексту (text mining) та великі мовні моделі (LLM) мають значний потенціал у цій сфері. Зокрема, text mining демонструє високу ефективність за умов якісної обробки текстових корпусів, тоді як LLM забезпечують глибший контекстуальний аналіз, хоча й потребують значних обчислювальних ресурсів.

Існують різні підходи до використання текстових даних у пошуку співавторів, кожен з яких має свої переваги та обмеження. Поєднання класичних статистичних методів із сучасними LLM виглядає перспективним для підвищення точності результатів. Однак вибір конкретного підходу залежить від доступних текстових даних, їх якості та поставлених дослідницьких завдань.

Дивлячись в майбутнє, перспективними напрямами подальших досліджень є розроблення гібридних методів, що поєднують text mining і LLM для підвищення якості рекомендацій, а також вдосконалення алгоритмів попередньої обробки текстових даних. Такі дослідження мають потенціал суттєво розширити можливості систем підтримки наукових досліджень, сприяючи створенню нових ефективних механізмів пошуку наукових колаборацій.

Література

1. Ramprasad, R., Batra, R., & Pilania, G. (2017). Machine learning in materials informatics: Recent applications and prospects. In *npj Computational Materials* (Vol. 3, pp. 54–56). Springer Nature. https://doi.org/10.1038/s41524-017-0056-5

2. Gao, X., Tan, R., & Li, G. (2020). Research on text mining of material science based on natural language processing. In *IOP Conference Series: Materials Science and Engineering* (Vol. 768, p. 072094). IOP Publishing. https://doi.org/10.1088/1757-899X/768/7/072094

3. Westergaard, D., Stærfeldt, H., Tønsberg, C., Jensen, L., & Brunak, S. (2017). Text mining of 15 million full-text scientific articles. *bioRxiv*. Cold Spring Harbor Laboratory. https://doi.org/10.1101/162099

4. Baghela, V. (2012). Text mining approaches to extract interesting association rules from text documents. In *JCSI International Journal of Computer Science Issues* (Vol. 9, No. 3(3)). The Science and Information Organization. ISSN 1694-0814

5. Kostoff, D., Losiewicz, P., & Oard, D. (2003). Science and technology text mining: Basic concepts (Technical Report A688514). Air Force Research Laboratory (AFRL), Rome, NY. http://www.stormingmedia.us/68/6885/A688514.html

6. Berry, M. W. (Ed.). (2004). *Survey of text mining I: Clustering, classification, and retrieval*. Springer. 261 p.

7. Hodhod, R., & Fleenor, H. (2018). A text mining based literature analysis for learning theories and computer science education. In *Advances in Intelligent Systems and Computing* (pp. 560–568). Springer. https://doi.org/10.1007/978-3-319-64861-3_52

8. Palo, F., Singhi, P., & Fadlallah, B. (2024). Performance-guided LLM knowledge distillation for efficient text classification at scale. *arXiv*. https://doi.org/10.48550/arXiv.2411.05045

9. Aquino-Brítez, S., García-Sánchez, P., Ortiz, A., & Aquino Brítez, D. A. (2025). Towards an energy consumption index for deep learning models: A comparative analysis of architectures, GPUs, and measurement tools. In *Sensors* (Vol. 25, p. 846). MDPI AG. https://doi.org/10.3390/s25030846

10. Živadinović, M. (2023). Application of large language models for text mining: The study of ChatGPT. In *ITEMA 2023* (pp. 73–80). https://doi.org/10.31410/ITEMA.S.P.2023.73

11. Xiang, J., Li, Y., He, Y., & Sun, Q. (2025). Local large language model-assisted literature mining for on-surface reactions. In *Materials Genome Engineering Advances* (Vol. 3). Wiley. https://doi.org/10.1002/mgea.88

12. Zhang, S., Zheng, D., Zhang, J., Zhu, Q., Song, X., Adeshina, S., Faloutsos, C., Karypis, G., & Sun, Y. (2024). Hierarchical compression of text-rich graphs via large language models. *arXiv*. https://doi.org/10.48550/arXiv.2406.11884

13. Kainer, D. (2025). The effectiveness of large language models with RAG for auto-annotating trait and phenotype descriptions. In *Biology Methods and Protocols* (Vol. 10). Oxford University Press. https://doi.org/10.1093/biomethods/bpaf016

14. Zhao, Y., & Goto, S. (2025). Can frontier LLMs replace annotators in biomedical text mining? Analyzing challenges and exploring solutions. *arXiv*. https://doi.org/10.48550/arXiv.2503.03261

15. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv*. https://arxiv.org/abs/1503.02531

16. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

17. Khine, M. (2025). *Text mining in educational research: Topic modeling and latent Dirichlet allocation* (pp. 5–24). Springer. https://doi.org/10.1007/978-981-97-7858-4

18. B, R. (2024). A text analytics approach to structured data text in a structured form. Key applications of text mining include text classification, sentiment analysis, topic modeling, clustering, named entity recognition (NER), text summarization, and measuring text similarity. *International Journal of All Research Education and Scientific Methods (IJARESM)*, 12(6), 2455–6211. ISSN 2455-6211

References

1. Ramprasad, R., Batra, R., & Pilania, G. (2017). Machine learning in materials informatics: Recent applications and prospects. In *npj Computational Materials* (Vol. 3, pp. 54–56). Springer Nature. https://doi.org/10.1038/s41524-017-0056-5

2. Gao, X., Tan, R., & Li, G. (2020). Research on text mining of material science based on natural language processing. In *IOP Conference Series: Materials Science and Engineering* (Vol. 768, p. 072094). IOP Publishing. https://doi.org/10.1088/1757-899X/768/7/072094

3. Westergaard, D., Stærfeldt, H., Tønsberg, C., Jensen, L., & Brunak, S. (2017). Text mining of 15 million full-text scientific articles. *bioRxiv*. Cold Spring Harbor Laboratory. https://doi.org/10.1101/162099

4. Baghela, V. (2012). Text mining approaches to extract interesting association rules from text documents. In *JCSI International Journal of Computer Science Issues* (Vol. 9, No. 3(3)). The Science and Information Organization. ISSN 1694-0814

5. Kostoff, D., Losiewicz, P., & Oard, D. (2003). Science and technology text mining: Basic concepts (Technical Report A688514). Air Force Research Laboratory (AFRL), Rome, NY. http://www.stormingmedia.us/68/6885/A688514.html

6. Berry, M. W. (Ed.). (2004). *Survey of text mining I: Clustering, classification, and retrieval*. Springer. 261 p.

7. Hodhod, R., & Fleenor, H. (2018). A text mining based literature analysis for learning theories and computer science education. In *Advances in Intelligent Systems and Computing* (pp. 560–568). Springer. https://doi.org/10.1007/978-3-319-64861-3_52

8. Palo, F., Singhi, P., & Fadlallah, B. (2024). Performance-guided LLM knowledge distillation for efficient text classification at scale. *arXiv*. https://doi.org/10.48550/arXiv.2411.05045

9. Aquino-Brítez, S., García-Sánchez, P., Ortiz, A., & Aquino Brítez, D. A. (2025). Towards an energy consumption index for deep learning models: A comparative analysis of architectures, GPUs, and measurement tools. In *Sensors* (Vol. 25, p. 846). MDPI AG. https://doi.org/10.3390/s25030846

10. Živadinović, M. (2023). Application of large language models for text mining: The study of ChatGPT. In *ITEMA 2023* (pp. 73–80). https://doi.org/10.31410/ITEMA.S.P.2023.73

11. Xiang, J., Li, Y., He, Y., & Sun, Q. (2025). Local large language model-assisted literature mining for on-surface reactions. In *Materials Genome Engineering Advances* (Vol. 3). Wiley. https://doi.org/10.1002/mgea.88

12. Zhang, S., Zheng, D., Zhang, J., Zhu, Q., Song, X., Adeshina, S., Faloutsos, C., Karypis, G., & Sun, Y. (2024). Hierarchical compression of text-rich graphs via large language models. *arXiv*. https://doi.org/10.48550/arXiv.2406.11884

13. Kainer, D. (2025). The effectiveness of large language models with RAG for auto-annotating trait and phenotype descriptions. In *Biology Methods and Protocols* (Vol. 10). Oxford University Press. https://doi.org/10.1093/biomethods/bpaf016

14. Zhao, Y., & Goto, S. (2025). Can frontier LLMs replace annotators in biomedical text mining? Analyzing challenges and exploring solutions. *arXiv*. https://doi.org/10.48550/arXiv.2503.03261

15. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv*. https://arxiv.org/abs/1503.02531

16. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

17. Khine, M. (2025). *Text mining in educational research: Topic modeling and latent Dirichlet allocation* (pp. 5–24). Springer. https://doi.org/10.1007/978-981-97-7858-4

18. B, R. (2024). A text analytics approach to structured data text in a structured form. Key applications of text mining include text classification, sentiment analysis, topic modeling, clustering, named entity recognition (NER), text summarization, and measuring text similarity. *International Journal of All Research Education and Scientific Methods (IJARESM)*, 12(6), 2455–6211. ISSN 2455-6211