УДК 004.89:311.2:616-036.22

Б. В. Бобко[1]
С. О. Жуков[1]

АВТОМАТИЗОВАНЕ ФОРМУВАННЯ ПРОСТОРУ ОЗНАК ІЗ НЕСТРУКТУРОВАНИХ МЕДИЧНИХ ЗАПИСІВ ДЛЯ ЗАДАЧ ВИЯВЛЕННЯ СТАТИСТИЧНИХ АНОМАЛІЙ

[1]Вінницький національний технічний університет

Анотація

Роботу присвячено вирішенню задачі автоматизованої структуризації текстових медичних даних для подальшого статистичного аналізу та виявлення аномалій. Проведено дослідження методів витягування іменованих сутностей (симптомів) з неструктурованих скарг пацієнтів українською мовою. Для експериментального аналізу сформовано синтетичний корпус текстів на основі класифікатора ICPC-2, що забезпечує семантичну валідність даних. Виконано порівняльний аналіз трьох підходів: на основі правил (rule-based), з використанням великої мовної моделі (LLM) та гібридного методу. Встановлено, що гібридний підхід забезпечує найкращий баланс між повнотою (Recall) та інтерпретованістю результатів, що є критично важливим для формування достовірної вибірки при пошуку аномалій у медичній статистиці.

Ключові слова: медична статистика, NLP, NER, великі мовні моделі, LLM.

Abstract

The paper addresses the problem of automated structuring of text-based medical data for further statistical analysis and anomaly detection. Methods for Named Entity Recognition (symptom extraction) from unstructured patient complaints in Ukrainian were investigated. A synthetic text corpus based on the ICPC-2 classifier was generated for experimental analysis. A comparative analysis of three approaches was performed: rule-based, Large Language Model (LLM), and a hybrid method. It was determined that the hybrid approach provides the best balance between Recall and interpretability, which is critical for forming a reliable dataset for anomaly detection in medical statistics.

Keywords: medical statistics, NLP, NER, large language models, LLM.

Вступ

Аналіз медичної статистики є важливим інструментом для моніторингу стану здоров’я населення, планування ресурсів системи охорони здоров’я та виявлення потенційних ризиків і аномалій у медичних даних. Значна частина інформації, що використовується для формування статистичних показників, надходить із неструктурованих текстових джерел, зокрема звернень пацієнтів, клінічних нотаток та амбулаторних записів. Автоматизоване опрацювання таких даних ускладнюється мовною варіативністю, неформальністю викладу та відсутністю стандартизованих формулювань.

Особливої складності набуває задача автоматизованого виявлення симптомів, оскільки клінічні тексти часто містять описові, неповні або метафоричні формулювання, що не відповідають канонічній медичній термінології. Це ускладнює застосування класичних rule-based підходів, які ґрунтуються на фіксованих словниках і правилах, та обмежує їхню здатність до масштабування і адаптації до нових мовних конструкцій [1].

Останніми роками значну увагу приділяють інтелектуальним методам аналізу текстів, зокрема великим мовним моделям (LLM), які демонструють високу ефективність у задачах обробки природної мови [2]. Завдяки здатності враховувати контекстні залежності та узагальнювати знання, такі моделі є перспективними для витягування медичної інформації з неструктурованих текстів. Водночас їх застосування у критично важливих доменах, зокрема медицині, потребує додаткової оцінки з погляду точності, інтерпретованості та контрольованості результатів [3].

Одним із підходів до підвищення надійності автоматизованих систем аналізу є поєднання статистичних та експертних методів у межах гібридних моделей [4]. Такі підходи дозволяють використовувати сильні сторони глибинного навчання для виявлення складних мовних патернів, водночас зберігаючи можливість доменної корекції результатів за допомогою правил.

Актуальність дослідження зумовлена необхідністю перетворення "сирого" тексту на структурований набір ознак (Feature Space), придатний для машинного навчання. Застосування

виключно класичних (rule-based) або виключно нейромережевих (LLM) підходів має свої обмеження: перші не покривають мовну варіативність, другі – схильні до "галюцинацій", що неприпустимо при формуванні точної статистики.

Метою даного дослідження є порівняльний аналіз різних підходів до автоматизованого виявлення симптомів у неструктурованих медичних текстах українською мовою, що забезпечить високу точність структуризації даних для подальшого аналізу типових закономірностей та аномалій. У роботі оцінюється ефективність кожного з підходів за стандартними метриками інформаційного витягування, а також аналізуються їхні переваги та обмеження в контексті задач аналізу медичної статистики.

Результати дослідження

У межах даного дослідження було виконано порівняльний аналіз трьох підходів до автоматизованого виявлення медичних симптомів у неструктурованих текстах звернень пацієнтів українською мовою: rule-based підходу, підходу із великої мовної моделі та гібридного, що поєднує обидва підходи. Метою експерименту було оцінювання ефективності кожного з методів за стандартними метриками інформаційного витягування та визначення їхньої придатності для задач аналізу медичної статистики.

У межах експериментального дослідження було сформовано корпус текстів звернень пацієнтів на основі ICPC-2 кодів, отриманих із реального медичного датасету. Зазначені коди використовувалися для генерації текстових описів симптомів у довільній природній формі. Оскільки цільові симптоми були визначені до етапу генерації текстів, процес ручної анотації корпусу не здійснювався, а розмітка формувалася автоматично на основі відомої відповідності між згенерованими текстами та симптомами. При цьому генерація враховувала мовні варіації, синонімію та неповні формулювання, характерні для подібних описів.

Rule-based підхід реалізовано на основі словниково-патернової моделі. Було сформовано доменний словник симптомів із урахуванням морфологічних варіацій та поширених описових конструкцій, який також базувався на ICPC-2 кодах симптомів та скарг.

Як основу підходу із використанням великої мовної моделі було використано модель XLM-RoBERTa (XLM-R), яка забезпечує контекстно-залежне кодування тексту та підтримує багатомовні сценарії, зокрема українську мову. Її було донавчено (fine-tuned) на спеціалізованому корпусі анотованих медичних текстів. Вона була адаптована до задачі витягування іменованих сутностей (NER) із фокусом на медичні симптоми.

Гібридний підхід поєднує використання мовної моделі із rule-based постобробкою. На першому етапі симптоми визначалися за допомогою мовної моделі, після чого результати проходили постобробку на основі правил. Такий підхід дозволяв зменшити кількість помилкових виявлень за рахунок контекстних обмежень та доменних знань.

Оцінювання ефективності підходів здійснювалося з використанням стандартних метрик precision (точності), recall (повноти) та оцінки F1-score. Precision характеризує частку коректно виявлених симптомів серед усіх знайдених, recall — частку знайдених симптомів серед усіх наявних у тексті, а F1-score використано як інтегральний показник якості. Експериментальні результати наведено в таблиці 1.

Таблиця 1. Результати перевірки різних підходів

	Precision	Recall	F1
Rule-based підхід	0.70	0.73	0.70
Підхід на основі LLM	0.80	0.95	0.87
Гібридний (LLM + rule-based)	0.77	0.91	0.83

Rule-based підхід демонструє відносно збалансовані значення precision та recall, що підтверджує його стабільність як базового методу. Водночас обмежене покриття мовних варіацій і залежність від повноти словника зумовлюють нижчий рівень F1-score порівняно з іншими методами.

Велика мовна модель показала найвищий рівень recall, що свідчить про її здатність виявляти симптоми навіть у складних або нетипових мовних контекстах. Високе значення F1-score підтверджує

ефективність подібних моделей для аналізу неструктурованих медичних текстів. Водночас деяке зниження precision вказує на схильність моделі до надмірного узагальнення.

Гібридний підхід продемонстрував компромісні результати між точністю та повнотою. Застосування правил постобробки дозволило зменшити кількість хибнопозитивних виявлень порівняно з чисто LLM підходом, зберігаючи при цьому високий рівень recall. Отримані результати показують доцільність поєднання статистичних і експертних методів у задачах аналізу медичної статистики. Більш наочно відмінності в результатах можна побачити на рисунку 1.

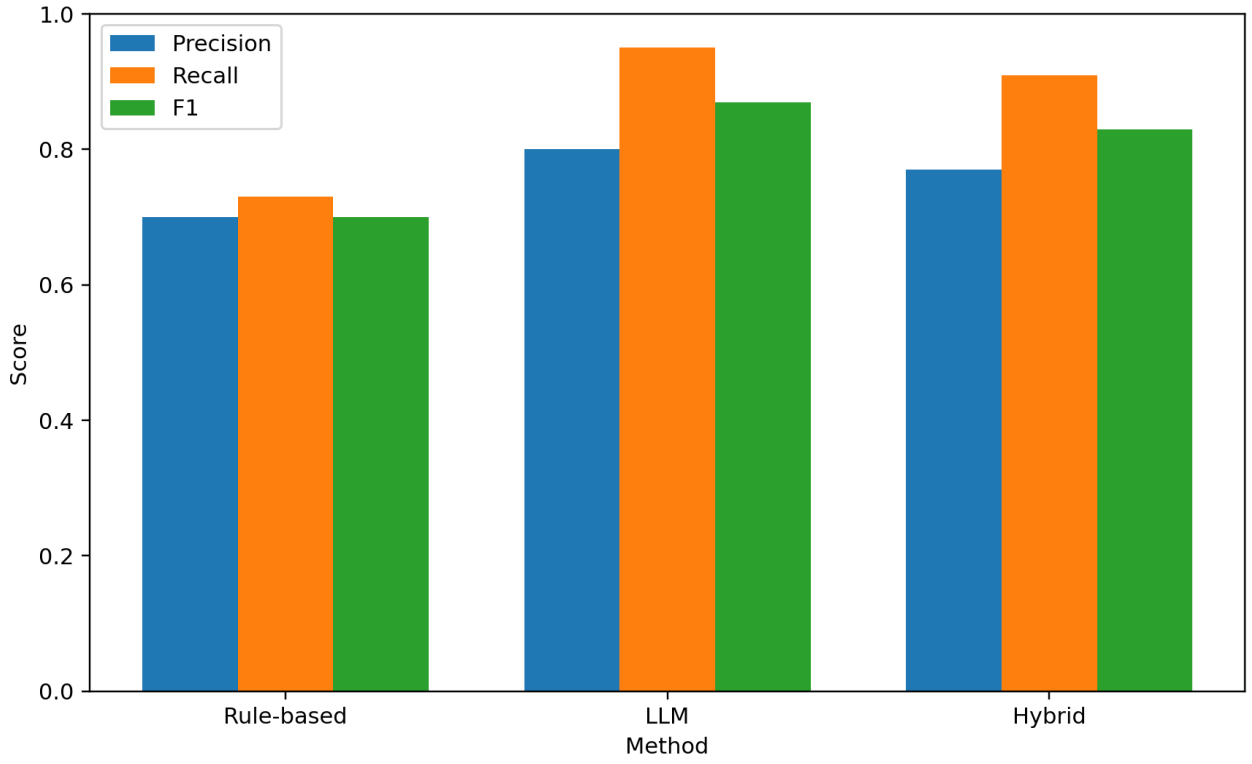

Рисунок 1 – Графічне представлення результатів тестування

Висновки

У роботі виконано порівняльне дослідження трьох підходів до автоматизованого виявлення медичних симптомів у неструктурованих текстах звернень пацієнтів українською мовою: rule-based, із використанням донавченої великої мовної моделі та гібридного. Експериментальні результати підтвердили суттєві відмінності між підходами за показниками точності та повноти, що відображає їхні концептуальні особливості.

Rule-based підхід продемонстрував стабільні, але обмежені результати, що зумовлено залежністю від повноти словників і заздалегідь визначених правил. Такий підхід може бути ефективним як базовий або допоміжний інструмент, однак його здатність до масштабування та адаптації до мовної варіативності є обмеженою.

Експериментально підтверджено, що використання донавчених великих мовних моделей (LLM) для української медичної лексики дозволяє досягти високих показників повноти (Recall = 0.95), виявляючи приховані симптоми, які пропускають словникові методи. Отримані результати підтверджують доцільність їхнього використання для аналізу неструктурованих медичних текстів, особливо в задачах, де пропуск релевантної інформації є критичним.

Гібридний підхід, що поєднує використання LLM із rule-based постобробкою, забезпечив компроміс між точністю та повнотою. Застосування експертних правил дозволило зменшити кількість хибнопозитивних спрацювань без істотної втрати recall, що є важливим для практичних систем аналізу медичної статистики. Для задач медичної статистики, де ціна помилки є високою, доцільно використовувати гібридну архітектуру.

Загалом результати дослідження підтверджують ефективність інтелектуальних методів

автоматизованого аналізу медичних текстів та показують можливість використання гібридних підходів як збалансованого рішення між якістю, інтерпретованістю та контрольованістю результатів у задачах медичної статистики. Запропонований підхід дозволяє автоматизувати процес розмітки великих масивів неструктурованих записів. Отримані структуровані дані формують репрезентативну вибірку для побудови профілів "типових закономірностей" захворюваності, що є необхідною умовою для подальшого виявлення статистичних аномалій алгоритмами машинного навчання.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Leibo Liu, Victoria Blake, Matthew Barman, Blanca Gallego, Timothy Churches, Georgina Kennedy, Sze-Yuan Ooi, Geoffrey P Delaney, Louisa Jorm. 2025. Using natural language processing to extract information from clinical text in electronic medical records for populating clinical registries: a systematic review, Journal of the American Medical Informatics Association,; ocaf176, https://doi.org/10.1093/jamia/ocaf176.
2. Luis B. Elvas, Ana Almeida, João C. Ferreira. 2025. Natural language processing in medical text processing: A scoping literature review, International Journal of Medical Informatics, Volume 204, 106049, ISSN 1386-5056, https://doi.org/10.1016/j.ijmedinf.2025.106049.
3. Білецький, Б. С. 2023. Використання методів оброблення природної мови для вилучення визначень слів із контексту. Вінницький національний технічний університет. https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2023/paper/download/18886/15651.
4. FNU Sudhakar Abhijeet. Hybrid Approaches for NER in Noisy OCR Medical Records. 2025. International Journal of Science and Research Archive, 16(03), 082-089. https://doi.org/10.30574/ijsra.2025.16.3.2499.

Бобко Богдан Володимирович — аспірант кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: bobko.bogdan@gmail.com

Жуков Сергій Олександрович — кандидат технічних наук, доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: sazhukov@gmail.com

Bobko Bohdan V. – Postgraduate student of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: bobko.bogdan@gmail.com

Zhukov Serhii O. **–** Candidate of technical sciences, associate professor of the department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: sazhukov@gmail.com