НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ОДЕСЬКА ЮРИДИЧНА АКАДЕМІЯ»

ФАКУЛЬТЕТ КІБЕРБЕЗПЕКИ ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

28 листопада 2025 року

КІБЕРБЕЗПЕКА В СУЧАСНОМУ СВІТІ: АКТУАЛЬНІ ВИКЛИКИ

МАТЕРІАЛИ VI МІЖНАРОДНОЇ НАУКОВО-ПРАКТИЧНОЇ КОНФЕРЕНЦІЇ

УДК 004.056

Відповідальній редактор:

О. В. Дикий, декан факультету кібербезпеки та інформаційних технологій, кандидат юридичних наук, доцент

Матеріали видано в авторській редакції.
Повну відповідальність за достовірність та якість поданого матеріалу несуть учасники конференції, їхні наукові керівники, які рекомендували ці матеріали до друку.

Рекомендовано до друку
Вченою радою Національного університету
«Одеська юридична академія»
(протокол №3 від 29.11.2025 р.)

Матеріали Міжнародної науково-практичної конференції «Кібербезпека в сучасному світі: актуальні виклики» (м. Одеса, 28 листопада 2025 р.). Одеса, 2025. 287 с.

У конференції взяли участь студенти, аспіранти, молоді вчені, викладачі та науковці. Конференція проводиться на базі Національного університету «Одеська юридична академія».

АНАЛІЗ МЕТОДІВ ВИЛУЧЕННЯ КЛЮЧОВОЇ ІНФОРМАЦІЇ З ЮРИДИЧНИХ ТЕКСТІВ

Гура Володимир

доцент кафедри інформаційних технологій Національного університету «Одеська юридична академія», кандидат технічних наук, доцент

Курчук Роман

студент 2-го курсу магістратури факультету кібербезпеки та інформаційних технологій Національного університету «Одеська юридична академія»

Обсяги юридичних документів зростають щодня, ускладнюючи швидке виділення ключової інформації, до якої належать учасники справи, застосовані норми права, факти, винесене рішення, мотиви та санкції. Дослідження присвячене методам автоматичного витягування такої інформації з тексту. Проаналізовано багато наукових публікацій з метою виявлення ефективних підходів. Експериментальна перевірка проведена на більш ніж 500 судових рішеннях українською мовою з ЄДРСР трьома методами. Найкращий результат показав метод доповнення контексту нормативними актами використанням пошуку відповідної статті в законодавчій базі з подальшим її використанням для інтерпретації судового тексту. Якість витягування сягнула 89%. Робота підкреслює практичну цінність інструментів для юристів, але вказує на потребу розширення україномовних анотованих наборів даних.

Цифровізація правосуддя в Україні досягла критичного рівня, Єдиний державний реєстр судових рішень (ЄДРСР) містить понад 15 мільйонів документів станом на 2025 рік, з щоденним приростом близько 3–5 тисяч [1]. Ручний аналіз одного рішення займає в середньому 20–40 хвилин, що призводить до перенавантаження суддів, адвокатів та їх співробітників. За даними Мін'юсту України (2024), понад 60% часу юристів витрачається на пошук фактів у текстах, а не на правовий аналіз [2]. Автоматизація цього процесу є не лише питанням ефективності, але й доступності правосуддя, скорочення часу аналізу, за міжнародними аналогами. Вона дозволить прискорити розгляд справ, зменшити судові черги та підвищити якість рішень. У країнах ЄС, наприклад в Естонії та Нідерландах, подібні системи вже інтегровані в судові платформи, скорочуючи час обробки на 50–60% [3]. В Україні, попри наявність відкритих даних, бракує інструментів для автоматичного витягування ключової інформації українською мовою – це створює технологічний розрив.

Ключова інформація – це основні факти, які визначають сутність документа та потрібні для прийняття юридичних рішень. За міжнародними стандартами Legal General Language Understanding Evaluation (LexGLUE) – це

бенчмарк-набір даних для оцінки моделей обробки природної мови Natural Language Processing (NLP) – це обробка природної мови, галузь штучного інтелекту (ШІ), що займається взаємодією комп'ютерів із людською мовою (текстом або усною мовою). Ключова інформація охоплює учасників справи (позивач, відповідач, суддя, свідки) через пошук імен та ролей [4, 5], застосовані норми права (статті законів чи кодексів) через виявлення посилань [6, 7], факти та обставини (дати, місця, події, суми збитків) через виділення подій та зв'язків [8], винесене рішення (визнання вини, задоволення позову, штраф) через визначення результату справи [9], мотиви та аргументи (правові обґрунтування, посилання на практику) через виділення головних тез [10, 11], а також санкції та наслідки (розмір штрафу, термін покарання) через пошук числових даних [12]. Усе це формує чіткий профіль справи, який дозволяє оцінити документ за кілька хвилин. У судових рішеннях України ці дані розкидані по тексту на 5–50 сторінок з різними формулюваннями, що суттєво ускладнює ручний пошук [13].

Цифровізація правової системи України супроводжується експоненційним зростанням обсягів судових рішень та законодавчих актів. Щоденно реєструються тисячі документів, аналіз яких вимагає від юристів значних часових витрат на пошук імен сторін, дат, норм права. Автоматизація цього процесу здатна скоротити час обробки з годин до секунд, зменшити помилки та підвищити доступність правосуддя. За кордоном аналогічні системи вже скорочують час аналізу на 40–60% [10, 11]. В Україні ЄДРСР містить відкриту базу рішень, однак спеціалізованих інструментів для їхньої автоматичної обробки бракує.

Метою дослідження є систематизація сучасних методів, їх апробація на україномовних текстах та формулювання рекомендацій щодо впровадження в судову та адвокатську практику.

Розвиток методів пройшов від простих правил до складних моделей штучного інтелекту. На початку використовували шаблони та словники, наприклад, пошук «ТОВ» як назви компанії чи «ст. 190» як посилання на закон. Такий підхід швидкий, але не витримує варіацій у формулюваннях і дає точність нижче 70% [4].

Далі з'явилися моделі, які навчаються на прикладах. Серед них Bidirectional Encoder Representations from Transformers (BERT) – це нейронна мережа, що читає текст у двох напрямках і враховує контекст. У договорах про злиття вона досягла 85% точності при пошуку сутностей [4]. Подібні результати отримано для судових рішень Індії [5].

Для оцінки ефективності методів автоматичного витягування інформації, з юридичних текстів, створено LexGLUE – набір тестів англійською

мовою для юридичних завдань, пошук імен, класифікація договорів, визначення результату справи [9]. Моделі з механізмом уваги перевершують старі методи на 15–20% у довгих текстах [15, 16].

Класичні моделі **Conditional Random Fields** (CRF) – це **статистична модель для послідовного розпізнавання** (тексту, тегів, сутностей), добре працюють з послідовністю слів і досягли 92% показників при пошуку цитат [6]. Сучасний підхід Retrieval-Augmented Generation (RAG) це ШІ, який не просто "вигадує" відповідь, а шукає її в реальних документах, а потім генерує текст на основі знайденого, це дозволяє шукати потрібні статті в законодавчій базі з подальшим її використанням для пояснення судового тексту. Такий підхід підвищив якість на 18% у текстах португальською мовою [9].

Для мов, відмінних від англійської, розроблено спеціалізовані набори даних. Наприклад, Labor Lex – це анотований датасет португальською мовою, що містить судові рішення та трудові контракти. Він дозволяє навчати моделі витягувати сутності (учасники, норми, факти) та зв'язки між ними з точністю до 85% [14]. Аналогічний підхід застосовано до текстів про COVID-19. Дані для витягування претензій та відповідних норм права забезпечують 80% точності підходу [15]. Попереднє структурування візуально багатих документів (PDF) з виділенням заголовків, абзаців та таблиць підвищує якість витягування на 12% [16], а врахування ієрархії документа – на 15% при узагальненні змісту [10]. Польська, як і українська, має складні закінчення слів. Для польських текстів використовують Rapid Automatic Keyword Extraction (RAKE) – швидкий пошук ключових фраз без навчання, з урахуванням стоп-слів (наприклад, для тендерів), з точністю 70–80% [17]. Статистичні методи на основі груп слів з попереднім спрощенням форм слів (приведення до основи за допомогою Snowball – це фреймворк для створення алгоритмів стемінгу (зведення слів до основи) різними мовами) застосовують для визначення типу справи [18]. Комбіновані моделі ШІ, наприклад Bidirectional Long Short-Term Memory + Conditional Random Field (BiLSTM-CRF) – це ШІ, який читає речення в обидва боки і розумно ставить мітки словам, враховуючи сусідів, з урахуванням нечіткої логіки, дозволяють знаходити неоднозначні терміни та метафори в контрактах, підвищуючи точність на 10–15% [19]. Пошук на основі словників права (онтології) використовують для приховування імен та класифікації [20]. Польська версія BERT добре аналізує законодавчу базу парламенту, з точністю 80–90% [21]. Ці приклади показують, що комбіновані підходи ефективні для слов'янських мов.

У медицині (PubMed, MIMIC-III) пошук на основі графів знань (з медичними словниками UMLS) виділяє симптоми, діагнози, ліки з записів, з точністю 88–92%. Можна адаптувати для зв'язків «порушення → санкція» [22]. У фінансах (звіти SEC) класифікація без навчання, за допомогою великих мовних моделей, визначає ризики та зобов'язання з точністю 85%. Підходить

для швидкого визначення типу судової справи [23]. У наукових статтях статистичний метод Yet Another Keyword Extractor (YAKE) – це швидкий, безнавчальний алгоритм для автоматичного виділення ключових слів і фраз з тексту, швидко знаходить головні теми з точністю 75–85%. Корисно для аналізу цитувань у рішеннях [24]. У соціальних мережах (Twitter/X) модель T5 це універсальна модель, яка перетворює будь-який текст, витягує події та учасників у реальному часі з точністю 80%. Вона також дозволяє виконувати моніторінг згадок судових справ у медіа [25]. У технічній документації, наприклад у описах патентів (офіційних документах про винаходи), аналіз зв'язків між словами з подальшим застосуванням класичної моделі CRF дозволяє виділяти назву винаходу та імена авторів з точністю 90%. Цей самий підхід можна адаптувати для пошуку норм права та їх застосування у судових рішеннях [26].

Ці підходи з інших сфер дозволяють порівнювати та переносити рішення, наприклад, з медичною сфери до юридичної сфери.

Існуючі інструменти з готовими моделями для україномовних ресурсів досягають 75–80% точності при витягуванні ключової інформації без додаткового навчання [27].

Сучасні інструменти акцентують увагу на виділенні головних тез, що узагальнюють документ, як у системах скорочення судових рішень [11]. Пошук на основі словників знань (онтології) структурує сутність документу [4]. У 2024 році запропоновано комбіновану модель BiLSTM-CRF з урахуванням нечіткості для пошуку метафор і складних термінів, що підвищує точність на 10–15% [9]. Для автоматичного виділення логіки рішень використовують BERT разом з BiLSTM-CRF, що дозволяє знаходити правила «якщо – то» для моделювання [30]. Ці методи доповнюють RAG, додаючи глибше розуміння змісту [10].

План проведеного дослідження включає відбір даних, їх підготовку, розмітку, навчання моделей та оцінку результатів. Експеримент відтворюваний на стандартному обладнанні з Python 3.10, бібліотеками Hugging Face Transformers, FAISS, scikit-learn та seqeval.

Набір даних сформовано з Єдиного державного реєстру судових рішень (ЄДРСР) за період 2023–2024 років. Загалом відібрано 500 унікальних судових рішень, що охоплюють цивільні (60%), кримінальні (25%) та адміністративні (15%) справи. Це забезпечує репрезентативність вибірки. Кожен документ – повний текст рішення у форматі TXT, очищений від метаданих (номер справи, дата публікації). Середня довжина – 500 слів, мінімальна – 150, максимальна – 2500. Загальний обсяг всіх документів сягає близько 250 тисяч слів.

Підготовка даних включала токенізацію (розбиття на слова та знаки), нормалізацію (приведення до нижнього регістру, видалення зайвих пробілів) та анонімізацію (заміна реальних імен на теги типу [ПОЗИВАЧ_1], [ВІДПОВІДАЧ_2] для захисту персональних даних). Кожне рішення

розмітовувалося на вісім типів сутностей, учасники (позивач, відповідач, суддя, свідок), норми права (стаття, частина, пункт), дати, місця, факти (події, суми), рішення (задоволено, відмовлено, частково), мотиви (аргументи), санкції (штраф, термін). Також позначено зв'язки між сутностями, наприклад «учасник → норма права» чи «норма права → санкція».

Набір поділено на три частини, 70% (350 документів) – для навчання моделей, 15% (75) – для перевірки під час налаштування, 15% (75) – для фінального тесту. Такий розподіл стандартний у машинному навчанні та запобігає перенавчанню.

Перевірено три різні методи витягування потрібної інформації. Перший – підхід на основі шаблонів, це простий пошук за заздалегідь підготовленими зразками (регулярними виразами) та словниками. Наприклад, шаблон для норм права — «ст.?\s*\d+» або «частина\s+\w+», словник організацій – «ТОВ», «ПАТ», «ФОП». Цей метод не потребує навчання, працює швидко, але жорсткий до варіацій у тексті. Другий метод являє собою модель на основі української версії BERT (UkBERT) з додатковим шаром для класифікації. Третій – метод доповнення контексту законом RAG, спочатку векторний пошук трьох найближчих статей законодавства з бази «Законодавство України» (FAISS, ембеддинги від Sentence-BERT), потім передача судового тексту + знайдених статей генеративній моделі (Llama-3-8B) з запитом, «Витягни ключову інформацію, учасники, норми, факти, рішення, мотиви, санкції».

Оцінка якості проводилася за трьома метриками, зрозумілими навіть без технічної підготовки. Точність – це відсоток правильних відповідей серед усіх, які система позначила як факти. Наприклад, якщо система знайшла 10 норм права, але 2 з них помилкові, точність – 80%. Повнота – відсоток знайдених фактів від усіх, що є в тексті насправді. Якщо в документі 15 учасників, а система знайшла 12 правильних, повнота – 80%. Загальна оцінка являє собою середнє гармонійне між точністю та повнотою, щоб уникнути переваги по кожному показнику. Розрахунок проводився окремо для сутностей (імена, дати, норми) та зв'язків між ними за допомогою бібліотек seqeval і scikit-learn. Кожен метод запускався тричі, а отримані результати усереднювалися.

Експеримент проведено на тестовій вибірці з 75 судових рішень, що становить 37500 слів. Середня довжина одного документа – 500 слів. Обробка виконувалася на сервері з графічним процесором NVIDIA A100. Кожен метод запускався тричі, результати усереднені. Час вимірювався від моменту завантаження тексту до отримання структурованої інформації.

Загальні результати такі, метод доповнення контексту законом RAG показав найкращі значення – 91% точності та 87% повноти для пошуку сутностей, що дає загальну оцінку 89%. Для зв'язків між сутностями – 86% точності, 82% повноти, загальна оцінка 84%. Час обробки – 0,35 секунди на документ.

Українська версія BERT досягла 88% точності та 84% повноти для сутностей, загальну оцінку 86%, для зв'язків – 83% і 79% загальна оцінка 81%, час – 0,25 секунди [28]. Підхід на основі шаблонів дав 75% точності та 69% повноти для сутностей загальна оцінка 72%, для зв'язків – 68% і 62% загальна оцінка 65%, але найшвидший – 0,05 секунди на документ.

За типами сутностей перевага RAG особливо помітна при пошуку норм права, 91% загальна оцінка проти 65% у шаблонів і 79% у BERT [29]. Це пояснюється тим, що модель бачить не лише посилання в рішенні, а й повний текст статті закону. Для учасників справи RAG досягає 94%, BERT – 90%, шаблони – 82%. Дати та місця, 90%, 87%, 78% відповідно. Рішення, 88%, 85%, 70%. Санкції, 87%, 82%, 68%.

Щодо довжини документа, RAG зберігає стабільність, для коротких текстів (менше 300 слів) – 90%, середніх (300–1000) – 88%, довгих (понад 1000) – 87%. BERT падає з 87% до 83%, шаблони – з 74% до 68%. Це показує, що зовнішній контекст допомагає навіть у складних випадках.

Приклади помилок ілюструють слабкі місця. Підхід на основі шаблонів не розпізнає «частина третя статті 190 Кримінального кодексу України», бо шукає лише «ст. 190 ч. 3». BERT плутає «стаття договору» і «стаття кодексу» у 12% випадків через схожість слів. RAG вирішує це, додаючи текст: «ст. 190 КК України – шахрайство».

Порівняння з ручним аналізом, юристи витрачають в середньому 22 хвилини на документ (за 10 експертами) [15]. Автоматизовані методи – від 0,05 до 0,35 секунди. Економія – від 3760 до 26 400 разів.

Проведене дослідження чітко доводить, що інтеграція зовнішнього нормативного контексту (RAG – це інструмент, який не вигадує, а шукає в твоїх документах і відповідає тільки за фактами) є найефективнішим способом автоматичного витягування ключової інформації з україномовних судових рішень. Цей підхід суттєво перевершує традиційні шаблони та моделі, які працюють лише з текстом документа, завдяки здатності «розуміти» зміст норм права, а не лише їхні посилання.

Головний висновок це якісне витягування інформації з юридичних текстів українською мовою неможливе без поєднання мовних моделей із базою законодавства. Самостійне використання готових моделей дає прийнятні результати, але додавання релевантних статей закону як контексту усуває більшість помилок інтерпретації, особливо при роботі з нормами права.

Важливим моментом є критична нестача анотованих даних для української юридичної мови. Існуючі набори орієнтовані на загальну мову, а спеціалізовані корпуси для судових рішень відсутні. Це гальмує розвиток моделей і пояснює, чому навіть найкращі підходи потребують додаткового навчання. Створення великого національного корпусу розмічених судових

рішень має стати стратегічним завданням для розвитку юридичного напрямку ШІ в Україні.

Нарешті, дослідження підкреслює необхідність співпраці між юристами, розробниками та державними органами. Юристи потрібні для розмітки законодавчих даних, розробники – для створення зручних інтерфейсів, держава – для забезпечення доступу до структурованих джерел законотворчої інформації.

Отже, проведена робота не просто демонструє ефективний метод, а відкриває шлях до системної цифровізації української юриспруденції, де штучний інтелект стане повсякденним помічником, забезпечуючи швидкість, точність і рівний доступ до правової інформації.

Список використаних джерел:

1. ЄДРСР. Статистика 2025. https://reyestr.court.gov.ua/.
2. Мін'юст України. Звіт про цифровізацію 2024. https://minjust.gov.ua/digital-report-2024.
3. EU e-Justice Portal. Digital Justice Systems. https://e-justice.europa.eu.
4. Bhattacharyya A. et al. (2023). Merger Agreement Dataset. EMNLP 2023. https://aclanthology.org/2023.emnlp-main.1019.pdf.
5. Kalamkar P. et al. (2022). Named Entity Recognition in Indian Court Judgments. NLLP 2022. https://aclanthology.org/2022.nllp-1.15.pdf
6. Martinez-Gil J. (2025). Legal Citation Detection in UK Judgments. EMNLP 2025. https://aclanthology.org/2025.emnlp-main.1361.pdf
7. Angelidis I. et al. (2023). Legal Citation Extraction. ACL 2023. https://aclanthology.org/2023.findings-acl.123
8. Glaser I. et al. (2024). Event Extraction in Legal Texts. NLLP 2024.
9. Chalkidis I. et al. (2022). LexGLUE: A Benchmark for Legal Language Understanding. ACL 2022. https://aclanthology.org/2022.acl-long.297.pdf
10. Zhong Y., Litman D. (2022). Structural Computation for Case Summarization. NLLP 2022. https://aclanthology.org/2022.nllp-1.30.pdf
11. Galgani F. et al. (2023). Catchphrase Extraction for Legal Document Summarization. NLLP 2023. https://aclanthology.org/2023.nllp-1.12.pdf
12. Sanchez J. et al. (2025). Monetary Sanction Extraction. EMNLP 2025.
13. Ivanenko O. (2025). Analysis of Ukrainian Court Decisions Structure. UNLP 2025.
14. Quinta de Castro P. V., Da Silva N. F. F. (2025). Labor Lex: Information Extraction from Brazilian Labor Law. NLLP 2025. https://aclanthology.org/2025.nllp-1.14.pdf
15. Panchenko E. et al. (2022). Claim and Regulation Extraction for COVID-19. LREC 2022. https://aclanthology.org/2022.lrec-1.50.pdf
16. Bhattacharyya A. et al. (2025). Information Extraction from Visually Rich Documents. ACL 2025. https://aclanthology.org/2025.acl-long.844.pdf
17. Jungiewicz M., Łopuszyński G. (2022). Keyword Extraction from Polish Legal Texts. Springer NLP.

18. Halama P. et al. (2022). Classification of Polish Court Judgments Using N-grams and Stemming. Archives of Acoustics.
19. IEEE Access. (2024). BiLSTM-CRF with Fuzzy Logic for Legal Metaphor Extraction.
20. Dragoni M. et al. (2023). Ontology-Based Extraction in Polish Legal Texts. Springer.
arXiv. (2025). Polish BERT for Legislative Analysis.
21. Wang Y. et al. (2024). Graph-Based Clinical Entity Extraction. ACL BioNLP.
22. Li H. et al. (2025). Zero-Shot Risk Classification in SEC Filings. NAACL Industry.
23. Campos R. et al. (2023). YAKE for Scientific Keyphrase Extraction. ECIR.
24. Zhang L. et al. (2024). Event Extraction from Social Media with T5. EMNLP.
25. Liu Z. et al. (2023). Dependency Parsing for Patent Claim Extraction. IP&M.
26. Codesignal. (2025). Legal Information Extraction with Text Processing Libraries. https://codesignal.com/learn/courses/practical-applications-of-text-processing-for-real-life-tasks/lessons/information-extraction-from-legal-documents
27. Osyvokon. (2025). Ukrainian Language Resources Catalog. https://github.com/osyvokon/ukrainian-language-resources
28. UNLP. (2025). Ukrainian Natural Language Processing Conference. https://unlp.org.ua/
29. IEEE Access. (2024). Decision Model Extraction Using BERT and BiLSTM-CRF.

КІБЕРБЕЗПЕКА В СУЧАСНОМУ СВІТІ: АКТУАЛЬНІ ВИКЛИКИ ТА ПРАКТИКА

МАТЕРІАЛИ
VI МІЖНАРОДНОЇ
НАУКОВО-ПРАКТИЧНОЇ
КОНФЕРЕНЦІЇ

28 листопада 2025 року
м. Одеса

Відповідальній редактор: О.В. Дикий

Дизайн та верстка:
М.Ю. Овчинников