УДК 340.113:004.8(043.2)

В.М. Синєглазов
(Національний авіаційний університет, Україна)
Д.О. Бондаренко
(НТУ «Київський політехнічний інститут імені Ігоря Сікорського», Україна)

Гібридна система на основі штучного інтелекту для класифікації кримінальних справ та аналізу доказів

У статті описано гібридну систему для класифікації кримінальних справ і аналізу доказів на основі штучного інтелекту. Система використовує моделі BERT, C4.5, LDA та SVM для підвищення ефективності автоматизації юридичного аналізу й пошуку аналогічних судових рішень.

Автоматизація юридичного аналізу відіграє важливу роль у забезпеченні справедливості та послідовності судових рішень. Використання сучасних технологій, зокрема систем штучного інтелекту для автоматичного пошуку схожих судових прецедентів, дозволяє значно прискорити процес прийняття рішень, знижуючи ризики людських помилок та забезпечуючи доступ до великого масиву юридичної інформації. Такий підхід підвищує якість аналізу справ і сприяє винесенню більш обґрунтованих і послідовних судових рішень. Загалом функціонал системи поділяється на 3 частини:

1) Виділення статей кримінального кодексу, до яких можна віднести справу.
2) Класифікація доказової бази за виділеними статтями, що дозволить прокурору вирішити чи достатньо цього для відкриття кримінального провадження
3) Пошук схожих судових рішень

Розглянемо кожне завдання системи більш детально.

Виділення статей кримінального кодексу

Одним із ключових завдань системи є автоматична класифікація кримінальних справ на основі відповідних статей законодавства, що можуть бути застосовані для висунення обвинувачень підозрюваним. Використання двоетапного підходу до класифікації дозволяє системі більш точно визначати перелік статей, які потенційно можуть бути застосовані до конкретної справи. Нижче наведено детальний опис процесу аналізу тексту на цьому етапі.

1) **Семантичне кодування:** На початковому етапі система здійснює семантичне кодування всіх текстових даних за допомогою мовної моделі BERT. Оскільки юридичні документи зазвичай характеризуються значною довжиною та складними формулюваннями, текст поділяється на параграфи, і кожен з них аналізується окремо. Модель BERT забезпечує розуміння контексту та значення слів у кожному параграфі, що дозволяє створити цілісне уявлення про зміст документа.

2) **Вилучення фактів:** На цьому етапі система витягує ключові факти з текстового файлу, застосовуючи класичні методи обробки природної мови (NLP), такі як токенізація, видалення стоп-слів, лематизація та стемінг. Додатково застосовується метод розпізнавання іменованих сутностей (NER) для ідентифікації ключових юридичних сутностей, таких як імена учасників (наприклад, жертва, обвинувачений), місця події, об'єкти (зброя, транспортні засоби), дії (удари, погрози), а також характеристик (тяжкість травм).
3) **Класифікація фактів за діагностичними питаннями:** Після вилучення фактів система використовує модульну штучну нейронну мережу (mANN) для їх класифікації на основі заздалегідь визначених діагностичних питань. Наприклад, система визначає "рівень насильства" або "тяжкість травм" на основі зібраних даних.

4) **Застосування юридичних елементів:** На наступному етапі система застосовує метод C4.5 (дерева рішень) для співставлення класифікованих діагностичних питань з юридичними елементами, що зберігаються в базі даних. Це дозволяє визначити, які статті закону можуть бути застосовані до справи. Наприклад, у випадку, якщо система ідентифікує високий рівень насильства, вона може рекомендувати більш серйозні звинувачення.

Класифікація доказової бази

Система буде мати здатність оцінювати достатність наявних доказів для порушення кримінальної справи, порівнюючи наявні факти з відповідними положеннями законодавства. Система групуватиме докази таким чином, щоб вони відповідали певній статті закону, на основі чого користувач зможе оцінити, чи достатньо доказової бази для порушення кримінальної справи за цією статтею. Оскільки система оперує обмеженим переліком статей, до яких може бути віднесена дана справа, завдання зводиться до класичної задачі класифікації. Для реалізації цього функціоналу необхідно використовувати розмічені дані у форматі «доказ – стаття», на основі яких можна навчити стандартні класифікаційні алгоритми, такі як SupportVectorMachines (SVM).

Пошук схожих судових рішень

Додатково, система також може бути використана для пошуку рішень у попередніх або схожих судових справах. Пошук аналогічних справ здійснюється за допомогою інтеграції кількох видів інформації: ключових термінів, знайдених у тексті, семантичного кодування тексту, а також тематичної інформації, яка міститься у документі. Комбінуючи ці три фактори, система формує комплексне представлення змісту тексту і використовує це представлення для порівняння енкодингів текстів з метою визначення ступеня їхньої подібності.

Процес виділення ключових термінів та семантичного кодування було розглянуто у попередніх пунктах, тому розглянемо більш детально етап тематичного кодування. Також до функціоналу проектованої системи відноситься генерація звітів, в яких надаються результати аналізу, включаючи

діагностичні питання, відповідні статті закону та запропоновані санкції. Це сприяє автоматизації процесу юридичного аналізу, роблячи його більш оперативним і менш схильним до суб'єктивних помилок.

Для навчання системи використовуються різні текстові документи, включаючи детальні описи інцидентів, результати розслідувань, свідчення та судові рішення. Перед тим як використовувати ці документи для навчання системи, вони повинні пройти через стандартні етапи попередньої обробки текстових даних:

- **Токенізація**: Поділ тексту на окремі слова або фрази (токени).
- **Видалення стоп-слів**: Видалення з тексту загальновживаних слів (наприклад, «і», «або», «в»), які не несуть значної інформації для аналізу.
- **Лематизація та стемінг**: Приведення слів до їх базових форм. Лематизація дозволяє звести всі форми слова до однієї базової форми (наприклад, "йшов", "йшла" — до "йти").
- **Розбивка на речення**: Текст розбивається на окремі речення для полегшення подальшого аналізу.

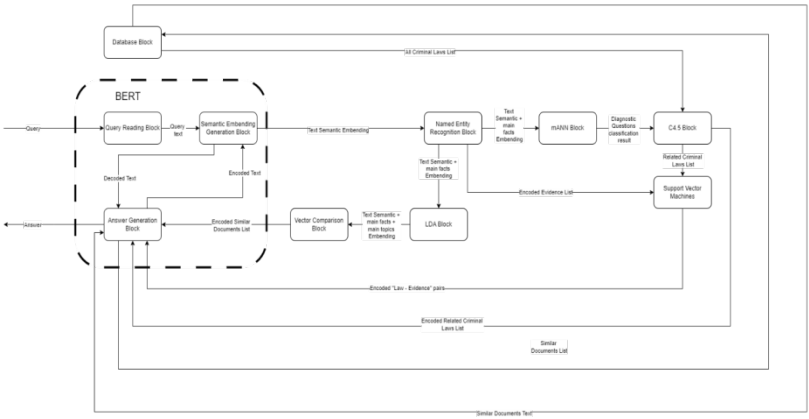

Рис.1 Архітектура системи

Система складається з кількох ключових блоків, кожен з яких виконує свою специфічну функцію в загальному процесі аналізу та класифікації кримінальних справ. Нижче наведено детальний опис функцій кожного блоку та їх взаємодії в рамках системи.

1) **BERT Block (Мовна модель)**: Цей блок є центральним елементом системи і відповідає за обробку та розуміння запитів користувачів. BERT (BidirectionalEncoderRepresentationsfromTransformers) використовується для семантичного кодування запитів та текстів документів. Він забезпечує глибоке розуміння контексту запиту та його відповідність наявній базі даних. Після аналізу запиту блок

BERT може генерувати відповіді на основі даних, отриманих від інших компонентів системи.

2) **NamedEntityRecognitionBlock (NER Block)**: Цей блок відповідає за розпізнавання та виділення ключових фактів з тексту, таких як імена учасників, місця подій, об'єкти злочинів та інші важливі сутності. NER використовується для створення структурованого представлення неструктурованих текстів, що дозволяє краще зрозуміти зміст документів і підготувати їх для подальшого аналізу в інших блоках. Результати роботи цього блоку передаються до наступних етапів класифікації.
3) **mANNBlock (Модульна штучна нейронна мережа)**: Після виділення ключових фактів, вони передаються до блоку mANN, який відповідає за класифікацію цих фактів згідно з діагностичними питаннями. Цей блок використовує штучну нейронну мережу для визначення того, які категорії діагностичних питань відповідають конкретним фактам, таким як "рівень насильства", "тяжкість травм", "наявність зброї" та інші. Це дозволяє системі краще зрозуміти контекст справи та підготувати її для наступного етапу класифікації.
4) **C4.5 Block (Дерева рішень C4.5)**: Після класифікації фактів за діагностичними питаннями, блок C4.5 відповідає за класифікацію кримінальних справ відповідно до статей кримінального кодексу. Метод дерев рішень C4.5 забезпечує інтерпретованість результатів класифікації, що важливо для правового аналізу. Він зіставляє отримані факти з правовими нормами та визначає, які статті закону можуть бути застосовані в даній справі.
5) **SupportVectorMachines (SVM Block)**: Цей блок відповідає за класифікацію доказів, виявлених під час аналізу текстів, відповідно до відповідних статей закону. SupportVectorMachines (SVM) використовуються для побудови класифікатора, який може зіставляти докази з певними правовими статтями на основі попередньо розмічених даних.
6) **LDA Block (LatentDirichletAllocation)**: LDA використовується для тематичного моделювання тексту, що дозволяє визначати основні теми, які обговорюються в документах. Цей блок класифікує текст на кілька тем, кожна з яких складається з набору пов'язаних понять. Тематичне моделювання дозволяє краще зрозуміти зміст справи, виявляючи ключові теми, які можуть бути важливими для подальшого аналізу та пошуку аналогічних справ.
7) **VectorComparisonBlock (Блок порівняння векторів)**: Після тематичного та семантичного кодування текстів документи представляються у вигляді векторів. Блок порівняння векторів відповідає за порівняння текстів на основі отриманих векторних представлень. Він визначає ступінь схожості між документами і створює список релевантних справ для подальшого аналізу. Це дозволяє системі знаходити аналогічні справи на основі змістовного

аналізу документів, навіть якщо термінологія або формулювання відрізняються.

8) **DatabaseBlock (Блок бази даних)**: Блок бази даних зберігає всі документи, включаючи статті кримінального кодексу та тексти попередніх справ. Він взаємодіє з іншими блоками, забезпечуючи доступ до необхідної інформації для порівняння та класифікації. Документи, знайдені як релевантні під час пошуку, витягуються з бази даних і використовуються для подальшого аналізу або створення звітів.

Висновки

Проектована гібридна система штучного інтелекту має на меті автоматизацію процесів класифікації кримінальних справ, аналізу доказів та пошуку аналогічних судових рішень, що значно покращить ефективність юридичного аналізу. Однією з основних переваг системи є її здатність поєднувати різноманітні методи, такі як семантичне кодування через BERT, тематичне моделювання за допомогою LDA, а також точну класифікацію фактів і доказів за допомогою алгоритмів C4.5 і SVM. Це забезпечить системі гнучкість і точність у роботі з великими масивами неструктурованих текстів, що є ключовим аспектом у роботі з кримінальними справами.Особливістю системи є її багаторівневий підхід до обробки інформації, що дозволяє проводити комплексний аналіз текстових даних з урахуванням як семантичного, так і тематичного контексту.У перспективі ця система здатна значно знизити ризик людських помилок, оптимізувати процес підготовки до судових розглядів і покращити доступ до релевантних юридичних прецедентів, що робить її інноваційним інструментом для роботи прокурорів та інших правозахисних органів.

Список літератури

1) Weifeng H. BERT_LF: A Similar Case Retrieval Method Based on Legal Facts [Електроннийресурс] / H. Weifeng, Z. Siwen, Z. Qiang // Wireless Communications and Mobile Computing. – 2022. – Режимдоступудоресурсу: https://doi.org/10.1155/2022/2511147.

2) Ezdihar B. Text mining and machine learning for crime classification: using unstructured narrative court documents in police academic [Електроннийресурс] / B. Ezdihar, B. Arwa, M. Rsha // Cogent Engineering. – 2024. – Режимдоступудоресурсу: https://doi.org/10.1080/23311916.2024.235985.

3) Sotarat T. A Framework of Multi-Stage Classifier for Identifying Criminal Law Sentences [Електроннийресурс] / T. Sotarat, W. Bunthit, C. Nipon // Procedia Computer Science 13. – 2012. – Режимдоступудоресурсу: https://www.sciencedirect.com/science/article/pii/S187705091200720X.