РОЗДІЛ 4. ТЕОРЕТИКО-ПРИКЛАДНІ, ПОРІВНЯЛЬНІ, ФІЛОСОФСЬКІ ТА ІСТОРИЧНІ ЗАСАДИ ПРАВОГО РЕГУЛЮВАННЯ

Ксенія ІВАНОВА
ORCID: 0000-0003-4696-2478

TEXT AND DATA MINING ЯК КЛЮЧОВИЙ ІНСТРУМЕНТ ДЛЯ МАШИННОГО НАВЧАННЯ: ЄВРОПЕЙСЬКИЙ ДОСВІД ПРАВОВОГО РЕГУЛЮВАННЯ

УДК 347.7

Постановка проблеми. Сучасне суспільство знаходиться на тому етапі розвитку, коли всі сфери людської діяльності зазнають впливу цифровізації. Причому це не тільки технологічний тренд, але й той чинник, який модифікує традиційні правові інститути та формує нові моделі правовідносин. Так, Інтернет, відкритий характер цифрового простору та нова – цифрова форма для об'єктів авторського права змінюють способи створення й розповсюдження творів, а активний розвиток систем штучного інтелекту (artificial intelligence, далі – AI) та алгоритмів, які здатні автоматизовано і непомітно збирати та обробляти дані з усіх можливих цифрових джерел для його навчання та вдосконалення, відкрили можливості для несанкціонованого втручання в легальну монополію правоволоділців. Безумовно, такі дії не зустрічають схвалення з боку творців, оскільки позбавляють їх справедливої винагороди за працю та розмивають межі виключних прав. З огляду на це, функціонування AI може нести в собі загрозу посилення напруженості між правоволоділцями й користувачами та порушення того балансу, який протягом десятиліть вибудовувало авторське право, намагаючись узгодити інтереси творців і суспільства.

У технологічному аспекті якість, точність та ефективність AI значною мірою визначаються алгоритмами та методами Text and

Data Mining (далі – TDM), які призначені для пошуку у великих обсягах цифрових джерел, а саме в текстах, даних, звуках, зображеннях або їх комбінаціях, неочевидних і корисних на практиці кореляцій та інших закономірностей. Безумовно, єдиної загальноприйнятої дефініції TDM не існує і в літературі можна зустріти різні визначення TDM – як процесу, технології, методу, алгоритму, інструменту тощо. Навіть самі дослідники використовують різні позначення (найменування) для TDM – майнінг даних (data mining), інтелектуальний аналіз тексту та даних, глибинний аналіз даних тощо.

Незважаючи на те, що TDM залежно від типу даних, які він опрацьовує, – структуровані або неструктуровані[1], може бути представлений як поєднання дата-майнінгу (data mining) та текст-майнінгу (text mining), головним залишається процес – операції з витягування (отримання) даних – майнінг. Сам термін «майнінг» був обраний недарма: метафора прийшла з англійського «mining», що в перекладі означає «видобуток руди». Як відмічають вчені, майнінг даних дуже схожий на процес видобутку золота та дорогоцінних металів, прихованих у надрах, оскільки знання – це щось дорогоцінне, що шукають та досліджують у великих блоках даних [2, с. 18]. Іншими словами, майнінг – це наочний термін, що характеризує процес знаходження невеликої кількості цінних «самородків» із великої кількості сирого матеріалу» [3, с. 6].

1 В контексті TDM розрізняють такі типи даних: структуровані (табличні) та неструктуровані (нетабличні). У той час як структуровані дані мають фіксовану структуру подання та організовані в табличній формі (наприклад, бази даних, електронні таблиці), що забезпечує їх пряму обробку, неструктуровані, навпаки, цього позбавлені. До неструктурованих даних відносять текстові документи, зображення, аудіо- та відеофайли, автоматична обробка та інтерпретація яких комп'ютерними системами суттєво ускладнена (наприклад, для обробки текстових файлів текст має бути перетворений у форму, придатну для автоматизованого аналізу комп'ютерною системою). Text mining фокусується виключно на аналізі текстових нетабличних даних, а data mining, відповідно, охоплює аналіз всіх інших даних – як структурованих таблиць і баз даних, так і інших неструктурованих об'єктів (зображень, аудіозаписів, відеозаписів тощо). Більш докладно див.: [1].

При цьому, попри відмінності дослідників у підходах, усі вони погоджуються з тим, що TDM – це автоматизована обробка даних з метою виявлення нових знань або прихованих закономірностей. TDM виступає сполучною ланкою між інформацією (даними), вилученої з цифрових джерел, та алгоритмічними процесами машинного навчання. Машинне навчання, AI та TDM становлять єдину технологічну систему: спочатку в результаті TDM зібрані дані автоматично структуруються в набори даних, на основі яких відбувається безпосереднє навчання нейронних мереж (машинне навчання); після етапу навчання модель AI вже самостійно може надавати конкретні результати за запитом (завданням) користувача (прогнозування, рекомендації, створення контенту тощо). Таким чином, TDM забезпечує обробку, підготовку та поєднання даних у набори, які придатні для аналізу; машинне навчання – їх аналітичне опрацювання й виявлення закономірностей, а AI – застосування отриманих знань на практиці. В силу органічної взаємопов'язаності цих процесів, неврегульованість у правовій площині TDM, як базового елементу, не створювало визначеності при використанні й інших елементів. Ця проблема країнами ЄС була вирішена ухваленням 17 квітня 2019 р. Директиви ЄС 2019/790 про авторське право і суміжні права на Єдиному цифровому ринку [4], також відомої як DSM Directive (далі – Директива). Вона доповнила раніше прийняті Директиви 96/9/ЄС та 2001/29/ЄС і в силу своєї правової природи вимагала від держав-членів ЄС імплементації її положень у національне законодавство.

Метою статті є дослідження теоретичних та практичних аспектів застосування положень щодо TDM, запроваджених Директивою, зокрема, наукового осмислення умов та меж, за яких його використання кваліфікується як правомірне і не порушує прав авторів (правоволоділців).

На відміну від країн ЄС, чинне українське законодавство взагалі не містить спеціальних норм щодо TDM, тим самим підкорюючи загальним положенням щодо вільного використання об'єктів авторського права. Однак такий підхід не відповідає зобов'язанням, які взяла на себе Україна в рамках Угоди про асоціацію між

Україною та ЄС [5] щодо адаптації національного законодавства до європейських стандартів у сфері авторського права. На сьогодні дослідження європейського підходу для розробки й впровадження аналогічної національної моделі правового регулювання TDM стає нагальною.

Аналіз останніх досліджень і публікацій. Актуальність і глибинний характер проблем, що породжує функціонування AI, привертає увагу всієї наукової спільноти, хоча природно, що більш широке висвітлення механізм правового регулювання TDM отримав саме в працях зарубіжних дослідників. В Україні правові аспекти машинного навчання в контексті TDM практично не досліджувались, хоча такі науковці, як С.О. Глотов, К.О. Зеров, Ю.М. Капіца, О.Е. Сімсон, І.Є. Якубівський, Є.О. Харитонов, О.І. Харитонова, К.О. Маслова-Юрченко та Т.О. Музика побічно торкались цієї проблематики, яка в сучасних умовах містить багато невирішених проблем.

Виклад основного матеріалу. Відповідно до Директиви TDM – інтелектуальний аналіз тексту та даних[1], означає будь-який автоматизований аналітичний метод, призначений для аналізу тексту та даних у цифровій формі з метою отримання інформації, яка охоплює, серед іншого, моделі, тенденції та кореляції (ст. 2(2) Директиви). На перший погляд, процес TDM, за визначенням Директиви, зав'язаний саме на *даних, інформації* у цифровій формі, що не є об'єктами авторського права, оскільки цей інститут охороняє не дані як такі (зокрема, текст), а оригінальну форму вираження. Між тим саме об'єкти авторського права в цифровій формі – тексти творів, зображення, інформація з баз даних тощо, є тим матеріалом (ресурсом), який піддається глибинному аналізу, щоб отримати ці дані. Саме тому слід погодитися з Е. Розаті (*Eleonora Rosati*), яка підкреслює, що «цінність даних полягає не в самих даних або тексті, взятих окремо, а у витягуванні цінності» [7, с. 2].

1 Зауважимо, що хоча в офіційному українському перекладі Директиви [6] вживається термін «глибинний аналіз тексту та даних», який в цьому контексті має бути еквівалентом загальновживаного терміну «інтелектуальний аналіз тексту та даних», перший не є усталеним в ІТ-сфері і може викликати хибні асоціації з «глибинним навчанням» (deep learning).

Враховуючи технічну складність процесів TDM, вчені, як правило, виокремлюють такі його етапи: (1) доступ до контенту (текстових чи інших даних); 2) вилучення та/або копіювання контенту; 3) аналіз тексту та/або даних з метою виявлення нових знань [8]. На етапі вилучення та/або копіювання контенту виділяють попередню обробку даних (перетворення даних на машиночитані формати, а також очищення даних шляхом видалення нерелевантних, надлишкових, неякісних) та перетворення даних (консолідацію та їх агрегацію) для опрацювання алгоритмами інтелектуального аналізу даних [9, с. 48]. З огляду на це, питання, чи відбувається використання об'єктів авторського права під час процесу TDM, можна вважати вичерпаним, оскільки вилучення та копіювання є діями, що вкладаються в поняття «використання» об'єкту права інтелектуальної власності і без згоди правоволодільця означають порушення його прав[1].

Проте Директива закріпила винятки з легальної монополії правоволодільців і передбачила дві законні підстави для використання TDM: (1) науково-дослідними організаціями та установами зі збереження культурної спадщини з метою проведенням наукових досліджень та (2) будь-якими іншими особами, в тому числі і з комерційною метою (статті 3 та 4).

Означені в Директиві винятки вимагають чіткого дотримання певних умов, що виявляються не тільки у суб'єктному складі осіб, які застосовують TDM, та меті, для якої це провадиться, а й інших аспектах використання охоронюваних об'єктів: способу використання, режиму доступу до об'єктів, строку зберігання матеріалу.

Як було зазначено, перша законна підстава стосується використання TDM науково-дослідними організаціями та установами зі збереження культурної спадщини з метою проведенням наукових досліджень (стаття 3 Директиви). Отже, щоб підпасти під дію вказаного винятку, використання має відбуватись спеціальним

1 Між тим Директивою передбачені випадки, коли при TDM хоча й має місце відтворення, але відтворені дані підпадають під обов'язковий виняток щодо тимчасових актів відтворення, передбачений статтею 5(1) Директиви 2001/29/ЄС (п.9 Преамбули) і повинен застосовуватися до TDM, якщо не передбачається виготовлення копій за межами сфери використання такого винятку.

суб'єктом – науково-дослідними організаціями та установами зі збереження культурної спадщини. За визначеннями, наведеними у ст. 2 Директиви, до «установ зі збереження культурної спадщини» належать загальнодоступні бібліотека, музей, архів або установа зі збереження кінематографічної чи аудіо спадщини (ст. 2(3)), а до «науково-дослідних організацій» – університет, в тому числі його бібліотеки, науково-дослідний інститут або будь-яка інша установа, основною метою якої є проведення наукових досліджень або здійснення освітньої діяльності, що включає також проведення наукових досліджень: на неприбутковій основі або шляхом повторного інвестування всього прибутку у свої наукові дослідження; або відповідно до місії, яка становить суспільний інтерес, визнаної державою-членом; у такий спосіб, що суб'єкт, який має вирішальний вплив на таку організацію, не може користуватися преференційним доступом до результатів, отриманих у ході таких наукових досліджень (ст. 2(1)).

Таким чином, не можуть послуговуватись цим винятком ті суб'єкти, які використовують TDM з комерційною метою, а організації, над якими здійснюють контроль комерційні структури, не можуть вважатися науково-дослідними. Так, зі сфери застосування ст. 3 виключені суспільні організації мовлення та комерційні дослідницькі інститути, проте вони можуть підпадати під дію іншого винятку, передбаченого у ст. 4 Директиви [10]. Між тим, якщо університети та науково-дослідні інститути в межах публічно-приватних партнерств співпрацюють з приватним сектором і покладаються на своїх приватних партнерів у застосуванні технологій TDM, у тому числі шляхом використання їхніх технологічних засобів, виняток статті 3 Директиви на них також розповсюджується.

Положення ст. 3 Директиви дозволяють науково-дослідними організаціям та установам використовувати технології TDM лише до об'єктів (контенту), до яких вони мають «законний доступ». Відповідно до п. 14 Преамбули Директиви, це об'єкти, що знаходяться у вільному онлайновому доступі (на основі політики відкритого доступу), та доступ до яких базується на договірних засадах з правоволоділець (наприклад, доступ за підпискою). При цьому контент має використовуватись лише з єдиною метою

– проведення наукових досліджень, якими охоплюються природничі та гуманітарні науки (п. 12 Преамбули). При цьому копії творів, інших об'єктів, які піддавались TDM, дозволяється зберігати з метою наукових досліджень, у тому числі для перевірки результатів досліджень (ст. 3(2)). Емпіричні наукові дослідження, як правило, вимагають, щоб дослідницькі дані залишались у доступі з метою їх підтвердження [10].

Слід зауважити, що здійснення вказаними організаціями TDM відносно захищеного контенту буде вважатись правомірним навіть у разі, якщо правоволоділець прямо заборонив використання TDM щодо результатів своєї інтелектуальної, творчої діяльності. Це зумовлено тим, що практика встановлення подібних заборон з боку правоволодільців буде нести певні загрози для розвитку науки і культури, що в підсумку впливатиме на все суспільство, тоді як потенційна шкода правоволодільцям через встановлення в законі винятку в інтересах дослідницьких та подібних їм організацій буде мінімальною. Тому відповідно до положень статей 3 та 7 Директиви, встановлення заборон правоволодільцями не матимуть обмежувального ефекту, якщо буде йтись про некомерційну діяльність у сфері освіти, науки і культури тих організацій та установ, які через наукові пошуки та дослідження мають на меті примножити знання та привнести щось нове в певну галузь. У такий спосіб забезпечується рівновага приватних інтересів правоволодільців та публічних інтересів, що є виправданим і навіть необхідним для критично важливих сфер життя людини, таких як медицина, охорона здоров'я, освіта.

З іншого боку, ст. 3(3) Директиви для убезпечення цілісності мереж та баз даних, у яких розміщені твори чи інші об'єкти, дозволяє правоволодільцям вживати певних заходів. І хоча межі застосування останніх вичерпуються означеною метою, не виключається, що на практиці це може здійснюватися в обхід імперативних приписів статті. Прикладом є API (application programming interface – інтерфейси прикладного програмування), що використовуються для взаємодії між різними програмами за допомогою запитів, які, за правило, визначають певні умови доступу, щоб убезпечити систему та дані від дій зловмисників.

Традиційно API використовується для недопущення неавторизованого доступу до програми (даних), перешкоджання DDoS-атакам тощо, а отже, він здатен істотно обмежити дії користувачів. Вчені відмічають, що технічно досить просто розробити API, який допускає лише певну кількість запитів або певну їх довжину, складність чи певну функцію пошуку. Слід погодитися з тим, що важко, або навіть взагалі неможливо спрогнозувати, коли ці обмеження перейдуть червону межу між заходами безпеки, які передбачені статтею 3(3) Директиви, і стануть формою (забороненого) обмеження прав, встановлених статтею [11, с. 127]. Хоча подібні дії з використанням API і маскуються під правомірні, але по своїй суті – це зловживання правом.

Друга законна підстава для TDM – виняток, передбачений статтею 4 Директиви, дозволяє всім іншим суб'єктам, що не охоплені статтею 3, використовувати TDM, в тому числі, і з комерційною метою. Залишаючи обов'язкову вимогу щодо законності доступу користувачів до об'єкту, положення ст. 4 Директиви закріплюють механізм «відмови від TDM», що також відомий як «opt-out» (інші позначення – «застереження про відмову», «обмеження», – тут і далі будуть використовуватись як синонімічні), який дозволяє правоволодільцю заборонити використання своїх творів як вихідних даних для навчання AI. Разом із тим відмова від TDM має бути встановлена «у належний спосіб»: наприклад, для онлайн контенту Директива орієнтує на засоби машинного зчитування, які згідно з п. 18 Преамбули, можуть включати метадані та положення й умови веб-сайту або сервісу.

З огляду на наведене, в Директиві запроваджено загальне правило – «TDM дозволяється», але для науково-дослідних організації цей загальний дозвіл діятиме незалежно від волі правоволодільця, а для інших суб'єктів – діятиме, якщо правоволоділець не встановить відповідну відмову від TDM щодо свого контенту. Таким чином, можна спостерігати, що сучасні технології зміщують акценти в авторському праві: пасивна поведінка правоволодільця, яка обумовлюється існуванням загальної законодавчої заборони будь-якого використання об'єкту без його дозволу, в контексті TDM змінюється на активну – щоб заборонити будь-яке

використання, необхідно про це прямо заявити. Тим самим положення ст. 4 Директиви передають сферу використання об'єктів під контроль правоволодільця, який має опікуватись своїми правами і обирати «належний» спосіб застереження.

В умовах сьогодення правоволодільці мають можливість убезпечити свій контент від посягань за допомогою широкого спектру засобів технічного характеру, серед яких найбільш поширені файли robots.txt, tdm.txt або ai.txt (як сучасний аналог robots.txt)[1], метатеги, API с контролем доступу тощо; та правового характеру – ліцензій. Однак відсутність єдиних загальновизнаних стандартів допустимих способів відмови[2], як зазначає С. Хавлікова (*Stepanka Havlikova*), «може бути ахіллесовою п'ятою винятку TDM» [13].

Очевидно, що узагальнюючі формулювання Директиви 2019/790 відносно механізму відмови не вносять ясність у розуміння як допустимих способів повідомлення про це, так і тих суб'єктів, які управомочені їх використовувати (якщо ними можуть бути інші особи, окрім правоволодільця).

Довгий час вважалось, що через технічну природу алгоритмів, які здійснюють в TDM аналіз тексту та даних, застереження про відмову повинно мати «машиночитаний» формат, тобто такий, що зможе «прочитати» машина. І хоча це загальне правило, але воно може коригуватись залежно від рівня технічного розвитку. Так, у справі *Robert Kneschke v. LAION e.V.* [14], в якому застереження було викладено на веб-сайті англійською мовою[3], суд витлумачив «машиночитаний» як «машинозрозумілий», тим самим розширив межі зрозумілих для «машини» мов за рахунок людських (природних) якостей. Суд був переконаний, що найсу-

1 Більш докладно про це див. : [12].

2 Коли Директива була прийнята, політики та зацікавлені сторони очікували, що такі стандарти поступово з'являться в галузевій практиці. Під час прийняття Директиви вчені розглядали протокол robots.txt як можливий вектор для вираження відмови або як приклад майбутнього стандарту [12].

3 Застереження мало наступний вигляд: «You may not use automated programs, applets, bots or the like to access the Bigstock.com website or any content thereon for any purpose, including, by way of example only, downloading Content, indexing, scraping or caching any content on the website.»

часніші технології, які використовуються для TDM, мають можливості обробляти природну мову і розуміти зміст наведених природною мовою текстів. Пославшись на Закон про штучний інтелект (EU Artificial Intelligence Act [15]), Суд нагадав про обов'язок постачальників моделей AI загального призначення виявляти та дотримуватись, в тому числі за допомогою самих сучасних технологій, застереження про права, виражені відповідно до ст. 4(3) Директиви (ст. 53(1)(c) зазначеного Закону). І хоча порушення прав, за захистом яких звернувся *Robert Kneschke*, мало місце у 2021 році, Суд припустив, що на той час відповідачу (LAION e.V.) вже були доступні такі технології.

Звертає на себе увагу і той факт, що в положеннях ст. 4 Директиви йдеться про встановлення обмеження «правоволоділь-цем», тобто при вузькому тлумаченні це мало б зводитись, по суті, виключно до встановлення обмеження автором. Такий підхід нівелював би правовий ефект тих самих дій, що вчиняються іншими особами в інтересах автора (наприклад, ліцензіатом або організаціями колективного управління). Тому імплементація положень Директиви до національного законодавства держав-членів ЄС припускала зважений підхід до цієї проблеми, не обмежуючи коло таких осіб лише авторами. Так, у згаданій справі *Robert Kneschke v. LAION e.V.* фотограф *Robert Kneschke* розмістив свої роботи на веб-сайті фотостокового агентства на умовах невиключної ліцензії, а застереження було розміщено оператором веб-сайту. Суд, аналізуючи положення Urheberechtsgesetze (UrhG – Закон про авторське право і суміжні права ФРН [16]), а саме ст. 44b(1), яка імплементує положення ст. 4 Директиви, визнав таке застереження прийнятним, оскільки до уваги мають братися не тільки застереження самого автора, але й його правонаступників або володільців прав, отриманих від автора.

Інший бік проблеми – це ті застереження, які заявлені правоволоділь-цями. Вони, по-перше, повинні поважатись користувачами технологій TDM, а, по-друге, користувачі повинні дотримуватись цих застережень. Між тим на практиці вони не завжди дотримуються цих правил, тому невипадково в літературі моделюються можливі ситуації, коли юристи радять розробникам

комерційного AI «вивчити все, а потім знищити навчальний матеріал». В такому випадку складно провести зворотну розробку навченої моделі, повернутися до навчального матеріалу і довести порушення прав [11, с. 125]. До того ж доказування порушення прав може ускладнюватись й тим, що на перший погляд дії щодо знищення даних, навчального матеріалу (копії творів, інших об'єктів, які піддавались TDM) вкладатимуться в канву ст. 4 Директиви, яка для комерційного використання даних обмежує строк їх зберігання часом, «необхідним для інтелектуального аналізу тексту та даних» (нагадаємо, що для некомерційного використання TDM такої жорсткої вимоги не існує).

Перешкодити таким недобросовісним практикам і забезпечити прозорість даних, що використовуються для навчання моделей AI загального призначення (General Purpose AI, GPAI), покликані положення статті 53(1)(d) EU Artificial Intelligence Act, які набудуть чинності 2 серпня 2025 р. Вони вимагають від постачальників таких моделей скласти та опублікувати докладний звіт щодо контенту, використаного для навчання моделі, відповідно до наданого Офісом AI (AI Office) шаблону. Такий звіт, згідно з п.107 Преамбули, повинен охоплювати як фазу попереднього навчання, так і основне тренування з особливим акцентом на контент, захищений авторським правом.

Аналіз положень статей 3 та 4 Директиви доводить наступне: 1) незважаючи на чіткі формулювання, питання щодо їх змісту вимагають відповідей не тільки на рівні доктрини, але й судової практики, яка може дати оцінку фактичному впливу вказаних положень на відносини між інноваторами, що широко застосовують TDM, та правовласниками. Як важливі елементи правової екосистеми, судова практика і доктрина невід'ємно пов'язані: в умовах правової невизначеності судові рішення допомагають витлумачити законні межі використання об'єктів авторського права і тим самим створюють простір для подальших наукових роздумів.

Справа *Robert Kneschke v. LAION e.V.*, яка стосується AI і машинного навчання, сколихнула світ як перша з тих, по якій суд

висловив свою позицію[1] і дав відповіді на релевантні питання. Вона цінна й тим, що породила нові дискусії як щодо тлумачення поняття «наукові дослідження», так і щодо суб'єктів, які підпадають під дію ст. 3 Директиви.

У вказаній справі позов був поданий фотографом *Robert Kneschke* щодо порушення його авторських прав на фотографію, розміщену за ліцензією на веб-сайті фотостоку, яку компанія *LAION* взяла з сайту і включила до набору даних для використання у навчанні AI. *LAION (Large-scale Artificial Intelligence Open Network)* – німецька некомерційна організація, що створює набори даних для AI, на яких в подальшому навчається ряд відомих моделей AI, що перетворюють текстовий запит в зображення (зокрема, Stable Diffusion та Imagen). *LAION* створила набір даних для пар «зображення – текст», і саме в цій частині її дії підпадають під визначення TDM. В набір даних була включена і фотографія позивача (*Robert Kneschke*), незважаючи на розміщену на сайті відмову від TDM.

Суд мав вирішити, чи були порушені права фотографа, зокрема, чи відбулось відтворення (копіювання) його фотографії при створенні набору даних. По суті йшлось про кваліфікацію відносин як таких, що підпадають під дію винятку ст. 3 або ст. 4 Директиви (в Законі ФРН UrhG § 60d імплементує статтю 3, а § 44b – статтю 4 [16]).

Ключовим фактом для Суду в цій справі стало те, що LAION – некомерційна організація, яка розмістила створений набір даних у вільному доступі. З огляду на ці факти Суд застосував виняток статті 3 Директиви: на думку Суду, хоча LAION не є типовою науково-дослідною організацією, вона вносить вклад в наукові дослідження – створює набори даних, які можуть бути основою для навчання AI. При цьому Суд зауважив: «наукові дослідження» не слід розуміти вузько, як такі, що охоплюють лише етапи роботи, безпосередньо пов'язані із набуттям знань; достатньо, щоб цей етап робіт був спрямований на набуття знань (пізніше), як напри-

1 Інші аналогічні справи, які стосуються використання об'єктів авторського права як матеріалу (контенту) для навчання AI, й досі перебувають на розгляді в судах (зокрема, справа Getty images v. Stability AI).

клад, у випадку збору даних, який спочатку треба провести, щоб пізніше зробити емпіричні висновки. Зокрема, термін «наукове дослідження» не передбачає подальшого дослідницького успіху, тому створення набору даних, який може стати основою для навчання систем AI, безумовно, можна розглядати як наукове дослідження. Створення набору даних – це фундаментальний крок, спрямований на його використання іншими дослідниками галузі штучних нейронних мереж для відповідних досліджень [14].

Разом із тим, хоча позивач і наполягав на тому, що дослідження мали комерційне спрямування, оскільки в подальшому набори даних використовувались комерційними організаціями (зокрема, Stability.AI), Суд виходив з того, що набір даних був переданий у вільний доступ для всього суспільства, а не тільки для комерційних організацій. Тісні зв'язки з останніми та отримання від них підтримки (в тому числі фінансової) також не переконали Суд в комерційному спрямуванні дослідження, оскільки не був доведений їх вирішальний вплив на LAION та надання їм переваг у доступі до результатів наукових досліджень, як того вимагає § 60d(2) UrhG.

Рішення суду у справі *Robert Kneschke v. LAION e.V.* (до речі, *Robert Kneschke* подав апеляцію на це рішення) наочно демонструє межі застосування на практиці положень Директиви щодо TDM і відповідних їй норм національного законодавства. Між тим, з точки зору доктрини, є й спірні моменти (окрім зазначених раніше). Так, наприклад, Суд витлумачив концепцію «законного доступу» (яка є вимогою щодо винятків TDM) як синонім «публічного доступу», хоча ці поняття суттєво різняться: наприклад, розміщення зображення на веб-сайті для загального перегляду робить його загальнодоступним, але не обов'язково також і «законно доступним», враховуючи, що особа, яка його розмістила, могла зробити це і без згоди відповідного правоволодільця [17, с. 852].

Крім того, аналіз положень Директиви 2019/790 дозволяє констатувати, що процес TDM має чіткі часові межі і триває до тих пір, поки відбувається аналітичний процес обробки даних, а звідси онлайн публікація набору даних, навіть якщо вона і здійс-

нена з загальнокорисною суспільною метою, унеможливлює кваліфікацію таких дій як правомірних.

Висновки. TDM відіграє важливу роль у процесі навчання AI, постачаючи великі обсяги текстових та числових даних, які необхідні для тренування моделей AI. Разом із тим аналіз правового регулювання винятків щодо TDM свідчить про незахищеність і вразливість авторів (правоволоділців) при застосуванні зазначених винятків на практиці. Незважаючи на те, що Директива 2019/790 чітко окреслила нові способи вільного використання захищених авторським правом цифрових джерел даних, їх практичне застосування вимагає більш детальної регламентації вимог, яким мають відповідати суб'єкти та їх науково-дослідницька та освітня діяльність, щоб підпадати під дію положень статті 3 Директиви і не допускати вільного тлумачення їх судом. Рішення суду у справі Robert Kneschke v. LAION e.V. виявило прогалини у правовому регулюванні і визначило вектор подальшого вдосконалення європейського законодавства з урахуванням інтересів всіх заінтересованих сторін.

1. Hassani H., Beneki C., Unger S., Mazinani M. T., & Yeganegi M. R. Text Mining in Big Data Analytics. Big Data and Cognitive Computing, 2020. Vol. 4. No.1. DOI : https://doi.org/10.3390/bdcc4010001

2. Rajab Asaad, R., & Masoud Abdulhakim, R. The Concept of Data Mining and Knowledge Extraction Techniques. Qubahan Academic Journal, 2021. Vol. 1. Iss: 2. pp. 17–20. DOI: https://doi.org/10.48161/qaj.v1n2a43

3. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques, 3rd edition. Morgan Kaufmann, 2011. 744 p.

4. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Text with EEA relevance.). URL : https://eur-lex.europa.eu/eli/dir/2019/790/oj

5. Угода про асоціацію між Україною, з однієї сторони, та Європейським Союзом, Європейським Співтовариством з атомної енергії та їхніми державами-членами, з іншої сторони, від 21 березня та 27 червня 2014 року. URL : https://zakon.rada.gov.ua/laws/show/984%20011/paran2820#n2820

6. Директива ЄС 2019/790 про авторське право і суміжні права на Єдиному цифровому ринку та про внесення змін до директив 96/9/ЄС та

2001/29/ЄС від 17.04.2019 № 2019/790. URL : https://zakon.rada.gov.ua/laws/show/984_022-19#Text

7. *Eleonora Rosati. The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market-Technical Aspects. URL: https://www.europarl.europa.eu/RegData/etudes/BRIE/2018/604942/IPOL_BRI(2018)604942_EN.pdf*
8. *Eleonora Rosati. Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and its Role in the Development of AI Creativity. Asia Pacific Law Review. 2019. Vol. 27. No. 2. pp. 198-217. URL : https://ssrn.com/abstract=3452376*
9. *Artha Dermawan. Text and data mining exceptions in the development of generative AI models: what the EU member states could learn from the Japanese "nonenjoyment" purposes?'. Journal of World Intellectual Property. 2024. Vol. 27, 1. pp. 44-68. DOI: https://doi.org/10.1111/jwip.12285*
10. *Hugenholtz P.B. The New Copyright Directive: Text and Data Mining (Articles 3 and 4). URL : https://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/*
11. *Kretschmer M., Margoni T., Oruç P. Copyright Law and the Lifecycle of Machine Learning Models. 2024. IIC-International Review of Intellectual Property and Competition Law. Vol. 55 (1). pp. 110-138. URL : https://link.springer.com/article/10.1007/s40319-023-01419-3*
12. *Keller P., Warso Z. Defining Best Practices for Opting out of ML Training (Open Future Foundation 2023). URL : https://openfuture.eu/wp-content/uploads/2023/09/Best-_practices_for_optout_ML_training.pdf*
13. *Stepanka Havlikova. Web Scraping and Text and Data Mining Exception: Could the CDSM Directive, Designed to Support the Reuse of Publicly Available Data, Have Had the Opposite Effect? URL : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4605551*
14. *LG Hamburg, Urteil vom 27.09.2024 - 310 O 227/23. URL : https://openjur.de/u/2495651.html*
15. *AI Act (Regulation (EU) 2024/1689. URL : https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689*
16. *Urheberrechtsgesetz. Gesetz vom 09.09.1965. URL : https://dejure.org/gesetze/UrhG*
17. *Eleonora Rosati, Author Notes. Is text and data mining synonymous with AI training? Journal of Intellectual Property Law & Practice, 2024, Vol. 19, No.12, p. 851-852. DOI: https://doi.org/10.1093/jiplp/jpae092*

Іванова К.Ю. Text and Data Mining як ключовий інструмент для машинного навчання: європейський досвід правового регулювання

Стаття присвячена дослідженню теоретичних та практичних аспектів застосування інтелектуального аналізу тексту та даних (TDM).

Відмічається, що у технологічному аспекті якість, точність та ефективність AI значною мірою визначаються алгоритмами та методами TDM, які призначені для пошуку у великих обсягах цифрових джерел неочевидних і корисних на практиці кореляцій та інших закономірностей. TDM виступає сполучною ланкою між інформацією (даними), вилученої з цифрових джерел, та алгоритмічними процесами машинного навчання. Машинне навчання, AI та TDM органічно взаємопов'язані: TDM забезпечує обробку, підготовку та поєднання даних, які придатні для аналізу, у набори даних, на основі яких відбувається безпосереднє навчання нейронних мереж (машинне навчання); машинне навчання забезпечує аналітичне опрацювання даних й виявлення певних закономірностей, а штучний інтелект – застосовує отримані знання на практиці.

Звертається увага, що застосування TDM безпосередньо пов'язане з використанням об'єктів авторського права як джерела даних (інформації), тому з ухваленням Директиви ЄС 2019/790 про авторське право і суміжні права на Єдиному цифровому ринку були регламентовані винятки з легальної монополії правоволодільців у разі використання TDM у наукових та комерційних цілях; встановлені вимоги щодо суб'єктного складу осіб, які застосовують TDM, режиму законного доступу до контенту, строків зберігання копій творів, механізму «відмови від TDM» (opt-out), який дозволяє правоволодільцю заборонити використання творів як вихідних даних для навчання штучного інтелекту.

Окрема увага приділена аналізу судового рішення у справі Robert Kneschke v. LAION e.V., у якій Суд тлумачив положення законодавства ФРН, в якому імплементовані положення Директиви ЄС 2019/790, зокрема, щодо суб'єктів, які можуть послуговуватись винятками для TDM, та характеру досліджень, які ними проводяться.

У зв'язку з тим, що в Україні TDM ще не отримав нормативного закріплення, звертається увага на необхідність вивчення європейського досвіду в цій площині з метою розробки й впровадження аналогічної національної моделі правового регулювання TDM як ключового інструменту машинного навчання.

Ключові слова: штучний інтелект, машинне навчання, інтелектуальний аналіз тексту та даних, авторське право, відмова від TDM.

Ivanova K.Yu. Text and Data Mining as a key tool for machine learning: European experience in legal regulation

The article is dedicated to the study of theoretical and practical aspects of applying text and data mining (TDM).

It is noted that, from a technological perspective, the quality, accuracy, and efficiency of AI largely depend on TDM algorithms and methods, which are designed to identify non-obvious and practically useful correlations and other patterns in large volumes of digital sources. TDM serves as a link between the information (data)

extracted from digital sources and the algorithmic processes of machine learning. Machine learning, AI, and TDM are organically interconnected: TDM provides the processing, preparation, and combination of data suitable for analysis into datasets, which are then used for direct neural network training (machine learning); machine learning ensures analytical processing of the data and the identification of certain patterns; and artificial intelligence applies the acquired knowledge in practice.

It is noted that the use of TDM is directly related to the use of copyright-protected works as data sources (information). Therefore, with the adoption of EU Directive 2019/790 on copyright and related rights in the Digital Single Market, exceptions to the rightholders' legal monopoly were regulated to permit the use of TDM for the purposes of scientific research and, separately, for commercial purposes. The Directive 2019/790 also establishes requirements regarding the types of entities that can carry out TDM, the lawful access to content, retention periods for copies of works, and the "TDM opt-out" mechanism, which allows rightholders to prohibit the use of their works as input data for AI training.

Special attention is given to the analysis of the court decision in Robert Kneschke v. LAION e.V., in which the court interpreted the provisions of German legislation implementing Directive 2019/790, particularly with respect to the entities entitled to rely on TDM exceptions and the nature of the research they conduct.

Since TDM has not yet been legally recognized in Ukraine, the article emphasizes the need to study the European experience in this field in order to develop and implement a similar national model for the legal regulation of TDM as a key tool for machine learning.

Keywords: artificial intelligence, machine learning, text and data mining, copyright, TDM opt-out.