УДК 004.8

Шашок Д.А., Микитюк М.О., Кліменко В.І.

Хмельницький національний університет

ІНТЕЛЕКТУАЛЬНА СИСТЕМА ДЛЯ НЕЙРОМЕРЕЖЕВОГО ВИЯВЛЕННЯ ГЕНДЕРНОЇ ДИСКРИМІНАЦІЇ У СОЦІАЛЬНО-ОРІЄНТОВАНИХ СЕРВІСАХ

У роботі обґрунтовується необхідність розроблення інтелектуальної системи для автоматизованого виявлення гендерної дискримінації в соціально-орієнтованих сервісах, що функціонують у середовищах інтенсивної цифрової комунікації. Підкреслюється, що значний обсяг діалогового контенту, висока варіативність мовних форм і наявність прихованих дискримінаційних патернів роблять традиційні підходи моніторингу малоефективними. Використання сучасних методів обробки природної мови на основі трансформерних моделей створює можливість контекстуального аналізу повідомлень, урахування семантичних і прагматичних особливостей, а також виявлення латентних форм упередженості..

The paper justifies the need to develop an intelligent system for automated detection of gender discrimination in socially oriented services operating in environments of intensive digital communication. It is emphasized that a significant amount of dialogic content, high variability of language forms and the presence of hidden discriminatory patterns make traditional monitoring approaches ineffective. The use of modern methods of natural language processing based on transformer models creates the possibility of contextual analysis of messages, taking into account semantic and pragmatic features, as well as the detection of latent forms of bias..

У сучасному цифровому середовищі соціально-орієнтовані сервіси, такі як гарячі лінії, чати підтримки, сервіси електронної демократії, платформи звернень громадян, тематичні спільноти у месенджерах і соціальних мережах стають ключовим посередником між людиною та інституціями, що забезпечують соціальний захист, психологічну підтримку, доступ до послуг та участь у суспільному житті. Саме в цих комунікаційних просторах особливо гостро проявляються як явні, так і латентні форми гендерної дискримінації: від відкрито агресивних образ і сексистського тролінгу до «м'яких» патерналістських висловлювань, знецінення компетентності, прихованих стереотипів у відповідях операторів сервісів чи модераторів спільнот [1]. Така дискримінація не лише травмує окремих користувачів, а й підриває довіру до інституцій, які позиціонують себе як інклюзивні та орієнтовані на права людини.

Традиційні підходи до виявлення гендерної дискримінації в подібних середовищах спираються або на ручний моніторинг, або на прості ключові слова та

поверхневі фільтри [2]. Це створює дві критичні проблеми. По-перше, людиноорієнтований моніторинг погано масштабується, поглинає значні ресурси й не може забезпечити постійний контроль у режимі реального часу. По-друге, поверхневі лексичні фільтри майже не виявляють контекстуальних, іронічних, завуальованих чи структурно складних висловлювань, які несуть дискримінаційний зміст без використання очевидних «заборонених» слів. У результаті виникає наукове й практичне протиріччя: з одного боку, соціально-орієнтовані сервіси декларують цінності інклюзивності, з іншого, не мають інструментів об'єктивного, систематичного й відтворюваного контролю за гендерно-дискримінаційним контентом, що циркулює в їхньому інформаційному просторі.

Наразі соціально-орієнтовані сервіси, зокрема онлайн-спільноти, форуми й платформи підтримки, акумулюють значні масиви текстових повідомлень, у яких прояви упереджень, стереотипів чи образ можуть бути як явними, так і прихованими [3]. Через великий обсяг даних і варіативність мовних конструкцій ручний моніторинг стає практично неможливим, що підсилює актуальність застосування методів автоматизованого аналізу природної мови [4].

Розвиток нейромережевих моделей обробки природної мови [5], зокрема трансформерів [6], відкрив можливість працювати з текстом не лише на рівні окремих слів [7], а й на рівні контексту [8], семантичних ролей [9], тональних і прагматичних характеристик [10]. Це дозволяє формулювати задачу виявлення гендерної дискримінації у соціально-орієнтованих сервісах як задачу контекстуальної класифікації висловлювань, послідовностей реплік чи цілих діалогів, де рішення базується на інтегральному аналізі мовних патернів [11, 12], а не на механічній перевірці окремих лексем [13]. Попередні розробки нейромережевих методів виявлення сексизму в медіаданих демонструють, що трансформерні моделі, адаптовані до спеціалізованих корпусів, здатні досить надійно відрізняти нейтральні повідомлення від дискримінаційних, включно з латентними формами, за умови коректного проєктування конвеєра підготовки даних, навчання моделей та організації програмної системи [14].

Сучасні підходи NLP забезпечують ефективну обробку великих текстових потоків і здатні виявляти не лише буквальні образливі висловлювання [15], а й приховані форми дискримінації [16], такі як мікроагресії, стереотипні припущення чи непряму об'єктивацію. Моделі глибокого навчання, включно з трансформерними архітектурами [17, 18], дозволяють аналізувати контекст на рівні цілого повідомлення, ураховувати семантичні й прагматичні зв'язки та виявляти залежності, недоступні для класичних лексичних методів. Розширені техніки, такі як контекстуальне ембедування, класифікація з багатьма мітками та моделі з пояснюваністю, підвищують точність визначення упереджених висловлювань і полегшують подальшу валідацію рішень фахівцями [19].

У випадку соціально-орієнтованих сервісів специфіка предметної області є суттєво складнішою, ніж у класичному аналізі відкритих медіа. По-перше, контент тут часто діалоговий: дискримінаційність репліки може виявлятися лише у взаємодії з попередніми повідомленнями, роллю мовця (користувач, консультант, модератор), статусом сервісу (державний, громадський, волонтерський) та очікуваннями користувача щодо підтримки. По-друге, у таких сервісах особливо важливий баланс між чутливістю й специфічністю: система не повинна надмірно «штрафувати» складні, але професійно виважені повідомлення консультантів, водночас має бути здатною виявляти навіть м'які форми знецінення чи стереотипного приписування ролей. По-третє, мова комунікації у соціальних сервісах часто змішана, неформальна, з жаргоном, емодзі, код-міксингом між українською, російською та англійською, що ускладнює застосування стандартних лінгвістичних моделей. Це вимагає спеціальної уваги до формування корпусів, анотування даних, мульти- чи крослінгвального навчання моделей та адаптації токенізаторів до реальної мовної практики.

Проєктування інтелектуальної системи для нейромережевого виявлення гендерної дискримінації у соціально-орієнтованих сервісах доцільно розглядати як багаторівневий процес, що починається з концептуалізації предметної області й закінчується побудовою програмної архітектури, готової до інтеграції у реальні сервіси. Схема запропонована на рисунку 1. На концептуальному рівні ядром системи є модуль аналізу тексту на основі трансформерної моделі, донавченої на корпусі повідомлень соціальних сервісів з позначками «гендерно-дискримінаційне» та «недискримінаційне», а також, за потреби, додатковими підкласами, які відображають різні типи дискримінаційних проявів, наприклад відкрите приниження, сексуалізовані образи, патерналістські формулювання чи скритий стереотипний розподіл ролей. Навколо цього ядра вибудовується інфраструктура, яка забезпечує захоплення й нормалізацію даних із сервісів, знеособлення персональної інформації, зберігання історій взаємодій, формування сигналів для модерації, аналітики та звітності, а також механізми аудиту і контролю якості самого алгоритму.

Систему доцільно організувати за клієнт-серверною моделлю з виділенням кількох логічних шарів: шару інтеграції з каналами комунікації, де реалізується приймання повідомлень із чат-ботів, веб-форм, API соціальних мереж чи контакт-центрів; шару обробки й зберігання даних, який відповідає за нормалізацію тексту, знеособлення, побудову діалогових контекстів та ведення бази даних; інтелектуального шару, де реалізовано пайплайн нейромережевого аналізу, обчислення індикаторів ризику й формування пояснень; а також прикладного шару, що надає інтерфейси для модераторів, адміністраторів сервісу, аналітиків і, у спрощеному вигляді, для користувачів, які можуть отримувати зворотний зв'язок

щодо неприйнятних формулювань. Така багарівнева організація дозволяє розвести проблеми масштабованості, захисту даних, навчання моделей і користувацької взаємодії, роблячи систему гнучкою щодо подальшого розширення.

На рівні моделі даних ключовим є представлення комунікації не як набору окремих повідомлень, а як мережі взаємопов'язаних сутностей. Кожна сесія взаємодії у соціально-орієнтованому сервісі може бути змодельована як діалог, що має час початку, тип каналу, анонімізовані профілі учасників та послідовність реплік. Кожна репліка має текст, метадані (час, канал, роль мовця, мовні налаштування, технічні параметри) і результати аналізу: оцінку ймовірності гендерної дискримінації, тип виявленого патерну, індикатори впевненості моделі та, за потреби, фрагменти тексту, які зробили найбільший внесок у рішення. На цьому ж рівні фіксуються випадки ручної корекції результатів модераторами, що дозволяє використовувати їх надалі в циклі донавчання моделі та контролю її упередженості.

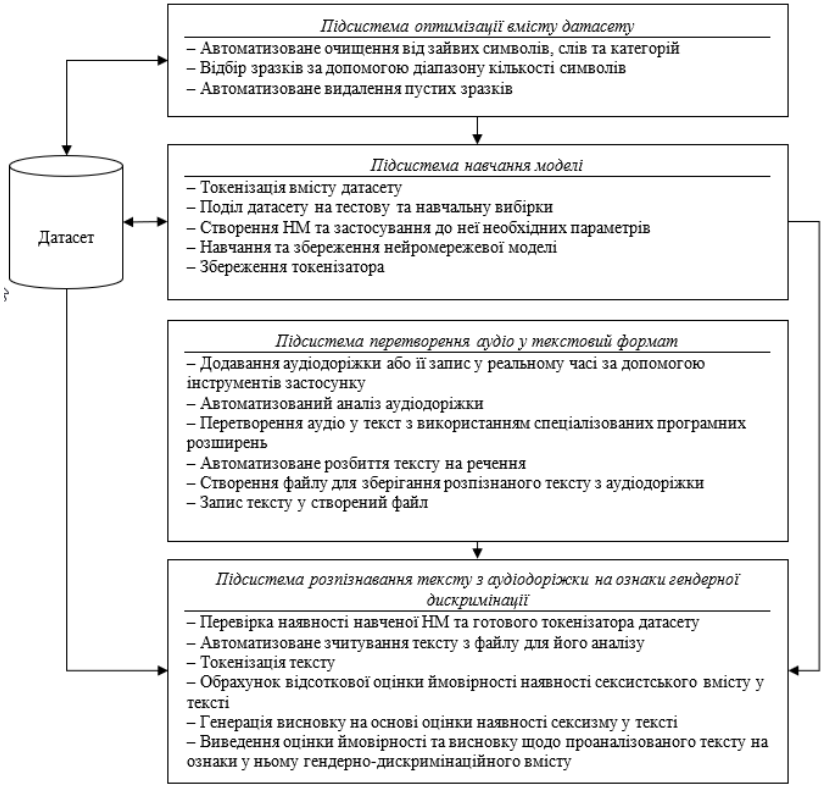

Рисунок 1 – Взаємодія складових інтелектуальної системи

Опис проєктування інтелектуальної системи неможливий без деталізації процесу побудови нейромережевого ядра. Базовим вибором є використання багатомовної трансформерної архітектури типу mBERT або XLM-R, здатної обробляти україномовні, російськомовні й англомовні фрагменти у межах однієї моделі. Початкове налаштування передбачає адаптацію токенізатора до специфіки текстів соціальних сервісів, з урахуванням характерних скорочень, емодзі, хештегів, посилань, слів із транслітерацією та кодуванням латиницею. Далі формується навчальний корпус, що включає анонімізовані дані з реальних сервісів (за умови дотримання етичних норм і правового регулювання), а також доповнюється відкритими корпусами сексистських висловлювань і нейтральних текстів із соціальних мереж, форумів, петицій чи коментарів до публікацій. Анотація здійснюється за чітко визначеною схемою, що передбачає розмежування гендерної дискримінації від інших видів токсичності, зокрема загального хейту, політичної агресії чи мови ворожнечі за іншими ознаками.

Сам процес навчання організовується як задача бінарної або багатокласової класифікації з можливістю подальшого переходу до багатоміткової постановки, якщо виникає потреба виділяти різні типи гендерної дискримінації паралельно. На вхід моделі подаються токенізовані репліки або короткі фрагменти діалогу, зафіксовані у певному часовому вікні. Модель видає вектор ймовірностей по класах, який далі використовується для прийняття рішення на рівні окремого повідомлення, а також агрегується на рівні діалогу, користувача чи сервісу. Для контролю якості використовуються стандартні метрики класифікації, але в інтерпретації, що враховує асиметрію вартості помилок: хибно-негативні рішення, коли дискримінаційний контент не виявлено, мають вищу критичність, ніж хибно-позитивні, що маркують нейтральні висловлювання як підозрілі.

Проєктування класів програмної системи відображається у вигляді діаграми класів (рисунок 2), яку доцільно організувати навколо кількох концептуальних центрів. Першим центром є доменні класи, що описують структуру соціально-орієнтованих сервісів і комунікацій у них. До них належать класи, які моделюють користувача сервісу з його анонімізованим профілем, роль у системі, історію взаємодій, а також клас оператора або консультанта, що репрезентує сторону сервісу. Діалог як процес відображається окремим класом, який утримує колекцію реплік, пов'язаних із певною сесією, та метадані каналу взаємодії. Репліка представлена окремим класом із полями для тексту, часових міток, посилань на автора, структуру контексту та результатів аналізу.

Другим концептуальним центром є класи, що реалізують інтелектуальний модуль. Центральне місце посідає клас, що інкапсулює нейромережеву модель, вміє завантажувати ваги, виконувати прямий прохід і повертати ймовірнісні оцінки за

класами. Допоміжні класи реалізують попередню обробку тексту, включно з нормалізацією, очищенням від технічних елементів, сегментацією діалогу та побудовою входів моделі, а також постобробку результатів, зокрема формування текстового або структурованого висновку про наявність і тип гендерної дискримінації. Окремий клас може відповідати за генерацію пояснень, використовуючи методи пояснюваного машинного навчання, наприклад виділення фрагментів тексту, які зробили максимальний внесок у рішення, у вигляді вагових коефіцієнтів чи теплових карт.

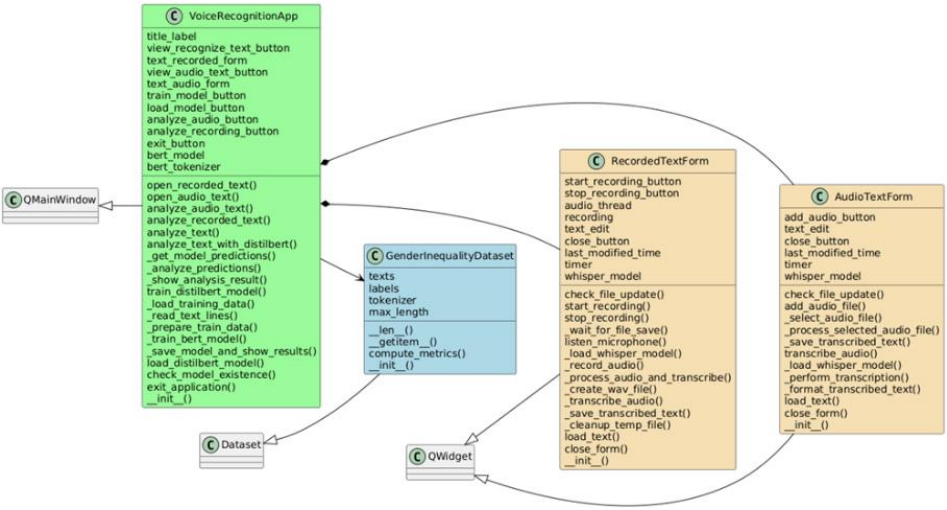

Рисунок 2 – Діаграма класів інтелектуальої системи

Третій центр діаграми класів стосується інфраструктури зберігання та інтеграції. Тут доцільно виокремити класи, які реалізують репозиторії для діалогів, користувачів, результатів аналізу та журналів модераційних дій. Ці класи виступають абстракцією над конкретною СКБД, забезпечуючи транзакційність, логування змін, підтримку історії перегляду та можливість відновлення процесу у разі помилок. Їм протистоять класи, відповідальні за інтеграцію з зовнішніми сервісами – адаптери до API месенджерів, модулі обробки вебхуків, компоненти для імпорту історичних даних. Такий поділ дозволяє мінімізувати зв'язність між бізнес-логікою й технічними засобами доставки даних.

Нарешті, четвертий центр об'єднує класи користувацького інтерфейсу та сервісних компонентів. Окремі класи відповідають за вікна модератора чи адміністратора, у яких відображаються черги повідомлень із високим ризиком

дискримінації, детальні пояснення рішень моделі, агреговані статистики за каналами й тематичними категоріями. Інші класи представляють налаштування системи, конфігурацію порогів спрацювання, сценарії обробки інцидентів, ролі й права доступу. Логічним центром цього шару є клас, що координує взаємодію між інтерфейсом, репозиторіями даних та інтелектуальним модулем, реалізуючи шаблон фасаду для спрощення внутрішніх залежностей.

Діаграма класів, побудована з урахуванням описаних центрів, демонструє низку важливих відносин. Клас діалогу агрегаційно пов'язаний із класом репліки, оскільки утримує множину реплік, але вони можуть зберігатися автономно у базі даних. Репліка асоційована з класом користувача і класом оператора, що відображає напрямок та суб'єкта висловлювання. Клас інтелектуального модуля композиційно містить класи попередньої обробки та генератора пояснень, оскільки без них не може коректно функціонувати. Класи репозиторіїв перебувають у відношенні залежності з класами доменної моделі, оскільки реалізують збереження їхніх екземплярів, тоді як інтерфейсні класи використовують доменну модель для відображення й редагування відповідних сутностей.

У процесі проєктування особливу увагу приділяється питанням етичності й мінімізації вторинної дискримінації, яку потенційно можуть відтворювати алгоритми. Це виявляється у необхідності контролю якості моделі не лише на глобальних метриках, а й окремо для різних підгруп користувачів, з урахуванням гендерної ідентичності, мови, типу сервісу. Система має зберігати результати валідаційних експериментів, журнали ручних корекцій модераторів, а також надавати інструменти для аудиту рішень у проблемних випадках, коли користувачі оскаржують дії модерації. Такі механізми не лише підвищують прозорість, а й створюють зворотний зв'язок, необхідний для поступового донавчання моделі на більш репрезентативних і збалансованих корпусах.

У підсумку проєктована інтелектуальна система поєднує концептуально виважений підхід до аналізу гендерної дискримінації у соціально-орієнтованих сервісах із чітко структурованою програмною архітектурою. Нейромережевий модуль на основі трансформерної моделі забезпечує здатність виявляти як явні, так і приховані форми дискримінації в умовах «шумної» реальної мовної практики, тоді як багаторівнева система класів, від доменної моделі до інтерфейсів адміністратора, робить можливим безперервний моніторинг, модерацію, аналітику та аудит. Така система здатна стати ключовим елементом цифрової інфраструктури соціально-орієнтованих сервісів, підтримуючи досягнення цілей гендерної рівності та побудови інклюзивних комунікаційних просторів у відповідності до глобальних орієнтирів сталого розвитку.

Перелік посилань

1. Jakob, M., Klauer, K. C., & Shechter, A. (2024). Detecting Gender Discrimination in a Context Disfavoring and Favoring Both Men and Women–a Signal Detection Approach.
2. Foley, S., Ngo, H. Y., Loi, R., & Zheng, X. (2015). Gender, gender identification and perceived gender discrimination: An examination of mediating processes in China. Equality, Diversity and Inclusion: An International Journal, 34(8), 650-665..
3. Молчанова М.О., Мазурець О.В., Собко О.В., Віт Р.В., Назаров В.В. Алгоритм виявлення аб'юзивного вмісту в україномовному аудіоконтенті для імплементації в об'єктно-орієтовану інформаційну систему. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №1 (331). С. 101-106.
4. Молчанова М.О., Мазурець О.В., Собко О.В., Кліменко В.І., Андрощук В.І. Метод нейромережевого виявлення кібербулінгу з використанням хмарних сервісів та об'єктно-орієнтованої моделі. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №2 (333). С. 200-206.
5. Денисенко Б.О., Молчанова М.О., Мазурець О.В. Інтелектуальна система виявлення дезінформації з застосуванням штучних нейронних мереж. Збірник наукових праць за матеріалами XVI Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2024». 15-16 листопада 2024. Хмельницький, 2024. с. 167-174.
6. Blazhuk V., Mazurets O., Zalutska O. An Approach to Using the mBERT Deep Learning Neural Network Model for Identifying Emotional Components and Communication Intentions. The Impact of Scientific Research on the Development of the Modern World. Proceedings of the XLIV International scientific and practical conference. October 23-25, 2024. Dubrovnik, Croatia. 2024. Pp. 79-84.
7. Sobko O., Mazurets O., Didur V., Chervonchuk I. Recurrent Neural Network Model Architecture for Detecting a Tendency to Atypical Behavior Of Individuals by Text Posts. Theoretical and Practical Aspects of Modern Research. Proceedings of XXVI International scientific and practical conference. June 5-7, 2024. International Scientific Unity. Ottawa, Canada. 2024. Pp. 113-117.
8. Tymofiiev I., Mazurets O., Hardysh D., Molchanova M. Neural Network Dual Architecture for Depression Detection Using Cloud Services. Scientific Research in the Era of Digital Technologies: Challenges and Opportunities. Proceedings of the XLVI International scientific and practical conference. November 6-8, 2024. Barcelona, Spain. 2024. Pp. 84-88.
9. Mazurets O., Molchanova M., Klimenko V., Prosvitliuk M Practice Implementation of Neural Network Model BART-Large-CNN for Text Annotation. Prospects of Scientific Research in the Conditions of the Modern World. Proceedings of XXVII International scientific and practical conference. June 12-14, 2024. Rotterdam, Netherlands. 2024. Pp. 97-102.
10. Shevchuk P., Molchanova M., Mazurets O. Software for Text Messages Reliability Analysis Based on the Machine Learning Models Ensemble. Proceedings of IV International Scientific and Practical Conference «Innovative research and perspectives of the development of science and technology». January 29-31, 2024. Stockholm, Sweden. 2024. Pp. 347-354.
11. Yurchenko D., Mazurets O., Didur V., Molchanova M. Approach to Using Cloud Services for Visual Analytics of Neural Network Analysis of Texts Emotional Tonality. The Future of Scientific

Discoveries: New Trends and Technologies. Proceedings of the XLVII International scientific and practical conference. November 13-15, 2024. Marseille, France. 2024. Pp. 108-113.
12. Hladun O., Mazurets O., Molchanova M., Sobko O. Real Time Detection the Person Emotion State Using Neural Network. Scientific Research: Modern Innovations and Future Perspectives. Proceedings of the 2 International scientific and practical conference. November 25-27, 2024. Montreal, Canada. 2024. Pp. 119-123.
13. Mazurets O., Sobko O., Vit R., Pasternak V. Practical Approach for Detection by Deep Learning of Target Objects of Subject Area Based on Semantic Connectivity Indicators in Audio Database. Proceedings of XXIV International Scientific and Practical Conference «Modern Scientific Challenges are the Driving Force of the Development of Scientific Research». May 22-24, 2024. Bruges, Belgium. International Scientific Unity. 2024. Pp. 91-96.
14. Овчарук О.М., Мазурець О.В. Нейромережевий метод діагностування психологічних розладів за аналізом повідомлень на основі роздільного підходу до класифікації. Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах». № 1, 2025. с. 210-216.
15. Mazurets O.V., Sobko O.V., Molchanova M.O., Zalutska O.O., Yurchak A.V. Practical Implementation of Neural Network Method for Stress Features Detection by Social Internet Networks Posts. Global Science: Prospects and Innovations. Proceedings of the II International Scientific and Theoretical Conference «Scientific Review of the Actual Events, Achievements and Problems». May 31, 2024. Berlin, Federal Republic of Germany: International Center of Scientific Research. 2024. Pp. 160-167.
16. Овчарук О.М., Мазурець О.В. Нейромережева архітектура з квантовим шаром для аналізу текстових повідомлень на прояви посттравматичного стресового розладу. Науковий журнал «Наука і техніка сьогодні». Київ, 2024. №13 (41). С. 1192-1204.
17. Віт Р.В., Мазурець О.В. Підхід до тематичної класифікації текстової інформації засобами обробки природної мови. Науковий журнал «Наукові праці Донецького національного технічного університету», серія «Проблеми моделювання та автоматизації проектування». 2025. №1 (21). С. 94-99.
18. Murava V., Zalutska O., Didur V., Mazurets O. Software architecture of information system for exchanging LLM thematic prompts. Global Trends in the Development of Information Technology and Science. Proceedings IV International Scientific and Practical Conference. June 25-27, 2025. Stockholm, Sweden. Pp. 121-127.
19. Мазурець О.В., Тимофієв І.А., Кліменко В.І., Тищенко О.О. Метод виявлення депресивного стану пов'язаного із навчанням у закладах освіти із використанням нейромережі дуальної архітектури. Науковий журнал «Вісник Херсонського національного технічного університету». 2024. №4 (91). С. 311-318.