УДК 004.8 DOI https://doi.org/10.35546/kntu2078-4481.2024.4.45

О. В. СОБКО
викладач кафедри комп'ютерних наук
Хмельницький національний університет
ORCID: 0000-0001-5371-5788

МЕТОД НЕЙРОМЕРЕЖЕВОГО ФОРМУВАННЯ РЕПРЕЗЕНТАТИВНИХ НЕДИСКРИМІНАЦІЙНИХ ТЕКСТОВИХ ДАТАСЕТІВ ЗГІДНО FATE-ПРИНЦИПУ СПРАВЕДЛИВОСТІ

У статті запропоновано метод нейромережевого формування репрезентативних недискримінаційних текстових датасетів згідно FATE-принципу справедливості. Запропонований метод акцентує увагу на створенні збалансованих датасетів, які точно відображають демографічні групи, враховуючи етичні аспекти, такі як гендер, вік, релігія та етнічність. Метод полягає в ідентифікації та коригуванні дисбалансів у датасеті шляхом розв'язання оптимізаційної задачі, що обирає дані для видалення або аугментації таким чином, щоб кінцевий датасет залишався репрезентативним і неупередженим. Для оцінки ефективності цього підходу було розроблено програмне забезпечення, яке використовує моделі машинного навчання, зокрема SVM для класифікації за віковим аспектом, LSTM для гендерної класифікації та BERT для релігійної класифікації, всі з яких показали високі статистичні результати.

Результати застосування цього методу показують, що після формування датасет став більш репрезентативним з точки зору справедливості за віковим та гендерним аспектами, з мінімальними відхиленнями (від 0.00% до 0.04%) від ідеального репрезентативного розподілу. Ці результати демонструють, що запропонований метод здатний ефективно аналізувати та формувати датасети, забезпечуючи їх відповідність стандартам справедливості за різними етичними категоріями. Крім того, цей підхід сприяє досягненню Цілей сталого розвитку, зокрема Цілі № 5 (гендерна рівність), Цілі № 10 (скорочення нерівності) та Цілі № 11 (сталий розвиток міст і громад). Забезпечення того, щоб датасети відображали різноманітне і інклюзивне представлення соціальних груп, сприяє створенню етичних та справедливих систем штучного інтелекту, що допомагає зменшити упередженість та дискримінацію в процесах прийняття рішень.

Ключові слова: *репрезентативність, етичні принципи, недискримінація, датасет, Цілі сталого розвитку.*

O. V. SOBKO
Lecturer of Computer Science Department
Khmelnytskyi National University
ORCID: 0000-0001-5371-5788

METHOD OF NEURAL NETWORK FORMATION OF REPRESENTATIVE NON-DISCRIMINATORY TEXT DATASETS ACCORDING TO THE FATE PRINCIPLE OF JUSTICE

Paper presents neural network method for generating representative non-discriminatory text datasets according to the FATE fairness principle. The proposed method focuses on creating balanced datasets that accurately reflect demographic groups, taking into account ethical aspects such as gender, age, religion, and ethnicity. The method consists of identifying and correcting imbalances in the dataset by solving an optimization problem that selects data for deletion or augmentation in such a way that the final dataset remains representative and unbiased. To evaluate the effectiveness of this approach, software was developed that uses machine learning models, in particular SVM for age classification, LSTM for gender classification, and BERT for religious classification, all of which showed high statistical results.

The results of this method show that after generation, the dataset became more representative in terms of fairness in terms of age and gender, with minimal deviations (from 0.00% to 0.04%) from the ideal representative distribution. These results demonstrate that the proposed method is able to effectively analyze and generate datasets, ensuring their compliance with fairness standards for different ethical categories. In addition, this approach contributes to the achievement of the Sustainable Development Goals, in particular Goal No. 5 (gender equality), Goal No. 10 (reduced inequality) and Goal No. 11 (sustainable urban and community development). Ensuring that datasets reflect a diverse and inclusive representation of social groups contributes to the creation of ethical and fair AI systems, which helps reduce bias and discrimination in decision-making processes.

Key words: *representativeness, ethical principles, non-discrimination, dataset, Sustainable Development Goals.*

Постановка проблеми

У сучасному світі проблема дискримінації в даних та алгоритмах машинного навчання є актуальною, оскільки вона безпосередньо впливає на якість прийняття рішень у різних галузях [1, С. 217–221; 2, С. 16–28]. Для запобігання упередженням та забезпечення етичного використання даних у машинному навчанні необхідно

дотримуватись принципів FATE (Fairness, Accountability, Transparency, Ethics). Наведені принципи спрямовані на забезпечення справедливості, відповідальності, прозорості та етики у процесах збору, обробки та аналізу даних. Дотримання FATE-принципів є основою для побудови систем, які сприяють соціальній справедливості та сталому розвитку [3, С. 344–356; 4, С. 262–265].

Також сьогодні дуже актуальними є Цілі сталого розвитку (SDG, Sustainable Development Goals). Ці цілі повинні поширюватися і на сферу штучного інтелекту, оскільки останній стає дедалі більш інтегрованим у повсякденне життя людей. Вплив рішень, які приймають алгоритми штучного інтелекту, відчувається у таких сферах, як охорона здоров'я, освіта, правосуддя, соціальна політика та багато інших. Тому важливо, щоб розвиток штучного інтелекту відповідав принципам сталого розвитку, спрямованим на створення справедливого, інклюзивного і стійкого суспільства [5, С. 91–112].

Текстові датасети, що використовуються у машинному навчанні, повинні також відповідати цілям сталого розвитку, зокрема Ціль № 5 (гендерна рівність), Ціль № 10 (скорочення нерівності) та Ціль № 11 (сталий розвиток міст та громад). Забезпечення гендерної рівності (Ціль № 5) в даних важливе для того, щоб уникнути упередженості, яка може спричинити дискримінацію за статевими ознаками в майбутніх рішеннях моделей машинного навчання. Врахування скорочення соціальних та економічних нерівностей (Ціль № 10) у створенні датасетів допомагає уникнути помилок, що можуть виникнути через недостатнє представлення меншин або маргіналізованих груп. Окрім того, створення збалансованих і репрезентативних даних підтримує сталий розвиток міст та громад (Ціль № 11), забезпечуючи рівні можливості для всіх груп населення і знижуючи ризик соціальної несправедливості у технологіях, які можуть бути застосовані в містах і громадах.

Таким чином, забезпечення репрезентативності та відповідність цілям сталого розвитку у текстових датасетах є важливим кроком для забезпечення прозорості та відповідальності у системах прийняття рішень.

Аналіз останніх досліджень і публікацій

Досягнення Цілей сталого розвитку та FATE-принципів вимагає розробки рішень, що сприяють соціальній рівності, економічному зростанню та екологічній стабільності. Далі наведено огляд наукових публікацій, які присвячені досягненню Цілей сталого розвитку та FATE-принципів у моделях штучного інтелекту.

У дослідженні [6, С. 969–985] автори створили корпус текстів щодо Цілей сталого розвитку (SDGs) японською мовою та використали модель BERT для розпізнавання та векторизації семантики пов'язаних з SDGs речень. Модель демонструє хороші результати в ідентифікації відповідних Цілей та прогнозуванні зв'язків між ними, що може бути використано для знаходження можливих партнерів для співпраці в рамках SDGs.

У роботі [7, С. 103249] автори досліджують формування репрезентативної підмножини тексту, в великих текстових наборах даних, що використовуються для навчання PreLM (переднавчених мовних моделей). Експериментальні результати показали, що RepSet, отриманий за допомогою методу на основі різниці ймовірностей, досягав 90% ефективності, навіть при тому, що розмір RepSet був на два-три порядки менший за розмір оригінального набору даних.

У статті [8, С. 1–4] автори аналізують проблему упередженості в системах штучного інтелекту та її негативний вплив на різні групи людей. Автори пропонують впровадження принципів різноманіття та інклюзії на всіх етапах розробки ШІ для створення справедливих і надійних систем, в тому числі і на етапі формування навчальних даних для моделей машинного навчання. Рекомендації включають врахування різних перспектив, зменшення упередженості в даних та забезпечення прозорості алгоритмів.

Стаття [9] описує проблему несправедливості в машинному навчанні через нерівномірне представлення класів і захищених груп (наприклад, за статтю, расою чи віком) у даних. Автори пропонують новий алгоритм Fair Oversampling, який одночасно зменшує дисбаланс класів і груп, покращуючи точність і справедливість моделей. Також вони розробляють метрику Fair Utility, яка об'єднує збалансовану точність із показниками справедливості.

Аналізуючи представлені дослідження, можна зробити висновок, що їх основна мета – вирішення різних аспектів підготовки навчальних даних для алгоритмів машинного навчання у контексті Цілей сталого розвитку та етичного використання штучного інтелекту. Однак не всі роботи повністю враховують ключові етичні засади, зокрема принципи недискримінаційності та забезпечення репрезентативності підгруп населення, що є важливими для створення збалансованих і справедливих моделей ШІ.

Формулювання мети дослідження

Метою роботи є забезпечення дотримання принципу справедливості FATE для навчальних датасетів, яке полягає у створенні методу нейромережевого формування репрезентативних недискримінаційних текстових датасетів для подальшого їх використання у навчанні нейромережевих моделей для вирішення різноманітних задач.

Викладення основного матеріалу дослідження

Метод нейромережевого формування репрезентативних недискримінаційних текстових датасетів згідно FATE-принципу справедливості подано у вигляді трьох послідовних етапів: перевірки коректності елементів датасету, аналізу репрезентативності за етичними аспектами та репрезентативне коригування датасету. Кожен етап складається з своїх кроків, які наведено на рисунку 1.

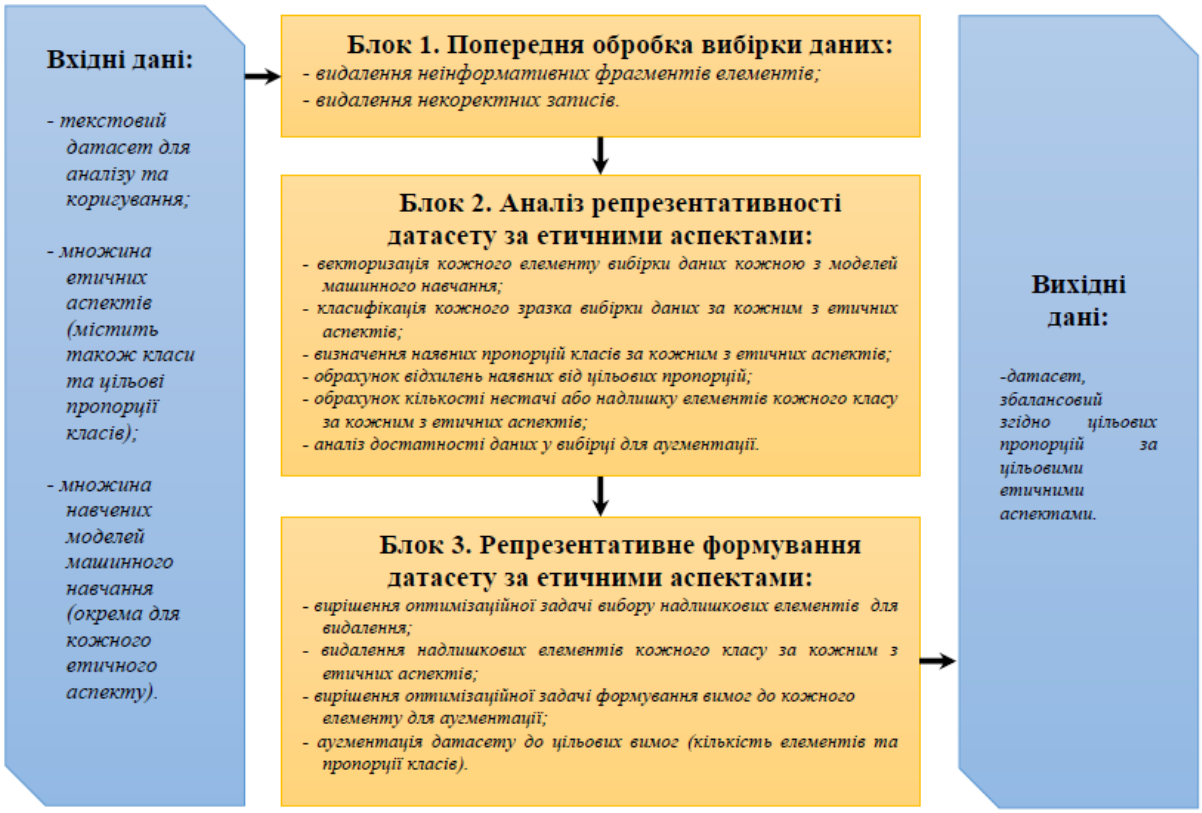

Рис. 1. Схема методу нейромережевого формування репрезентативних недискримінаційних текстових датасетів згідно FATE-принципу справедливості

Метод нейромережевого формування репрезентативних недискримінаційних текстових датасетів згідно FATE-принципу справедливості складається з 3 блоків, кожен із яких виконує важливі кроки.

У Блоці 1 здійснюється попередня обробка вхідних даних. На цьому етапі видаляються неінформативні фрагменти, які не несуть корисної інформації для подальшого аналізу, а також некоректні записи, які можуть включати помилки чи невідповідності у структурі даних. Цей блок виконує функцію очищення даних, створюючи основу для точності наступних етапів.

У Блоці 2 проводиться аналіз репрезентативності вибірки за етичними аспектами. Спочатку всі елементи вибірки проходять процес векторизації, що перетворює текстові дані у формат, придатний для обробки моделями машинного навчання. Потім кожен зразок класифікується за етичними аспектами, які можуть включати різні категорії, що відображають конкретні проблеми, пов'язані з недискримінаційністю. На основі класифікації визначаються наявні пропорції вибірки для кожного етичного аспекту, що дозволяє виявити відхилення від цільових пропорцій, які були задані як критерій справедливості. Паралельно проводиться підрахунок нестачі або надлишкових елементів у кожному класі, що дозволяє оцінити обсяги необхідних коригувань. Завершується цей блок аналізом достатності даних у вибірці, щоб визначити, чи відповідає вона вимогам для наступних етапів.

У Блоці 3 здійснюється формування репрезентативного датасету з урахуванням етичних аспектів. Спочатку вирішується завдання оптимізації вибірки шляхом видалення надлишкових елементів у тих класах, які мають більшу представленість, ніж це передбачено цільовими пропорціями. Далі відбувається видалення надлишкових елементів окремих класів за кожним із етичних аспектів. Після цього визначаються оптимальні вимоги для кожного елемента, який може бути використаний для аугментації, тобто штучного розширення вибірки. Завершується цей блок процесом аугментації, де до датасету додаються нові елементи або коригуються існуючі таким чином, щоб отримати збалансований набір даних, який повністю відповідає цільовим пропорціям.

Після виконання усіх трьох блоків формується збалансований датасет, який враховує принципи справедливості відповідно до етичних аспектів і може використовуватися для навчання моделей, що працюють із текстовими даними.

Для створення набору навчених моделей машинного навчання, які відповідатимуть окремим етичним аспектам, необхідно навчити кожну модель-класифікатор для аналізу репрезентативності датасету згідно з другим блоком, описаним на рисунку 1. Щоб отримати ці класифікатори, які формуватимуть множину моделей, орієнтованих на етичні аспекти, слід виконати послідовність дій, представлених на рисунку 2.

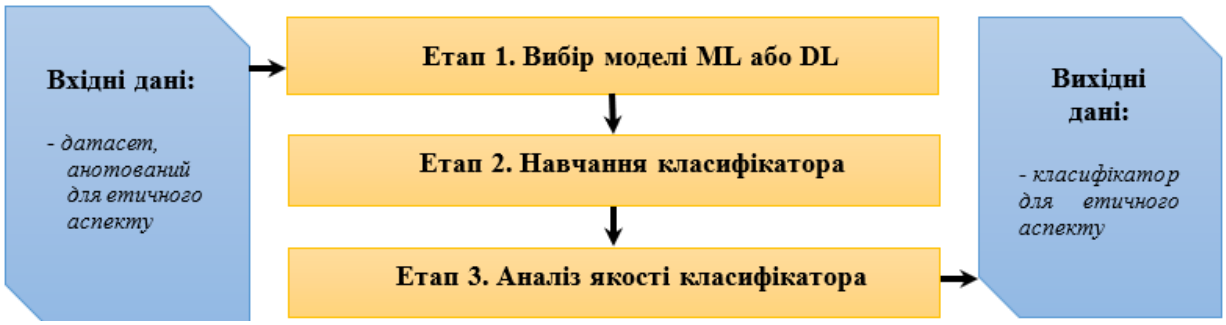

Рис. 2. Схема отримання навченої моделі машинного навчання для окремого етичного аспекту

Першим етапом є обрання моделі машинного навчання, яка підходить для класифікації текстів за етичними аспектами. Для цього можуть застосовуватися як моделі глибокого навчання, такі як BERT, GPT, LSTM, GRU, так і традиційні алгоритми класифікації, наприклад, Logistic Regression, Naive Bayes, Support Vector Machines або k-Nearest Neighbors. Після вибору моделі наступним етапом здійснюється її навчання на анотованому датасеті, підготовленому для аналізу конкретного етичного аспекту.

На завершальному етапі виконується оцінка якості отриманої моделі на основі таких метрик, як Accuracy, Precision, Recall та F1-score. Якщо результати аналізу якості є незадовільними, необхідно повернутися до етапу вибору моделі та повторити процес. У разі досягнення прийнятної якості формується класифікатор, який здатен оцінювати репрезентативність текстових даних відповідно до заданого етичного аспекту.

Таким чином, для кожного етичного аспекту створюється окрема модель машинного навчання, що в підсумку формує множину етичних моделей, кількість яких відповідає кількості проаналізованих аспектів. Це забезпечує створення репрезентативної текстової вибірки, враховуючи етичні вимоги.

Для перевірки ефективності методу нейромережевого формування репрезентативних недискримінаційних текстових датасетів згідно FATE-принципу справедливості було сформовано вхідний датасет, який об'єднує два набори даних: «Cyberbullying Classification» [10] та «Cyberbully Detection Dataset» [11]. Перший з них включає 46,017 твітів, розмічених за шістьма категоріями видів кібербулінгу. Другий датасет містить 99,989 твітів, також класифікованих за типами кіберзалякування. Жоден із цих датасетів не має інформації щодо статі, вікової групи, релігії чи етнічного походження авторів повідомлень.

Для тренування моделей машинного навчання, які використовуватимуться для розмітки вхідного датасету, було застосовано додаткові набори даних, що охоплюють три етичні аспекти принципу справедливості: гендер, вік і релігію. Оскільки класи у цих наборах даних були нерівномірно представлені за кількістю зразків, що могло погіршити якість навчання моделей, було проведено балансування всіх класів за кількістю зразків у них.

Для оцінки ефективності методу нейромережевого формування репрезентативних недискримінаційних текстових датасетів згідно FATE-принципу справедливості було розроблено програмну реалізацію з використанням мови Python. Для класифікації текстів із вхідного датасету кіберзалякувань за ознаками гендеру, віку та релігії застосовано бібліотеку TensorFlow. Програмна реалізація також надає можливість переглядати класи даних у наборі відповідно до міток, які відображають обрані етичні аспекти, що продемонстровано на рисунку 3.

Для оцінки ефективності запропонованого методу нейромережевого формування репрезентативних недискримінаційних текстових датасетів згідно FATE-принципу справедливості було навчено кілька моделей машинного навчання. Показники статистичних метрик, таких як Accuracy, Precision, Recall і F1-score, для моделей різних етичних аспектів подано на рисунку 4.

Для різних класів було досягнуто різних рівнів лінійної роздільної здатності. Для релігійної ознаки класифікатор BERT, який продемонстрував найвищу ефективність серед моделей, забезпечив добре роздільні дані. За гендерною ознакою найкращі результати показав класифікатор LSTM, однак рівень роздільної здатності залишився середнім. Для вікової ознаки використання класифікатора SVM показало найгіршу роздільну здатність.

Для аналізу та створення репрезентативної вибірки текстових даних, що відповідає цільовим пропорціям класів за віковими та гендерними категоріями, було використано дані про популяція України. Інформація для цього була взята з даних Інституту демографії та соціальних досліджень імені М. В. Птухи Національної академії наук України.

Встановлено, що датасет не є репрезентативним, оскільки кількість текстових зразків у класах різних етичних аспектів не відповідає пропорціям демографічних груп населення України. Це вимагає балансування даних для забезпечення їхньої репрезентативності.

Внаслідок розв'язання оптимізаційної задачі для створення репрезентативної вибірки за віковими та гендерними етичними аспектами на прикладі демографічних груп популяції України, було отримано репрезентативну вибірку текстових даних шляхом аугментації. Баланс класів цієї вибірки наведено в таблиці 1.

Dataset markup example

Labeling dataset

	Name	Label
▶	Not_Cyberbullying	0
	Cyberbullying_Gender	1
	Cyberbullying_Religion	2
	Other_Cyberbullying	3
	Cyberbullying_Age	4
	Cyberbullying_Ethnicity	5
*		

Results

	Sentence	Cyberbullying_typ	Cisgender	Religion	Age
▶	In other words ...	0	Woman	Jew	40-49
	Why is #aussiet...	0	Woman	Muslim	0-19
	@XochitlSuckkk...	0	Man	Jew	30-39
	@Jason_Gio me...	0	Man	Muslim	30-39
	@RudhoeEnglis...	0	Man	Muslim	0-19
	@RajaSaab @Q...	0	Man	Buddhism	30-39
	Itu sekolah ya b...	0	Man	Jew	0-19
	Karma. I hope it...	0	Woman	Muslim	50-100
	@stockputout e...	0	Woman	Jew	0-19
	Rebecca Black ...	0	Man	Jew	0-19
	@Jord_Is_Dead ...	0	Man	Jew	0-19

Рис. 3. Перегляд міток зразків з вибірки за гендерним, релігійним та віковим етичними аспектами

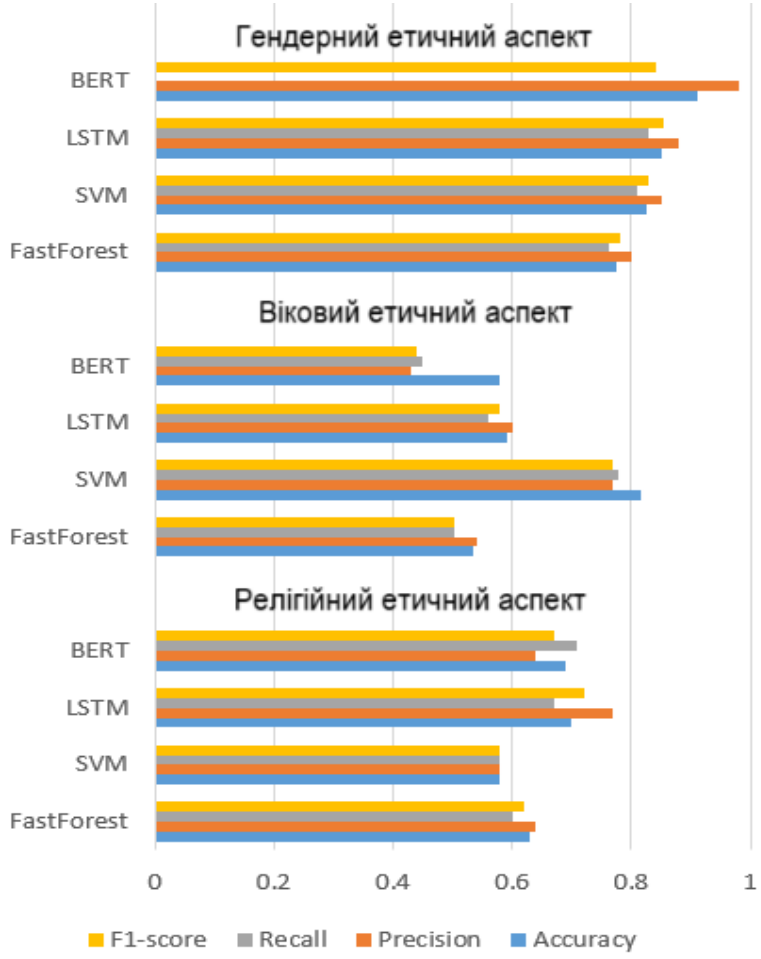

Рис. 4. Перегляд міток зразків з вибірки за гендерним, релігійним та віковим етичними аспектами

Таблиця 1

Розподіл зразків у сформованому репрезентативному даасеті після аугментації даних

Вікові демографічні підгрупи	0-19 років	20-29 років	30-39 років	40-49 років	50-100 років
Відсоткове відношення демографічних груп за гендером та віком у популяції України					
Чоловіки	9.67%	5.64%	8.96%	7.79%	15.56%
Жінки	9.04%	4.53%	7.96%	7.47%	23.38%
Відсоткове відношення демографічних груп за гендером та віком у текстовій вибірці					
Чоловіки	9.65%	5.62%	8.94%	7.80%	15.57%
Жінки	9.05%	4.57%	7.97%	7.45%	23.38%
Одержане відхилення від репрезентативного розподілу					
Чоловіки	0.02%	0.02%	0.02%	0.01%	0.02%
Жінки	0.01%	0.04%	0.01%	0.02%	0.00%

В результаті формування датасету за розробленим методом було отримано відхилення розподілів зразків за класами вікового та гендерного етичних аспектів від ідеального репрезентативного розподілу, які склали: мінімум 0.00%, максимум 0.04%, середнє 0.02%.

Таким чином, завдяки виконанню всіх кроків методу нейромережевого формування репрезентативних недискримінаційних текстових датасетів згідно FATE-принципу справедливості було створено недискримінаційний і неупереджений датасет, який пропорційно відображає демографічні підгрупи популяції України. Це відповідає принципу справедливості FATE. Крім того, такий підхід підтримує досягнення Цілей сталого розвитку, зокрема в частинах, що стосуються рівності (Ціль № 5), зменшення нерівності (Ціль № 10) та створення справедливих і сталих міст та громад (Ціль № 11). Використання збалансованих, репрезентативних даних сприяє формуванню технологій, які підтримують соціальну інклюзивність, забезпечують рівний доступ до можливостей і враховують потреби всіх соціальних груп.

Висновки

Було розроблено метод нейромережевого формування репрезентативних недискримінаційних текстових датасетів, який відповідає принципам справедливості FATE та враховує етичні аспекти, такі як гендер, вік, релігія, етнічність тощо. Метод дозволяє коригувати датасети для забезпечення їх репрезентативності за етичними аспектами, що включає оптимізацію вибору елементів для видалення та аугментації даних відповідно до цих принципів. Для тестування методу було створено програмне забезпечення, яке використовує машинне навчання для класифікації текстів за різними етичними категоріями. Практичне застосування методу дозволило значно покращити репрезентативність датасетів за віковим та гендерним аспектами, з мінімальними відхиленнями від ідеального репрезентативного розподілу, що свідчить про високу ефективність методу.

Цей підхід також підтримує реалізацію Цілей сталого розвитку ПРООН, зокрема Ціль № 5, Ціль № 10 та Ціль № 11, забезпечуючи більш рівне та справедливе представлення різних соціальних груп у даних. Врахування цих цілей допомагає забезпечити, щоб моделі машинного навчання були етично збалансованими і неупередженими, що є важливим для створення інклюзивних та справедливих технологій в інформаційному середовищі.

Список використаної літератури

1. Собко О. В. Дослідження ефективності методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості. *Перспективи сучасної науки: теорія і практика: Матеріали VIII Міжнар. наук.-практ. конф.,* 2024. С. 217–221.

2. Krak I., Zalutska O., Molchanova M., Mazurets O., Bahrii R., Sobko O., Barmak O. Abusive Speech Detection Method for Ukrainian Language Used Recurrent Neural Network. *CEUR Workshop Proceedings. 2024. Vol. 3688*. С. 16–28.

3. Zalutska O., Molchanova M., Sobko O., Mazurets O., Pasichnyk O., Barmak O., Krak I. Method for sentiment analysis of Ukrainian-language reviews in e-commerce using RoBERTa neural network. *CEUR Workshop Proceedings. 2023. Vol. 3387.* С. 344–356.

4. Собко О. В. Метод інтелектуального пошуку та класифікації кіберзалякувань у текстовому контенті. *Інформаційні управляючі системи та технології ІУСТ-ОДЕСА-2024: Матеріали XII Міжнар. наук.-практ. конф. Одеса*, 2024. С. 262–265.

5. Jungwirth D., Haluza D. Artificial intelligence and the sustainable development goals: an exploratory study in the context of the society domain. *Journal of Software Engineering and Applications. 2023. Vol. 16, No. 4.* С. 91–112. https://doi.org/10.4236/jsea.2023.164006.

6. Matsui T., Suzuki K., Ando K., Kitai Y., Haga C., Masuhara N., Kawakubo S. *A natural language processing model for supporting sustainable development goals: translating semantics, visualizing nexus, and connecting stakeholders. Sustainability Science. 2022. Vol. 17, No. 3*. С. 969–985. https://doi.org/10.1007/s11625-022-01093-3.

7. Suzuki J., Zen H., Kazawa H. Extracting representative subset from extensive text data for training pre-trained language models. *Information Processing & Management. 2023. Vol. 60, No. 3.* С. 103249. https://doi.org/10.1016/j.ipm.2022.103249.

8. Zowghi D., Bano M. AI for all: Diversity and Inclusion in AI. *AI and Ethics.* 2024. С. 1–4. https://doi.org/10.1007/s43681-024-00485-8.

9. Dablain D., Krawczyk B., Chawla N. Towards a holistic view of bias in machine learning: Bridging algorithmic fairness and imbalanced learning. *arXiv preprint arXiv:2207.06084.* 2022. https://doi.org/10.48550/arXiv.2207.06084.

10. Kaggle.com. Cyberbullying Classification. 2021. URL: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification?resource=download (дата звернення: 24.11.2024).

11. Kaggle.com. CyberBullying Detection Dataset. 2024. URL: https://www.kaggle.com/datasets/sayankr007/cyberbullying-data-for-multi-label-classification (дата звернення: 24.11.2024).

References

1. Sobko, O. V. (2024). Doslidzhennia efektyvnosti metodu otsiniuvannia ta koryhuvannia reprezentatyvnosti datasetu za FATE-pryntsypom spravedlyvosti [Research on the efficiency of dataset representativeness evaluation and correction method based on the FATE fairness principle]. *In Perspektyvy suchasnoi nauky: teoriia i praktyka: Proceedings of the VIII International Scientific-Practical Conference* (pp. 217–221) [in Ukrainian].

2. Krak, I., Zalutska, O., Molchanova, M., Mazurets, O., Bahrii, R., Sobko, O., & Barmak, O. (2024). Abusive speech detection method for Ukrainian language used recurrent neural network. *CEUR Workshop Proceedings, 3688,* pp. 16–28.

3. Zalutska, O., Molchanova, M., Sobko, O., Mazurets, O., Pasichnyk, O., Barmak, O., & Krak, I. (2023). Method for sentiment analysis of Ukrainian-language reviews in e-commerce using RoBERTa neural network. *CEUR Workshop Proceedings, 3387*, pp. 344–356.

4. Sobko, O. V. (2024). Metod intelektualnoho poshuku ta klasyfikatsii kiberzaliakuvan u tekstovomu kontenti [Method of intelligent search and classification of cyberbullying in textual content]. *In Informatsiini upravliaiuchi systemy ta tekhnolohii IUST-Odesa-2024: Proceedings of the XII International Scientific-Practical Conference* (pp. 262–265) [in Ukrainian].

5. Jungwirth, D., & Haluza, D. (2023). Artificial intelligence and the sustainable development goals: An exploratory study in the context of the society domain. *Journal of Software Engineering and Applications, 16(4),* pp. 91–112. https://doi.org/10.4236/jsea.2023.164006

6. Matsui, T., Suzuki, K., Ando, K., Kitai, Y., Haga, C., Masuhara, N., & Kawakubo, S. (2022). A natural language processing model for supporting sustainable development goals: Translating semantics, visualizing nexus, and connecting stakeholders. *Sustainability Science, 17(3),* pp. 969–985. https://doi.org/10.1007/s11625-022-01093-3

7. Suzuki, J., Zen, H., & Kazawa, H. (2023). Extracting representative subset from extensive text data for training pre-trained language models. *Information Processing & Management, 60(3),* pp. 103249. https://doi.org/10.1016/j.ipm.2022.103249

8. Zowghi, D., & Bano, M. (2024). AI for all: Diversity and inclusion in AI. *AI and Ethics,* pp. 1–4. https://doi.org/10.1007/s43681-024-00485-8

9. Dablain, D., Krawczyk, B., & Chawla, N. (2022). Towards a holistic view of bias in machine learning: Bridging algorithmic fairness and imbalanced learning. *arXiv preprint arXiv:2207.06084*. https://doi.org/10.48550/arXiv.2207.06084

10. Kaggle.com. (2021). Cyberbullying Classification. https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification?resource=download (Accessed: November 24, 2024).

11. Kaggle.com. (2024). CyberBullying Detection Dataset. https://www.kaggle.com/datasets/sayankr007/cyberbullying-data-for-multi-label-classification (Accessed: November 24, 2024).