УДК 004.8
DOI https://doi.org/10.26661/2786-6254-2024-2-09

МЕТОД АНАЛІЗУ ТА ФОРМУВАННЯ РЕПРЕЗЕНТАТИВНИХ ВИБІРОК ТЕКСТОВИХ ДАНИХ ІЗ ВИКОРИСТАННЯМ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

Собко О. В.
викладач кафедри комп'ютерних наук
Хмельницький національний університет
вул. Інститутська, 11, Хмельницький, Україна
orcid.org/0000-0001-5371-5788
olena.sobko.ua@gmail.com

Бармак О. В.
доктор технічних наук, професор,
завідувач кафедри комп'ютерних наук
Хмельницький національний університет
вул. Інститутська, 11, Хмельницький, Україна
orcid.org/0000-0003-0739-9678
alexander.barmak@gmail.com

Ключові слова: *NLP, етична коректність даних, етичні принципи, недискримінація, репрезентативність текстових наборів даних.*

Розроблено метод аналізу та формування репрезентативних вибірок текстових даних, призначений для аналізу та формування репрезентативних текстових вибірок даних за принципом справедливості FATE для предметних областей. Метод виконує аналіз репрезентативності вибірки даних за етичними аспектами, за результатом чого виконується репрезентативне коригування датасету за етичними аспектами. При коригуванні датасету відбувається вирішення оптимізаційної задачі як для вибору надлишкових елементів для видалення, так і для формування вимог щодо приналежності за етичними аспектами до кожного елементу для аугментації даних.

Для дослідження ефективності методу аналізу та формування репрезентативного подання текстового датасету було створено програмне забезпечення, яке використовує моделі машинного навчання для класифікації текстів за різними етичними аспектами – віку, гендеру, релігії, етнічності тощо. Для класифікації текстових зразків за етичними аспектами у вибірці було використано моделі машинного навчання: за віковим етичним аспектом SVM, гендерним – LSTM, релігійним – BERT, які кращі показники статистичних метрик. В результаті практичного застосування розробленого методу тестовий нерепрезентативний порівняно з об'єктивними даними демографічної статистики датасет було трансформовано у репрезентативний за віковим та гендерним етичними аспектами. Одержані відхилення розподілів зразків за класами етичних аспектів датасету, трансформованого за створеним методом, від ідеального репрезентативного розподілу склали: мінімальне – 0.00%, максимальне – 0.04%, середнє – 0.02%, за умов початкового обсягу датасету 47 692 елементів, мінімальної початкової кількості зразків у класі 1007 елементів, максимальної початкової кількості зразків у класі 28 112 елементів. Досліджена ефективність доводить, що розроблений метод дозволяє виконувати аналіз репрезентативності текстових датасетів та приведення їх до репрезентативного вигляду за різними аспектами принципу справедливості FATE.

METHOD OF ANALYSIS AND FORMATION OF REPRESENTATIVE SETS OF TEXT DATA USING MACHINE LEARNING MODELS

Sobko O. V.
Lecturer at the Department of Computer Sciences
Khmelnytskyi National University
Institutska str., 11, Khmelnytskyi, Ukraine
orcid.org/0000-0001-5371-5788
olena.sobko.ua@gmail.com

Barmak O. V.
Doctor of Engineering Sciences, Professor,
Head of the Department of Computer Sciences
Khmelnytskyi National University
Institutska str., 11, Khmelnytskyi, Ukraine
orcid.org/0000-0003-0739-9678
alexander.barmak@gmail.com

Key words: *NLP, data ethical correctness, ethical principles, non-discrimination, text datasets representative.*

The method of analysis and formation of representative sets of text data using machine learning models was developed, intended for analysis and formation of representative text samples of data according to the principle of fairness of FATE for subject areas. The method performs an analysis of representativeness of data sample according to ethical aspects, as result of which a representative adjustment of the dataset according to ethical aspects is performed. When adjusting the dataset, the optimization problem is solved both for the selection of redundant elements for removal, and for formation of requirements for ethical aspects of belonging to each element for data augmentation.

To investigate the effectiveness of analysis method and the formation of a representative presentation of the text dataset, software was created that uses machine learning models to classify texts according to various ethical aspects - age, gender, religion, ethnicity, etc. Machine learning models were used to classify text samples by ethical aspects in the sample: by age ethical aspect SVM, gender – LSTM, religious – BERT, which are the best indicators of statistical metrics. As a result of the practical application of the developed method, the test dataset, unrepresentative compared to the objective data of demographic statistics, was transformed into a representative one in terms of age and gender ethical aspects. The obtained deviations of the sample distributions by classes of ethical aspects of the dataset transformed according to the created method from the ideal representative distribution were: minimum 0.00%, maximum 0.04%, average 0.02%, under the conditions of the initial volume of the dataset 47,692 elements, the minimum initial number of samples in the class 1007 elements, the maximum initial number of samples in the class is 28,112 elements. The studied efficiency proves that developed method allows performing the analysis of the representativeness of text datasets and bringing them to a representative look according to various aspects of FATE principle of justice.

Вступ. У сучасному світі активно розробляються численні рішення з використанням штучного інтелекту, покликані вирішувати різноманітні завдання, з якими люди стикаються щодня. Відповідно, результати, що генеруються штучним інтелектом, залежать від навчальних датасетів, на яких вони навчалися, іншими словами – вміст цих датасетів безпосередньо впливає на кінцевий результат. Відсутність прозорості щодо джерел і характеристик даних, які використовуються для навчання алгоритмів штучного інтелекту, зменшує довіру до отриманих результатів. В такому випадку часто користувачі не можуть оцінити потенційні упередження чи дискримінаційні елементи, вбудовані у ці алгоритми. Недостатня інформованість про вміст навчальних датасетів збільшує ризик поширення несправедливих або неточних рішень, які можуть мати серйозні наслідки для окремих осіб та суспільства в цілому [1].

Засоби для оцінювання репрезентативності текстового набору даних відповідно до принципів етичної недискримінації є наразі відсутніми. Відомі датасети для навчання нейромереж, наприклад [2] та [3], активно використовуються дослідниками, адже мають великий обсяг даних, проте вони не валідувались авторами щодо репрезентативності за принципом справедливості, а отже, використання таких датасетів для навчання алгоритмів штучного інтелекту можуть потенційно порушувати етичні принципи та, звідси, мати низьку достовірність прийнятих рішень.

Репрезентативність даних у датасетах не лише впливає на точність результатів та моделей, але й тісно пов'язана з принципами FATE (Fairness, Accountability, Transparency, Ethics) у використанні даних і розробці технологій штучного інтелекту. Якщо датасет не включає належного представлення всіх соціальних, демографічних або культурних груп, це може призвести до дискримінаційних моделей, які надають пріоритет одній групі над іншою, тобто не є справедливими. Репрезентативність датасетів за етичним принципом FATE може бути досягнута шляхом коректного балансування за різними етичними аспектами: расового, гендерного, релігійного, вікового тощо [4].

Основним внеском статті є розробка й апробація підходу до аналізу та формування репрезентативних текстових вибірок даних за принципом справедливості FATE для предметних областей.

Огляд літератури. Дослідженню репрезентативності текстових вибірок та справедливому і неупередженому представленню демографічних груп у них присвячено багато робіт, оскільки поняття репрезентативності, справедливості та неупередженості є важливими у створенні етичних і справедливих моделей машинного навчання [5].

Так, у [6] автори піднімають важливу проблему репрезентативності вибірок у контексті машинного навчання та штучного інтелекту, акцентуючи увагу на необхідності точного відображення популяційних даних. Основною стратегією, яку автори пропонують для досягнення високої якості моделей, є використання стратифікованих вибірок, що дозволяють зменшити варіативність між підгрупами та точно відобразити пропорції між різними категоріями у популяції.

Автори [7] розглядають упередження, що виникають як через дисбаланс класів у даних, так і через чутливі (захищені) ознаки, такі як раса чи стать. Автори пропонують новий метод, Fair Oversampling, який поєднує популярний метод для роботи з дисбалансом даних SMOTE із модифікаціями, що допомагають знизити вплив чутливих ознак. Підхід збільшує точність моделі за рахунок балансування класів і зменшує залежність від чутливих ознак, що покращує групову справедливість.

У [8] розглядається проблема гендерної упередженості в моделях обробки природної мови, вирішуючи її за допомогою двох основних підходів: статистичного та каузального забезпечення справедливості. Дослідники застосовують такі техніки, як counterfactual data augmentation для каузального дебіасингу, а також методи ресемплінгу та ревагінгу для статистичного дебіасингу. Результати показали, що поєднання цих технік дозволяє значно упередження у моделях як за статистичними, так і за каузальними метриками.

У наведених роботах показано, що формування репрезентативних та неупереджених вибірок є актуальним напрямком дослідження, проте більшість робіт присвячено або виявленню неупередженості, або аналізу репрезентативності або неупередженості вибірок даних, однак вибірки даних повинні бути модифіковані для досягнення відповідності FATE-принципам [9]. Узагальнюючи, можна виділити особливості сучасного підходу, який застосовується до розробки моделей ШІ (Рис.1). Однак такий підхід не враховує існуючі етичні принципи та недискримінаційне, репрезентативне подання існуючих підгруп популяції, які повинні застосовуватись для отримання моделей ШІ.

Метою роботи є забезпечення дотримання етичних аспектів (гендерного, релігійного, вікового тощо) принципу справедливості FATE [4] для навчальних датасетів, яке полягає у створенні методу аналізу та формування репрезентативних (за означеними аспектами) текстових вибірок даних.

Для досягнення означеної мети, потрібно запропонувати метод, який реалізовуватиме наступні завдання дослідження:

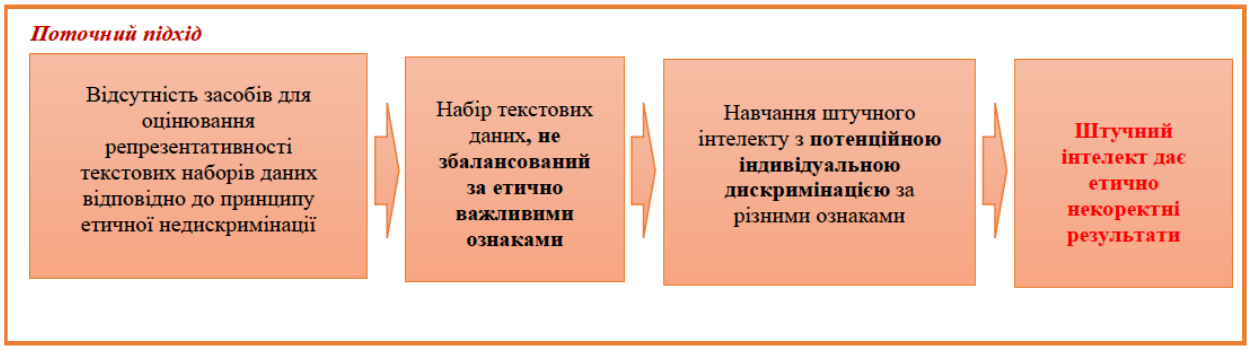

Рис. 1. Існуючий підхід до навчання моделей ШІ

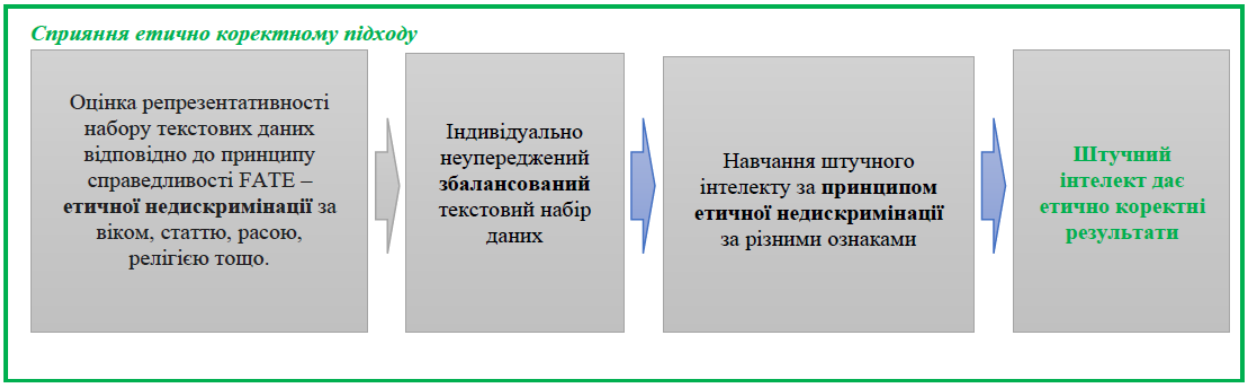

Рис. 2. Запропонований підхід до формування репрезентативних за етичними принципами датасетів

1) розробити підхід до аналізу та формування відповідних репрезентативних датасетів за принципом справедливості FATE для предметних областей;

2) дослідити ефективність запропонованого підходу, шляхом використання його для прикладного аналізу текстового датасету та приведення його до репрезентативного вигляду за аспектами принципу справедливості FATE: гендеру, віку й релігії.

Метод аналізу та формування репрезентативних вибірок текстових даних. На противагу існуючому підходу до навчання моделей ШІ (див. Рис. 1) у дослідженні запропоновано новий підхід (Рис. 2), який забезпечить репрезентативність та етичну коректність датасетів, які використовуються для навчання моделей ШІ.

Проблему одержання репрезентативного, неупередженого за етичними принципами текстового датасету можна подати у рамках інформаційної моделі наступного вигляду:

$$\{D, D', C, A, M, F\}, \qquad (1)$$

де D – текстовий датасет для аналізу та коригування, D' – текстовий датасет після коригування, C – множина класів предметної області датасету (наприклад, види кібербулінгу), A – множина етичних аспектів, M – множина навчених моделей машинного навчання (окрема для кожного етичного аспекту), F – цільова функція мінімізації відхилення між поточними та бажаними співвідношеннями для всіх етичних аспектів.

У дослідженні пропонується звести задачу побудови репрезентативного, неупередженого за етичними принципами датасету до задачі багатокритеріальної оптимізації. Задача оптимізації полягає у мінімізації відхилення між поточними та бажаними співвідношеннями класів, враховуючи обмеження на кількість зразків у класах і можливостей генерації синтетичних даних.

Вхідні дані: текстовий датасет D, множина етичних аспектів A, вимоги до репрезентативного розподілу D'.

Мета задачі: створення репрезентативної вибірки за всіма етичними аспектами, яка досягає цільових пропорцій класів для кожного етичного аспекту $D \Rightarrow D'$.

Змінні: x_{ij} – кількість зразків класу C_j в аспекті A_i після секвестрування та аугментації.

Цільовою функцією F є мінімізація відхилення між поточними та бажаними співвідношеннями для всіх етичних аспектів одночасно з урахуванням обмежень (3) – (6):

$$F = argmin\sum_{i-1}^{m}\sum_{j-1}^{n_i}\left|\frac{x_{ij}}{n'} - T_{ij}\right|. \quad (2)$$

Обмеження задачі:

1) сума всіх зразків класів в межах одного аспекту дорівнює цільовій кількості зразків для цього аспекту (4):

$$\sum_{j=1}^{n_i} x_{ij} = n', \forall i \in \{1,2,\ldots,m\}, \quad (3)$$

де n_i – кількість класів в аспекті A_i;

2) кількість зразків для кожного класу повинна відповідати цільовій пропорції класів:

$$\frac{x_{ij}}{n'} \approx T_{ij}, \forall i \in \{1,2,\ldots,m\}, \forall j \in \{1,2,\ldots,n_i\}; \quad (4)$$

3) розрахункова кількість зразків не може бути від'ємною:

$$x_{ij} \geq 0, \forall i \in \{1,2,\ldots,m\}, \forall j \in \{1,2,\ldots,n_i\}; \quad (5)$$

4) можливість додавання нових зразків повинна відповідати можливостям генерації нових даних для кожного класу та аспекту:

$$x_{ij} \leq g_{ij}, \forall i \in \{1,2,\ldots,m\}, \forall j \in \{1,2,\ldots,n_i\}, \quad (6)$$

де g_{ij} – максимально можлива кількість зразків класу C_j в аспекті A_i, яку можна додати.

Виходячи з поставленої оптимізаційної задачі формування репрезентативного датасету (2), наведемо кроки методу аналізу та формування репрезентативних вибірок текстових даних.

Метод аналізу та формування репрезентативних вибірок текстових даних із використанням моделей машинного навчання подаємо 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 3.

Вхідними даними методу є текстовий датасет для аналізу та коригування D, множина етичних аспектів A, множина навчених моделей ML M для етичних аспектів A, вимоги до датасету D′ (обсяг, пропорції за етичними аспектами).

На етапі 1 здійснюється перевірка коректності елементів датасету, а саме першим кроком видалення неінформативних фрагментів елементів та другим кроком видалення некоректних елементів у D.

На 2 етапі відбувається аналіз репрезентативності за етичними аспектами. На першому кроці цього етапу відбувається векторизація кожного ∀d ∈ D за кожною з моделей ∀m ∈ M. На

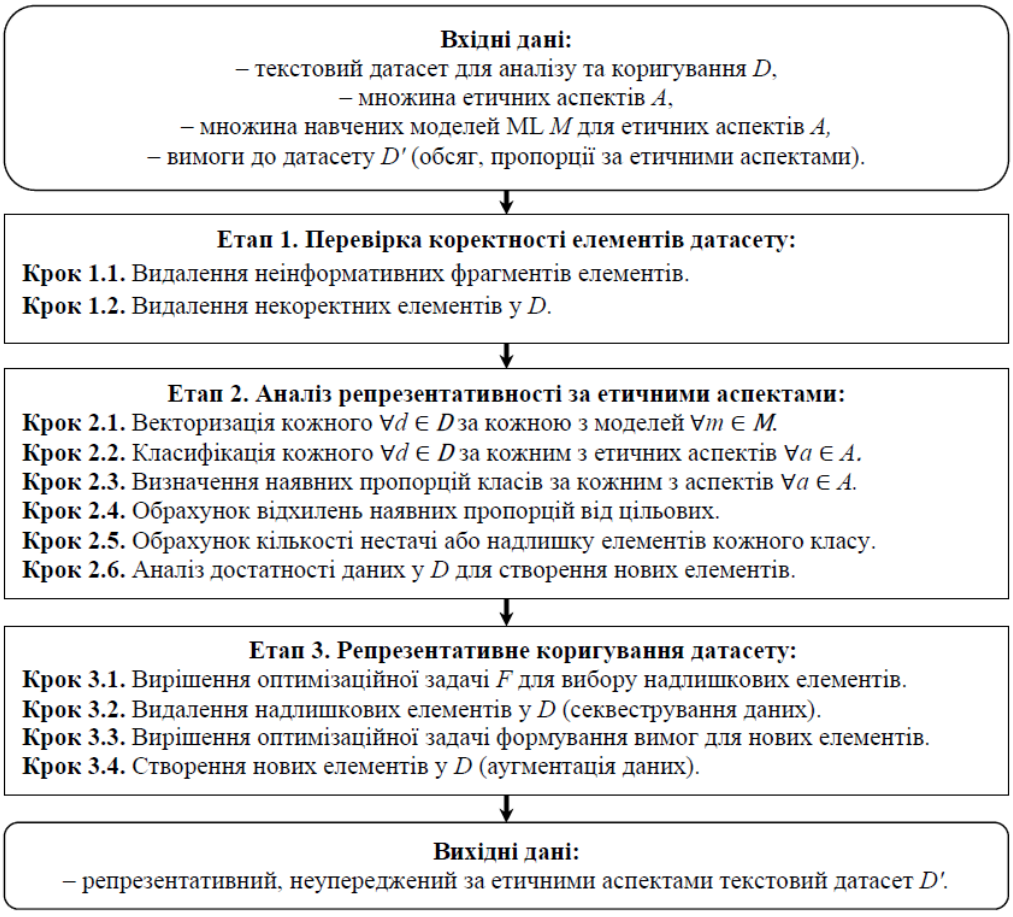

Рис. 3. Кроки of method for analysis and formation of representative text datasets

другому кроці класифікація кожного ∀d ∈ D за кожним з етичних аспектів ∀a ∈ A. На третьому кроці визначаються наявні пропорції класів за кожним із аспектів ∀a ∈ A. Далі на кроці чотири обраховуються відхилення наявних пропорцій від цільових, а також на п'ятому кроці обраховуються кількість нестачі або надлишку елементів кожного класу. Завершальним на цьому етапі є крок 6, на якому аналізується достатність даних у D для створення нових елементів.

Крок 3 передбачає репрезентативне коригування вибірки даних з урахуванням етичних аспектів. Коригуванням є видалення та додавання.

Операція секвестрування виконується для видалення надлишкових елементів кожного класу за кожним з етичних аспектів з мінімальною шкодою для інших розподілів, для чого вирішується оптимізаційна задача вибору надлишкових елементів в рамках (2), які мають бути видалені для досягнення цільових пропорцій класів.

Операція аугментації виконується для створення нових елементів за допомогою одного з відомих способів, наприклад за методикою SMOTE. Створюються вимоги в вигляді потрібної комбінації класів кожного з етичних аспектів для кожного нового елементу, для чого вирішується оптимізаційна задача формування вимог до відсутніх елементів в рамках (2).

Вихідними даними методу є репрезентативний, неупереджений за етичними аспектами текстовий датасет D′.

Виконання кроків методу аналізу та формування репрезентативних вибірок текстових даних дозволить формувати текстові вибірки, які є недискримінаційними та неупередженими та відображають пропорційне до реальних демографічних підгруп популяції представлення зразків вибірки, що впливатиме на точність та прозорість навчання моделей машинного навчання для вирішення різноманітних задач.

Для формування множини навчених моделей машинного навчання, які є окремими для кожного етичного аспекту, необхідно навчити кожну модель класифікатора, що аналізуватиме репрезентативність вхідної текстової вибірки згідно кроку 2 на Рис. 3. Для отримання таких класифікаторів, що і будуть формувати множину навчених етичних моделей машинного навчання, необхідно виконати кроки, що подані на Рис 4.

Першим кроком є вибір моделі ML для класифікації текстів за етичними аспектами. Для таких цілей використовуються як моделі глибокого навчання, наприклад, BERT, GPT, LSTM, GRU, так і класифікатори, наприклад, Logistic Regression, Naive Bayes, Support Vector Machines k-Nearest Neighbors тощо. Після цього наступним кроком відбувається навчання класифікатора за вибраною моделлю ML на анотованому датасеті для етичного аспекту.

Останнім кроком є аналіз якості отриманого класифікатора за статистичними показниками, такими як Accuracy, Precision, Recall та F1-score [10], якщо якість моделі згідно отриманих показників незадовільна, то необхідно повернутися до кроку вибору моделі ML, в іншому випадку – отримано класифікатор для аналізу датасету на репрезентативність за етичним аспектом, який розглядається.

Таким чином формується така кількість моделей машинного навчання для множини EML, яка відповідає кількості обраних етичних аспектів для аналізу та формування репрезентативної вибірки текстових даних.

Для апробації методу аналізу та формування репрезентативних вибірки текстових даних сформовано вхідний датасет на основі двох датасетів «Cyberbullying Classification» [2] та «Cyberbully Detection Dataset» [3]. Датасет «Cyberbullying Classification» містить 46017 твітів, які промарковані за видами кіберзалякувань на 6 класів. Датасет «Cyberbully Detection Dataset» містить 99989 твітів, який також промаркований за видами кіберзалякувань. Обидва датасети не містять міток щодо статі, вікової категорії, релігії та етнічності автора повідомлення.

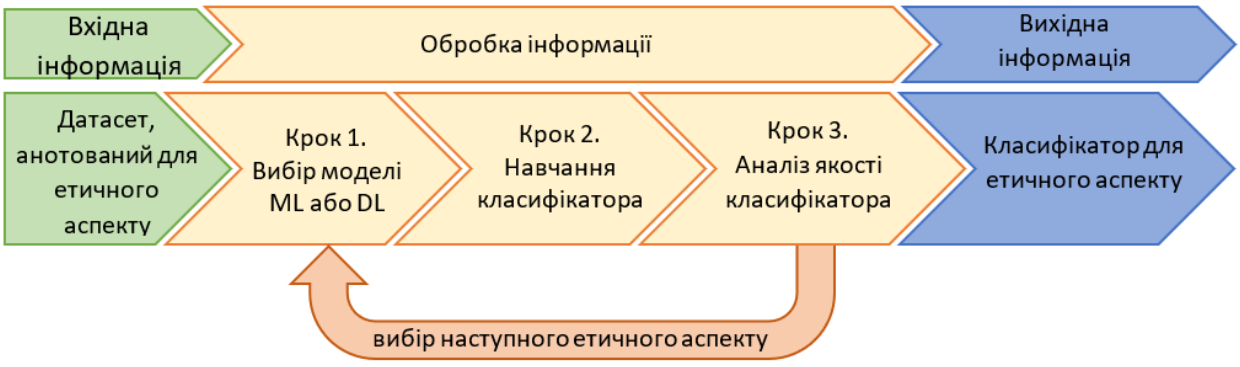

Рис. 4. Кроки для отримання моделей класифікаторів ML для етичних аспектів

Для навчання моделей машинного навчання, які використовуватимуться для розмітки вхідного датасету використано датасатети на прикладі трьох етичних аспектів принципу справедливості: гендеру [11] (34146 унікальних текстових записів), віку [12] (20109 унікальних текстових записів) та релігії [13] (21948 унікальних текстових записів). Так як класи в наведених датасетах не збалансовані та мають різну кількість зразків, що негативно впливатиме на якість навчання моделей машинного навчання, то усі класи у датасетах були збалансовані за кількістю. Остаточну кількість зразків у кожному класі навчальних вибірок для навчання EML за етичними аспектами показано на Рис. 5.

В результаті роботи зі створення навчальних вибірок, отримано збалансовані за кількістю текстових повідомлень у класах датасети. Такі датасети дозволять коректно оцінювати репрезентативність робочих текстових датасетів.

Результати та обговорення. Для аналізу та формування репрезентативної вибірки текстових даних за цільові пропорції класів для формування репрезентативної вибірки текстових даних за віком та статтю взято популяцію України. За даними Інституту демографії та соціальних досліджень імені М. В. Птухи Національної академії наук України (https://idss.org.ua/forecasts/nation_pop_proj), станом на липень 2023 року загальна чисельність населення України оцінюється в 35596216 осіб. У кожній віковій підгрупі представлено наступну кількість осіб: вікова група 0-19 років 6 659 068 осіб, 20-29 років 3 623 143 осіб, 30-39 років – 6 022 345 осіб, 40-49 років – 5 431 140 осіб, 50-100 років – 13 860 520 осіб. Щодо гендерної структури населення України на 2023 рік: 16951527 – жінки, а 18644689 – чоловіки (idss.org.ua). Зауважимо, що в межах цієї роботи при аналізі гендерного етичного аспекту розглядається цисгендерна група.

Для дослідження ефективності описаного в роботі методу аналізу та формування репрезентативної вибір текстових даних навчено декілька моделей машинного навчання. Результати обчислення статичних метрик як Accuracy, Precision, Recall та F1-score [10] моделей машинного навчання для гендерного, вікового та релігійного етичних аспектів наведено в таблиці 1.

Для різних класів було отримано різні рівні лінійної роздільної здатності [14]: за релігійною ознакою з використанням класифікатора BERT, який показав найкращий результат з навчених

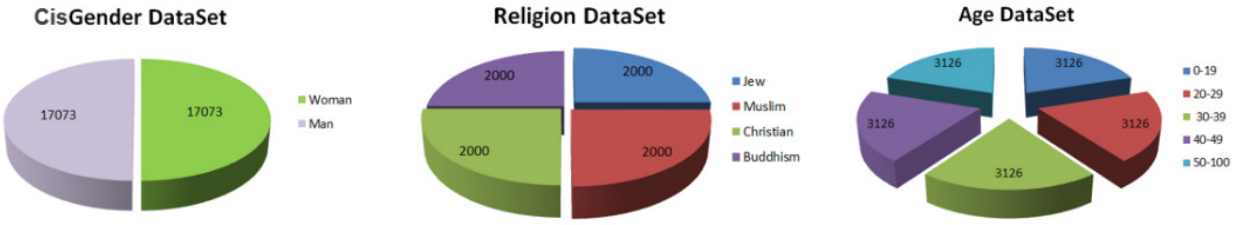

Рис. 5. Класи та кількість зразків у датасетах для навчання EML за етичними аспектами

Таблиця 1

Статистичні метрики Accuracy, Precision, Recall та F1-score моделей машинного навчання

Модель машинного навчання	*Accuracy*	*Precision*	*Recall*	*F1-score*
		Гендерний етичний аспект		
FastForest	0.630	0.640	0.600	0.620
SVM	0.580	0.580	0.580	0.580
LSTM	0.70	0.770	0.670	0.720
BERT	0.690	0.640	0.710	0.670
		Віковий етичний аспект		
FastForest	0.535	0.542	0.504	0.504
SVM	0.815	0.770	0.779	0.770
LSTM	0.590	0.600	0.560	0.580
BERT	0.580	0.430	0.450	0.440
		Релігійний етичний аспект		
FastForest	0.775	0.800	0.762	0.780
SVM	0.825	0.850	0.810	0.829
LSTM	0.850	0.880	0.830	0.854
BERT	0.910	0.980	0.74 0	0.840

моделей машинного навчання для задачі класифікації текстових зразків за релігійним етичним аспектом, дані виявились добре роздільні, за гендерною ознакою з використанням класифікатора LSTM, який показав найкращу ефективність порівняно з іншими моделями, дані виявились середньо роздільні та за віковою ознакою з використанням класифікатора SVM – погано роздільні [15].

Окрім того, виявлено що датасет не є репрезентативним, адже класи різних етичних аспектів мають кількість текстових зразків, що не відповідає пропорціям демографічних підгруп населення України, таким чином потребують збалансовування для набуття репрезентативного вигляду. Тому, згідно з кроками методу аналізу та формування репрезентативних вибірки текстових даних, вибірка текстових даних потребує аугментації даних для формування репрезентативної вибірки. Для цього необхідно вирішити оптимізаційну задачу (2) для коректного видалення надлишкових елементів кожного класу за кожним з етичних аспектів з подальшою аугментацією вибірки даних до цільових вимог (кількість елементів та пропорції класів).

У таблиці 2 подано відсоткові відношення зразків за віком у вибірці текстових даних та осіб популяції у вікових демографічних підгрупах, а також обчислено новий розподіл класів вибірки, якби враховувався лише один етичний аспект – віковий.

Після трансформації датасету за розробленим методом одержано відхилення розподілів зразків за класами вікового етичного аспекту датасету, трансформованого за створеним методом, від ідеального репрезентативного розподілу склали: мінімальне – 0.01%, максимальне – 0.04%, середнє – 0.02%, а для гендерного етичного аспекту: мінімальне – 0.03%, максимальне – 0.03%, середнє – 0.03%.

Проте оптимізаційна задача з формування репрезентативної вибірки текстових даних є багатокритеріальна, критеріями в якій є формування вибірки за віковим та гендерним етичним аспектом, тому метою є мінімізація відхилення між поточними та бажаними співвідношеннями класів, враховуючи обмеження на кількість зразків і можливості генерації нових даних. В результаті вирішення оптимізаційної задачі для формування репрезентативної вибірки за віковим та гендерним етичними аспектами на прикладі демографічних підгруп популяції України отримано шляхом аугментації репрезентативну вибірку текстових даних, баланс класів якої подано у таблиці 2.

Отримано відхилення розподілів зразків за класами вікового та гендерного етичних аспектів датасету одночасно, трансформованого за створеним методом, від ідеального репрезентативного розподілу склали: мінімальне 0.00%, максимальне 0.04%, середнє 0.02%.

Отже, в результаті виконання кроків методу аналізу та формування репрезентативних вибірок текстових даних сформовано датасет, який є недискримінаційним і неупередженим й відображає пропорційне до реальних демографічних підгруп популяції України представлення зразків вибірки.

Висновки. Було розроблено метод аналізу та формування репрезентативних вибірок текстових даних, призначений для аналізу та формування репрезентативних текстових вибірок даних за принципом справедливості FATE для предметних областей. Метод виконує аналіз репрезентативності вибірки даних за етичними аспектами, за результатом чого виконується репрезентативне коригування датасету за етичними аспектами. При коригуванні датасету відбувається вирішення оптимізаційної задачі як для вибору надлишкових елементів для видалення, так і для формування вимог щодо приналежності за етичними аспектами до кожного елементу для аугментації даних.

Для дослідження ефективності методу аналізу та формування репрезентативного подання текстового датасету було створено програмне забезпечення, яке використовує моделі машинного

Таблиця 2

Розподіл зразків у сформованій репрезентативній вибірці після аугментації даних в результаті розв'язку багатокритеріальної оптимізаційної задачі

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%

навчання для класифікації текстів за різними етичними аспектами – віку, гендеру, релігії, етнічності тощо. Так, для класифікації текстових зразків у вибірці за віковим етичним аспектом використано SVM, гендерним – LSTM, релігійним – BERT, які кращі показники статистичних метрик. В результаті практичного застосування розробленого методу було трансформовано датасет у репрезентативний за віковим та гендерним етичними аспектами. Одержані відхилення розподілів зразків за класами етичних аспектів датасету, трансформованого за створеним методом, від ідеального репрезентативного розподілу склали: мінімальне – 0.00%, максимальне – 0.04%, середнє – 0.02%. Досліджена ефективність доводить, що розроблений метод дозволяє виконувати аналіз репрезентативності текстових датасетів та приведення їх до репрезентативного вигляду за різними аспектами принципу справедливості FATE.

Подальшими планами щодо покращення методу аналізу та формування репрезентативних вибірок текстових даних є формування не тільки недискримінаційної вибірки за кількістю зразків, а й пошук та видалення у вибірках текстових зразків, що містять упереджене ставлення до представників різних демографічних підгруп, відповідно до етичних аспектів FATE-принципу справедливості.

ЛІТЕРАТУРА

1. Shah M., Sureja N. A Comprehensive Review of Bias in Deep Learning Models: Methods, Impacts, and Future Directions. *Arch Computat Methods Eng.* 2024. DOI: https://doi.org/10.1007/s11831-024-10134-2.
2. Kaggle.com. Cyberbullying Classification, 2021. URL: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification?resource=download (дата звернення: 26.10.2024).
3. Kaggle.com. CyberBullying Detection Dataset, 2024. URL: https://www.kaggle.com/datasets/sayankr007/cyber-bullying-data-for-multi-label-classification (дата звернення: 26.10.2024).
4. Memarian B., Doleck T. Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and Higher Education: A Systematic Review. *Computers and Education: Artificial Intelligence*. 2023, Vol. 5. DOI: https://doi.org/10.1016/j.caeai.2023.100152.
5. Manziuk E., Krak I., Barmak O., Mazurets O., Kuznetsov V., Pylypiak O. Structural Alignment Method of Conceptual Categories of Ontology and Formalized Domain. *CEUR Workshop Proceedings*. 2021, Vol. 3003, pp. 11–22.
6. Clemmensen L. K. H., Rune D. K. Data Representativity for Machine Learning and AI Systems. 2022. URL: https://ar5iv.labs.arxiv.org/html/2203.04706 (дата звернення: 26.10.2024).
7. Dablain D., Krawczyk B., Chawla N. Towards a Holistic View of Bias in Machine Learning: Bridging Algorithmic Fairness and Imbalanced Learning. *Discov Data*. 2024, Vol. 2(4). DOI: https://doi.org/10.1007/s44248-024-00007-1.
8. Chen H., Ji Y., Evans D. Addressing Both Statistical and Causal Gender Fairness in NLP Models. *In Findings of the Association for Computational Linguistics: NAACL 2024.* 2024, pp. 561–582. DOI: https://doi.org/10.48550/arXiv.2404.00463.
9. 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*. 2024. № 2(333). С. 200–206. DOI: https://doi.org/10.31891/2307-5732-2024-333-2-32.
10. Rainio O., Teuho J., Klén R. Evaluation Metrics and Statistical Tests for Machine Learning. *Scientific Reports*. 2024, Vol. 14(1). DOI: https://doi.org/10.1038/s41598-024-56706-x.
11. Kaggle.com. Tweet Files for Gender Guessing, 2019. URL: https://www.kaggle.com/datasets/aharless/tweet-files-for-gender-guessing (дата звернення: 26.10.2024).
12. Kaggle.com. CyberBullying Detection Dataset, 2024. URL: https://www.kaggle.com/datasets/sayankr007/cyber-bullying-data-for-multi-label-classification (дата звернення: 26.10.2024).
13. Live.european-language-grid.eu. TAG-it Dataset Distribution, 2024. URL: https://live.european-language-grid.eu/catalogue/corpus/8112/download (дата звернення: 26.10.2024).
14. 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, pp. 16–28.
15. Slobodzian V., Kovalchuk O., Molchanova M., Sobko O., Mazurets O., Barmak O., Krak I. Text Data Vectorization Model of Ukrainian-Language Internet Communication Content. *CEUR Workshop Proceedings*. 2022. Vol. 3171, pp. 561–571.

REFERENCES

1. Shah M., Sureja N. A Comprehensive Review of Bias in Deep Learning Models: Methods, Impacts, and Future Directions. *Arch Computat Methods Eng.* 2024. DOI: https://doi.org/10.1007/s11831-024-10134-2
2. Kaggle.com. Cyberbullying Classification, 2021. URL: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification?resource=download (accessed: 26.10.2024)
3. Kaggle.com. CyberBullying Detection Dataset, 2024. URL: https://www.kaggle.com/datasets/sayankr007/cyber-bullying-data-for-multi-label-classification (accessed: 26.10.2024)
4. Memarian B., Doleck T. Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and Higher Education: A Systematic Review. *Computers and Education: Artificial Intelligence*. 2023, Vol. 5. DOI: https://doi.org/10.1016/j.caeai.2023.100152
5. Manziuk E., Krak I., Barmak O., Mazurets O., Kuznetsov V., Pylypiak O. Structural Alignment Method of Conceptual Categories of Ontology and Formalized Domain. *CEUR Workshop Proceedings*. 2021, Vol. 3003, pp. 11–22.
6. Clemmensen L. K. H., Rune D. K. Data Representativity for Machine Learning and AI Systems. 2022. URL: https://ar5iv.labs.arxiv.org/html/2203.04706 (accessed: 26.10.2024)
7. Dablain D., Krawczyk B., Chawla N. Towards a Holistic View of Bias in Machine Learning: Bridging Algorithmic Fairness and Imbalanced Learning. *Discov Data*. 2024, Vol. 2(4). DOI: https://doi.org/10.1007/s44248-024-00007-1
8. Chen H., Ji Y., Evans D. Addressing Both Statistical and Causal Gender Fairness in NLP Models. *In Findings of the Association for Computational Linguistics: NAACL 2024.* 2024, pp. 561–582. DOI: https://doi.org/10.48550/arXiv.2404.00463
9. Molchanova M.O., Mazurets O.V., Sobko O.V., Klimenko V.I., Androshchuk V.I. Metod neiromerezhevoho vyiavlennia kiberbulinhu z vykorystanniam khmarnykh servisiv ta obiektno-oriientovanoi modeli. *Naukovyi zhurnal «Visnyk Khmelnytskoho natsionalnoho universytetu», seriia: Tekhnichni nauky*. 2024. №2 (333). S. 200–206. DOI: https://doi.org/10.31891/2307-5732-2024-333-2-32. (in Ukrainian)
10. Rainio O., Teuho J., Klén R. Evaluation Metrics and Statistical Tests for Machine Learning. *Scientific Reports*. 2024, Vol. 14(1). DOI: https://doi.org/10.1038/s41598-024-56706-x
11. Kaggle.com. Tweet Files for Gender Guessing, 2019. URL: https://www.kaggle.com/datasets/aharless/tweet-files-for-gender-guessing (accessed 26.10.2024)
12. Kaggle.com. CyberBullying Detection Dataset, 2024. URL: https://www.kaggle.com/datasets/sayankr007/cyber-bullying-data-for-multi-label-classification (accessed: 26.10.2024)
13. Live.european-language-grid.eu. TAG-it Dataset Distribution, 2024. URL: https://live.european-language-grid.eu/catalogue/corpus/8112/download (accessed: 26.10.2024)
14. 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, pp. 16–28.
15. Slobodzian V., Kovalchuk O., Molchanova M., Sobko O., Mazurets O., Barmak O., Krak I. Text Data Vectorization Model of Ukrainian-Language Internet Communication Content. *CEUR Workshop Proceedings*. 2022. Vol. 3171, pp. 561–571.