ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ МЕТОДУ ОЦІНЮВАННЯ ТА КОРИГУВАННЯ РЕПРЕЗЕНТАТИВНОСТІ ДАТАСЕТУ ЗА FATE-ПРИНЦИПОМ СПРАВЕДЛИВОСТІ

Собко Олена Віталіївна
аспірант
Хмельницький національний університет, Україна

Вступ. Штучний інтелект сьогодні активно розвивається для вирішення різних завдань, з якими люди стикаються щодня. Проте його результати залежать від навчальних датасетів, які не завжди створюються з урахуванням етичних норм. Часто такі набори даних можуть бути нерепрезентативними, що призводить до дискримінаційних результатів, наприклад, під час виявлення кіберзалякувань. Відсутність прозорості щодо змісту датасетів також підриває довіру до результатів, адже користувачі не можуть оцінити можливі упередження. Щоб уникнути цього, необхідно впроваджувати практики прозорості та підзвітності на всіх етапах розробки ШІ.

Наразі дослідницькі групи розробляють етичні стандарти для використання ШІ, такі як FATE-принципи (справедливість, підзвітність, прозорість, етика). Справедливість вимагає, щоб усі групи були належно представлені у навчальних даних, щоб уникнути дискримінації [1-3].

Метою роботи є дослідження ефективності методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості.

Матеріали та методи. Для обробки природної мови необхідні великі масиви текстових даних, які збираються з різних джерел, як-от соціальні мережі, інтернет чи наукові публікації. Важливим етапом є підготовка цих даних: очищення від некоректної інформації, такої як HTML-коди, та видалення дуже коротких текстів. Також потрібно забезпечити рівномірний розподіл даних по різних категоріях, щоб модель працювала без упереджень.

Існують датасети [4, 5], які використовуються для виявлення кіберзалякувань у тексті, але вони часто не є репрезентативними з етичної

точки зору. Зокрема, це стосується FATE-принципу справедливості, який вимагає рівного представлення різних соціальних груп у даних. Нерепрезентативні датасети можуть містити упередженість щодо статі, раси, віку чи інших характеристик, що призводить до неточних або дискримінаційних результатів при виявленні кіберзалякувань.

Відомі декілька досліджень, у яких застосовуються пояснений та етичний штучний інтелект у вирішенні задачі виявлення кіберзалякувань у текстових повідомленнях [6, 7]. У цих дослідженнях наведені розробниками моделі не тільки виявляють образливий контент, але й пояснюють, як вони дійшли до такого висновку, що підвищує довіру до результатів. Крім того, дослідники звертають увагу на етичні аспекти, зокрема на мінімізацію упередженості в класифікації, щоб уникнути дискримінації певних груп користувачів.

FATE-принцип справедливості передбачає, що розроблена модель для виявлення кіберзалякувань у текстових повідомленнях повинна бути безупередженою і не дискримінувати жодну групу осіб на основі таких ознак, як релігія, гендер, раса чи вік, тощо, що забезпечується шляхом належного представлення всіх груп у даних для навчання моделі.

Розроблений метод формування репрезентативного оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості може мати таку кількість кроків, яка відповідає кількості ознак принципу справедливості. Далі представлені кроки методу на основі трьох ознак_релігійної, вікової та гендерної. На першому кроці використовуються два набори даних: робочий набір даних №1 для виявлення кіберзалякувань, позначений за видами кіберзалякувань, і набір даних №2, класифікований за релігійною ознакою. Нейромережева модель навчається на основі набору даних №2 для оцінки та коригування репрезентативності робочого набору даних за релігійною ознакою відповідно до принципу справедливості FATE. Після цього етапу робочого набору даних коригується за релігійною ознакою, щоб кількість зразків була рівною для всіх релігійних класів.

На другому кроці модифікований робочий набір даних №1 (з доданою

релігійною класифікацією) та набір даних №3, класифікований за гендерною ознакою, використовуються для навчання моделі, що дозволяє оцінити та скоригувати робочий набір даних за гендерними ознаками, щоб забезпечити рівне представлення для всіх гендерних класів. Третій крок передбачає використання робочий набір даних № 1, який вже містить релігійну та гендерну класифікацію, і набору даних № 4, класифікованого за віком. Навчання моделі допомагає оцінити та коригувати репрезентативність за віковими ознаками, щоб кількість зразків для кожної вікової групи була збалансованою.

Кількість таких кроків відповідає кількості ознак, які потрібно врахувати для досягнення справедливості відповідно до FATE-принципу справедливості. Підсумковий результат – це оцінка репрезентативності робочого набору даних №1 за всіма обраними ознаками.

Результати та обговорення Для дослідження ефективності методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості було розроблено програмне забезпечення, яке реалізовано за допомогою мов програмування C# та Python. Після коригування репрезентативності датасету за гендерною, релігійною та віковою ознаками відповідно до FATE-принципу справедливості було отримано модифікований робочий набір даних № 1. Результати дослідження ефективності методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості наведені на рисунку 1.

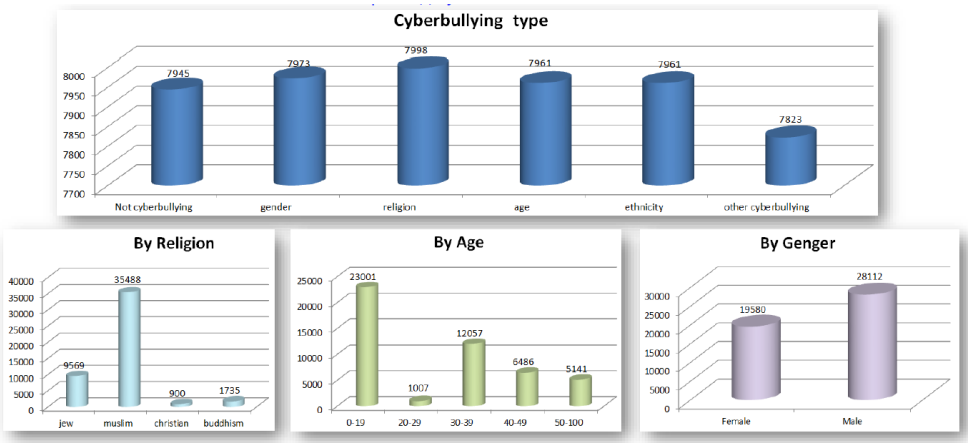

Рис. 1. Розподіл робочого набору даних №1 за видом кіберзалякування, ознаками гендеру, релігії та віку згідно FATE-принципу справедливості

Для різних класів було досягнуто різних рівнів лінійної роздільної здатності: дані за релігійною ознакою є добре роздільними; дані за гендерною ознакою FastForest, показали середню роздільність, тоді як дані за гендерною ознакою і дані за віковою ознакою є погано роздільними.

Висновки. Отже, важливість впровадження етичних принципів у розробку штучного інтелекту, зокрема забезпечення справедливості та рівного представництва різних груп у датасетах, полягає в тому, що це сприяє створенню систем, які не лише ефективно вирішують завдання, але й дотримуються соціальних і моральних норм. Впровадження таких принципів допомагає уникнути дискримінації та упередженості, що може виникнути через нерепрезентативність даних, на яких тренуються моделі.

Дослідження ефективності методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості було проведено за допомогою програмної реалізації на мовах програмування C# та Python, з використанням трьох наборів даних, що відображають три ознаки FATE-принципу: гендер, вік і релігія, а також два датасети, розмічені за видами кіберзалякувань, для створення робочого набору даних.

ПОСИЛАННЯ

1. 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.

2. 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.

3. Tang, Lin, and Yu-Sheng Su. Ethical Implications and Principles of Using

Artificial Intelligence Models in the Classroom: A Systematic Literature Review. International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 5, 2024, pp. 25.

4. Kaggle. Cyberbullying Classification. URL: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification?resource=download.

5. Kaggle. Cyberbullying Dataset. URL: https://www.kaggle.com/datasets/ashiqnazir/cbtweets.

6. Orelaja, A., Ejiofor, C., Sarpong, S., Imakuh, S., Bassey, C., Opara, I., Tettey, J.N.A. and Akinola, O. Attribute-specific Cyberbullying Detection Using Artificial Intelligence. Journal of Electronic & Information Systems. 6, 1 (Apr. 2024), 10–21.

7. Islam, M. R., Bataineh, A. S., Zulkernine, M. Detection of Cyberbullying in Social Media Texts Using Explainable Artificial Intelligence. Ubiquitous Security. UbiSec 2023. Communications in Computer and Information Science, vol 2034.