УДК: 004.8

# ПРИКЛАДНЕ ЗАСТОСУВАННЯ МЕТОДУ АНАЛІЗУ ТА ФОРМУВАННЯ РЕПРЕЗЕНТАТИВНИХ ВИБІРОК ТЕКСТОВИХ ДАНИХ

Собко О.В. (olena.sobko.ua@gmail.com)
Хмельницький національний університет (Україна)

*Розглянуто метод аналізу та формування репрезентативних вибірок текстових даних та його прикладне застосування. Прикладне програмне застосування методу дозволяє проводити аналіз вибірок текстових даних на репрезентативність за етичними аспктами принципу справедливості FATE, а також формувати репрезентативні вибірки за критеріями.*

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 тренувалися, іншими словами вміст цих датасетів безпосередньо впливає на кінцевий результат.

На даний час текстові датасети створюються з метою досягнення певної кінцевої цілі, часто без врахування етичних аспектів. Відсутні засоби для оцінювання репрезентативності текстового набору даних відповідно до принципів етичної недискримінації, що особливо актуально для соціально важливих та чутливих задач, до прикладу виявлення кіберзалякувань, визначення емоційного стану людей за текстовими дописами, тощо. Це призводить до того, що отримані результати можуть бути потенційно індивідуально дискримінаційними за різними ознаками, наприклад, такими як вік, стать, раса, релігія тощо.

Репрезентативність даних у датасетах не лише впливає на точність результатів та моделей, але й тісно пов'язана з етичними принципами FATE (Fairness, Accountability, Transparency, Ethics) – справедливістю, підзвітністю, прозорістю та етикою 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. Забезпечення репрезентативності є важливим для того, щоб моделі були справедливими щодо всіх демографічних груп і уникали системних упереджень, що можуть дискримінувати окремі групи населення [1].

Метод передбачає не тільки аналіз на репрезентативність, а й формування репрезентативної вибірки. При чому просте доповнення вибірки зразками, згенерованими, наприклад, за методикою SMOTE не є оптимальним, так як багатокритеріальне (за кількома етичними аспектами одночасно) формування репрезентативного датасету, призведе до нерепрезентативного представлення даних вибірки за окремими етичними аспектами.

Вхідними даними для методу аналізу та формування репрезентативних вибірок текстових даних є вибірка текстових даних для аналізу, цільова кількість елементів у вибірці, множина етичних аспектів, яка містить також класи та цільові пропорції класів, відповідно навчена множина моделей машинного навчання для кожного етичного аспекту, яка для навчання використовує збалансовані вибірки для кожного етичного аспекту.

На першому кроці здійснюється попередня обробка вибірки текстових даних, а саме видалення неінформативних фрагментів тексту, таких як знаки пунктуації, цифри та спеціальні символи [2]. Знаки пунктуації, як-от крапки, коми, знаки оклику та питання, зазвичай не несуть змістового навантаження при автоматизованій обробці тексту і тому видаляються для уникнення зайвого ускладнення процесу аналізу. Цифри також видаляються, так як не мають ключового значення для контексту вибірки, наприклад, коли йдеться про випадкові числові дані, які не є предметом дослідження. До таких елементів також відносяться спеціальні символи, зокрема знаки «@» або «#», які в більшості випадків не несуть аналітичного інтересу. Видалення смайлів під час попередньої обробки текстових даних в даному випадку є недоцільним.

На кроці 2 здійснюється аналіз репрезентативності вибірки текстових даних з урахуванням етичних аспектів. Спершу необхідно здійснити векторизацію кожного елемента вибірки даних, використовуючи кілька моделей машинного навчання для кожного з етичних аспектів. Далі кожен

зразок вибірки піддається класифікації машинною моделлю за визначеним етичним аспектом. Після цього на основі класифікації необхідно визначити наявні пропорції класів для кожного з етичних аспектів, що дозволить визначити, наскільки рівномірно або нерівномірно представлені різні категорії даних у вибірці щодо обраних етичних аспектів. Далі потрібно обчислити відхилення наявних пропорцій класів від цільових. Таким чином буде визначено ступінь несправедливого, упередженого представлення одних демографічних підгруп порівняно з іншими. Цільові пропорції визначають на бажаних показниках справедливості або на реальних демографічних пропорціях підгруп певного населення: міста, країни, тощо. Після цього здійснюється обчислення кількості елементів кожного класу, яких не вистачає або які перевищують необхідні пропорції, що необхідно для визначення, які саме дані демографічних підгруп потрібно збільшити або зменшити для досягнення репрезентативного вигляду вибірки текстових даних. Останнім на цьому кроці, необхідно оцінити достатність даних для аугментації, що полягає в перевірці мінімальної кількості зразків кожного класу для коректної представлення демографічної підгрупи у вибірці. Якщо певні демографічні підгрупи за певним етичним аспектом представлені недостатньо, то для такого класу вибірки даних відбувається аугментація даних, тобто штучне розширення вибірки для досягнення потрібної пропорції.

Третій крок передбачає репрезентативне коригування вибірки даних з урахуванням етичних аспектів. Перш за все, вирішується оптимізаційна задача вибору надлишкових елементів, які мають бути видалені для досягнення цільових пропорцій класів. Для цього необхідно виконати аналіз надлишкових даних, тобто тих елементів, які перевищують встановлені цільові пропорції класів за обраними етичними аспектами. Після цього здійснюється видалення надлишкових елементів, яке б не порушувало внутрішню структуру вибірки з метою збереження репрезентативності за усіма обраними етичними аспектами. Видаляються тільки ті елементи, які не є необхідними для забезпечення цільових пропорцій за етичними аспектами. Метою цього етапу є забезпечення збалансованої вибірки, де кожен клас представлений у межах оптимальних пропорцій демографічних підгруп. Далі формуються вимоги до кожного елемента для аугментації. Завершальним на цьому кроці є здійснення безпосередньо аугментації вибірки до цільових вимог. Аугментація повинна не призводити до штучного спотворення вибірки – недостатньо представлені класи доповнюються новими елементами, що відповідають заздалегідь визначеним критеріям, а саме цільовій кількості елементів у вибірці та множині етичних аспектів, яким повинна відповідати вхідна вибірка. Така аугментація дозволяє досягти необхідної кількості елементів для кожного класу за кожним обраним етичним аспектом, що забезпечує репрезентативне формування вибірки. Таким чином вихідними даними є вибірка даних, збалансована згідно цільових пропорцій за цільовими етичними аспектами.

Для дослідження ефективності методу аналізу та формування репрезентативних вибірки текстових даних було створено прикладну програмну реалізацію з використанням мови програмування Python. Для реалізації класифікації вхідного датасету з кіберзалякувань за ознаками гендеру, віку та релігії використано бібліотеку tensorflow. На Fig. 6 наведено приклад класифікації за релігійною ознакою FATE-принципу справедливості.


Рисунок 1 – Прикладна реалізація класифікації датасету з кіберзалякувань за релігійним етичним аспектом

Отже, було створено прикладну програмну реалізацію методу аналізу та формування репрезентативної вибірки текстових даних. Практичне застосування розробленого методу дозволяє встановити чи є репрезентативним за принципом справедливості FATE досліджуваний датасет. Якщо датасет не є репрезентативним, то за допомогою програмного забезпечення можна трансформувати датасет у репрезентативний за етичними аспектами (віковим, гендерним, релігійним, расовим, тощо) вигляд. Таким чином отримана в результаті роботи вибірка текстових даних, яка забезпечує репрезентативність та етичну коректність даних, дозволить шляхом навчання систем штучного інтелекту, формувати етичні за принципами FATE моделі машинного навчання.

**Список використаних джерел**

[1] Manziuk E., Barmak O., Krak I., Mazurets O., Skrypnyk T. Formal model of trustworthy artificial intelligence based on standardization. CEUR Workshop Proceedings, 2021, vol. 2853, pp. 190–197.

[2] 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, pp. 344–356.

УДК 004.8

## ПЕРСПЕКТИВИ ВИКОРИСТАННЯ ШТУЧНОГО ІНТЕЛЕКТУ В СФЕРІ КОРИСТУВАЦЬКИХ МУЗИЧНИХ СЕРВІСІВ

Цаплін О.О., Ізвалов О.В. (cabiturient@gmail.com, alexey@globalgamejam.org)
Економіко-технологічний інститут імені Роберта Ельворті (Україна)

*У тезах розглядається поняття адаптивності та адаптивних систем та їх значення в сучасних інформаційних технологіях та, зокрема, у сфері користувацьких музичних сервісів. Описується перспективна ніша для підвищення рівня адаптивності із застосуванням штучного інтелекту (ШІ), наводиться принцип роботи системи, основні елементи штучного інтелекту, що забезпечують адаптивність, та обов'язкові ознаки їх використання.*

**Вступ.** Сучасні підходи до інтелектуального аналізу аудіоконтенту (ІААК) спрямовані на автоматизацію обробки звукових сигналів за допомогою обробки природної мови (NLP). Алгоритми, які використовуються для обробки команд користувачів, також можуть аналізувати музичні треки, визначати їхні характеристики та підбирати їх відповідно до індивідуальних уподобань. Це дозволяє створювати персоналізовані рекомендації, покращуючи досвід користувачів і задовольняючи попит на адаптивні музичні послуги. Адаптивні системи, що враховують потреби користувачів, здатні значно підвищити ефективність взаємодії та задовольнити індивідуальні вимоги. З урахуванням даних про значну частку користувачів, які регулярно слухають аудіоконтент, стає очевидним, що адаптивні технології відіграють ключову роль у задоволенні попиту на персоналізовані музичні послуги.

**Актуальність.** Популярність аудіотехнологій зростає: за звітами 2023 року, понад 70% інтернет-користувачів щонайменше раз на тиждень слухають аудіоконтент, а кількість активних користувачів потокових музичних сервісів перевищила 500 мільйонів. Голосові асистенти, такі як Siri та Google Assistant, активно використовуються — близько 40% дорослих у США регулярно звертаються до них для пошуку інформації та управління пристроями. Очікується, що до 2025 року 75% домогосподарств у США матимуть принаймні один пристрій з вбудованим голосовим помічником. Ці дані підтверджують необхідність розробки адаптивних систем для покращення якості взаємодії та забезпечення персоналізованого досвіду.

У даному контексті актуальною вважається потреба у розробці адаптивного плеєру з елементами штучного інтелекту. Проект отримав робочу назву «YouPlayer». Інтеграція штучного інтелекту в додаток «YouPlayer» дозволяє покращити обробку звукових сигналів та адаптувати