REFERENCES

1. I. S. Nevludov, et al. “Cloud giants: AWS, Azure and GCP: дис.,” 2023 2nd International Conference on Innovative Solutions in Software Engineering Ivano-Frankivsk. 2023, pp. 18-24.

2. S. Sotnik, et al. “Overview: PHP and MySQL features for creating modern web projects,” International Journal of Academic Information Systems Research. 2023, pp. 11-17.

3. S. Sotnik, et al. “Development Features Web-Applications,” International Journal of Academic and Applied Research. 2023, pp. 79-85.

4. V. G. Kaponkin, et al. “The role of big data in improving functionality of search engines,” The 8th International scientific and practical conference “European congress of scientific achievements” (August 12-14, 2024) Barca Academy Publishing, Barcelona, Spain. 2024, pp. 69-76.

5. S. V. Sotnik “Implementation of game-based learning method,” Комп'ютерні ігри та мультимедіа як інноваційний підхід до комунікації - 2024 / Матеріали IV Всеукраїнської науково-технічної конференції молодих вчених, аспірантів і студентів, Одеса, 26-27 вересня 2024 р. 2024, pp. 19-22.

6. A. Tverdokhlib, et al. “Intelligent tools for optimizing information and search engines,” Manufacturing & Mechatronic Systems 2024: Proceedings of VIII st International Conference, Kharkiv, October 25-26, 2024, pp. 28-31.

7. S. V. Sotnik “Features of using REST architecture for development of ARS for information systems,” Міжнародна науково-практична конференція «Інформаційні системи в управлінні проєктами та програмами», Коблево, 9–13 вересня 2024 р. Збірник праць. – Харків: ХНУРЕ. 2024, pp. 42-45.

8. I. A. Borysenko, et al. “Chat gpt features in data search,” The 9th International scientific and practical conference “Scientific progress: innovations, achievements and prospects” (May 29-31, 2023) MDPC Publishing, Munich, Germany. 2023, pp. 139-143.

УДК 004.8
Собко О.В.[1]
[1] викл. кафедри комп’ютерних наук, Хмельницький національний університет

МЕТОД АНАЛІЗУ ТА ФОРМУВАННЯ РЕПРЕЗЕНТАТИВНИХ ДАТАСЕТІВ ДЛЯ ВИЯВЛЕННЯ КІБЕРЗАЛЯКУВАНЬ У ТЕКСТОВОМУ КОНТЕНТІ

Відсутність прозорості щодо джерел і характеристик даних, які використовуються для навчання алгоритмів штучного інтелекту, підриває 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.

Відсутність уваги до етичних компонентів при створенні та використанні датасетів призводить до упередженості в алгоритмах, що негативно впливає на справедливість та достовірність прийнятих рішень [1].

Метою роботи є розробка методу аналізу та формування репрезентативних датасетів для виявлення кіберзалякувань у текстах.

Метод аналізу та формування репрезентативних датасетів передбачає не тільки аналіз на репрезентативність, а й формування репрезентативної вибірки [2]. 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. Наприклад, необхідно сформувати текстову вибірку репрезентативну за двома етичними аспектами – гендерним та віковим. Виявивши нерепрезентативне представлення за віковим аспектом, необхідно доповнити текстову вибірку таким чином, щоб не порушити реперезентативність вибірки за гендерним аспектом, таким чином необхідно розв’язати задачу оптимізації формування репрезентативного текстової вибірки за усіма обраними етичними аспектами одночасно. Схему методу аналізу та формування репрезентативних вибірок текстових даних наведено на рисунку 1.

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 [3]. 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. Смайли слугують важливими емоційними індикаторами, які можуть значно змінити смисл речення. У багатьох випадках смайли можуть слугувати своєрідними маркерами настроїв або ставлень до теми. Включення смайлів в аналіз дозволяє покращувати точність моделей машинного навчання, що використовуються для класифікації текстів за емоційним або настроєвим змістом.

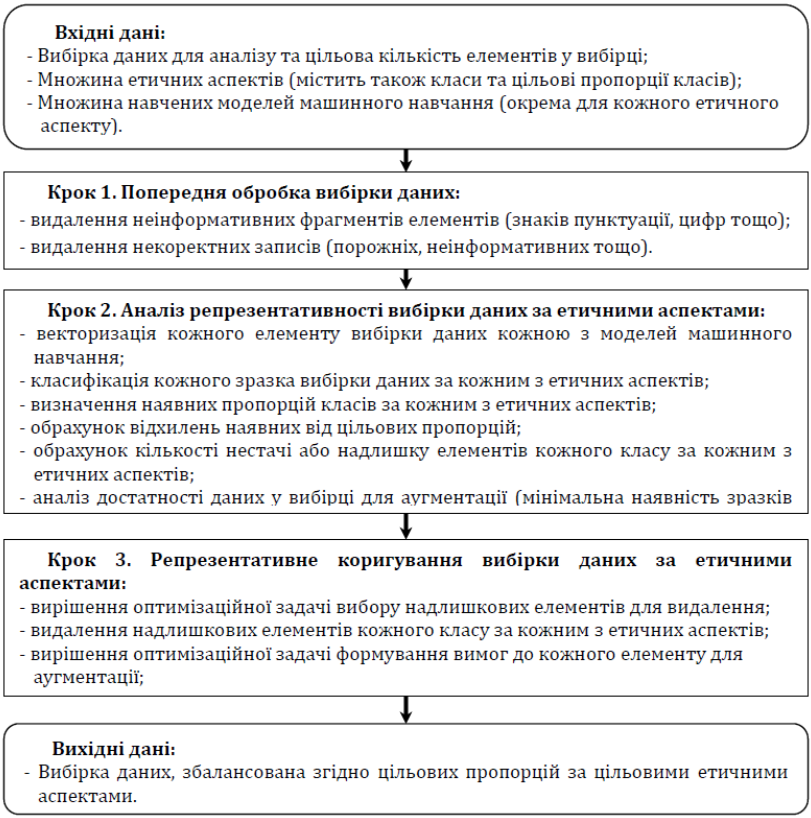

Рисунок 1 – Схема методу аналізу та формування репрезентативних вибірок текстових даних.

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 коректного представлення демографічної підгрупи. Якщо певні 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. Після вирішення оптимізаційної задачі 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, що також є оптимізаційною задачею, яка вимагає визначення, яким чином збільшити кількість елементів недостатньо представлених класів, щоб досягти бажаних пропорцій. 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. Така аугментація дозволяє досягти необхідної кількості елементів для кожного класу за кожним обраним етичним аспектом, що забезпечує репрезентативне формування вибірки.

Таким чином, вихідними даними методу є вибірка даних, збалансована згідно цільових пропорцій за цільовими етичними аспектами.

Отже, було розроблено метод аналізу та формування репрезентативних вибірок текстових даних. 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.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. 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.

2. Собко О. В. Дослідження ефективності методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості. Матеріали VIII Міжнародної науково-практичної конференції «Перспективи сучасної науки: теорія і практика». 16-18.09.2024. Львів – 2024. с. 217-221.

2. Собко О. Метод інтелектуального виявлення та класифікації кіберзалякувань у текстовому контенті. Матеріали XII Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології ІУСТ-ОДЕСА-2024». 23-25.09.2024. Одеса. 2024. С.262-265.

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, pp. 344–356.

УДК 004.9

Тарасов О.Ф.[1], Алтухов О.В. [2], Васильєва Л.В.[3]
[1] проф. Донбаської державної машинобудівної академії
[2] старш. викл. Донбаської державної машинобудівної академії
[3] доц. Київського національного економічного університету ім. В.Гетьмана

ЯДРО СИСТЕМИ МОДЕЛЮВАННЯ ПРОЦЕСІВ ІНТЕНСИВНОЇ ПЛАСТИЧНОЇ ДЕФОРМАЦІЇ

Процеси інтенсивної пластичної деформації (ІПД) дозволяють отримувати матеріали з дрібнозернистою структурою та покращеними фізико-механічними властивостями. Процеси ІПД активно розвиваються в