

ВИЯВЛЕННЯ ТА КЛАСИФІКАЦІЯ КІБЕРЗАЛЯКУВАНЬ У ТЕКСТОВОМУ КОНТЕНТІ ЗАСОБАМИ ШТУЧНОГО ІНТЕЛЕКТУ

Олена СОБКО, Олександр БАРМАК, Арчіл ЧОЧІА, 2025 рік

1. Method for Analysis and Formation of Representative Text Datasets / O. Sobko, O. Mazurets, M. Molchanova, I. Krak, O. Barmak. CEUR Workshop Proceedings, 2025, vol. 3899, pp. 84–98. *(індексована в наукометричній базі Scopus)*.
2. Abusive Speech Detection Method for Ukrainian Language Used Recurrent Neural Network / I. Krak, O. Zalutska, M. Molchanova, O. Mazurets, R. Bahrii, O. Sobko, O. Barmak. CEUR Workshop Proceedings, 2024, vol. 3688, pp. 16–28. *(індексована в наукометричній базі Scopus)*.
3. Method for Neural Network Cyberbullying Detection in Text Content With Visual Analytic / I. Krak, O. Sobko, M. Molchanova, I. Tymofiiiev, O. Mazurets, O. Barmak. CEUR Workshop Proceedings, 2025, vol. 3917, pp. 298–309. *(індексована в наукометричній базі Scopus)*.

Підхід до виявлення та класифікації типів кіберзалякувань у текстовому контенті



Обмеження підходу до виявлення кіберзалякувань у текстовому контенті

- Мовні особливості** – розроблені методи працюють виключно з україномовними текстами; для використання в іншомовному середовищі потрібна адаптація моделей і відповідні датасети.
- Використання машинного перекладу** – у зв'язку з відсутністю автентичних україномовних датасетів застосовувався переклад англomовних датасетів, що призводить до невеликих втрат значення, тональності й стилістичних особливостей.
- Обмеження довжини тексту** – аналіз здійснюється лише для текстів довжиною 100–300 символів, що обумовлено характеристиками навчальних датасетів.
- Обмежений перелік типів кіберзалякувань** – моделі розпізнають лише визначені категорії (вікове, гендерне, етнічне, релігійне, інші типи), через обмеження навчального датасету.

Інформаційна модель кіберзалякування

Інформаційну модель кіберзалякування CB можна подати у вигляді:

$$CB = \{TT, OC, RC, OTC, TC', VA, MetaData\}, \quad (1)$$

де TT – тестовий текст для аналізу; OC – числова оцінка наявності кіберзалякувань в тексті TT ; RC – рівень кіберзалякувань в тексті TT ; OTC – множина числових оцінок сили прояву кожного з досліджуваних типів кіберзалякувань TC ($|TC| = |OTC|$); TC' – множина наявних у тексті типів кіберзалякувань ($TC' \subset TC$, де TC – множина всіх досліджуваних типів кіберзалякувань); VA – множина візуальних подань інтерпретації результатів виявлення кіберзалякувань; $MetaData$ – множина метаданих моделі кіберзалякування CB .

У (1) **метадані моделі кіберзалякувань** CB можна подати у вигляді:

$$MetaData = \{MetaData1, MetaData2, MetaData3\}, \quad (2)$$

де $MetaData1$ – множина метаданих для оцінювання репрезентативності датасетів, $MetaData2$ – множина метаданих для виявлення кіберзалякувань; $MetaData3$ – множина метаданих для інтерпретації результатів.

Множина метаданих $MetaData2$ для виявлення та класифікації кіберзалякувань у текстовому контенті у (2) містить:

$$MetaData2 = \{TT_{2,3}, NN_{cbr}, NN_{cbd}\}, \quad (3)$$

де $TT_{2,3}$ – текст після препроцесингу $TT \rightarrow TT_{2,3}$ (ідентичний для $MetaData2$ та $MetaData3$); NN_{cbr} – нейромережева модель для виявлення кіберзалякувань (бінарна класифікація); NN_{cbd} – нейромережева модель для класифікації кіберзалякувань (мультилейблова класифікація).

Інформаційна модель кіберзалякування

Множина метаданих $MetaData3$ для інтерпретації результатів виявлення кіберзалякувань у текстовому контенті засобами у (2) подається у вигляді:

$$MetaData3 = \{TT_{2,3}, W, IM_{Interpet}\}, \quad (4)$$

де $IM_{Interpet}$ – інтерпретаційна модель для пояснення результатів класифікації NN_{cbd} (візуальна аналітика), W – множина значущих слів для типів кіберзалякувань у тексті $TT_{2,3}$.

У результаті застосування інтерпретаційної моделі $IM_{Interpet}$ формуються кортежі вигляду $W=\{(w_1, \beta_1), (w_2, \beta_2), \dots, (w_n, \beta_n)\}$, де w_i – i -те слово; β_i – вага i -го слова; n – кількість значущих слів у тексті $TT_{2,3}$ для обраного типу кіберзалякувань.

М1. Метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості

Поточний підхід

Відсутність засобів для оцінювання репрезентативності текстових наборів даних відповідно до принципу етичної недискримінації

Текстовий датасет, **незбалансований** за етично важливими ознаками

Навчання штучного інтелекту з **потенційною індивідуальною дискримінацією** за різними типами кіберзалякувань

Штучний інтелект дає етично некоректні результати



Етично коректний підхід

Оцінка репрезентативності наборів текстових даних відповідно до принципу справедливості FATE – етичної недискримінації за віком, статтю, расою, релігією тощо

Індивідуально неупереджений **збалансований** текстовий датасет

Навчання штучного інтелекту за **принципом етичної недискримінації** за різними типами кіберзалякувань

Штучний інтелект дає етично коректні результати

М1. Метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості

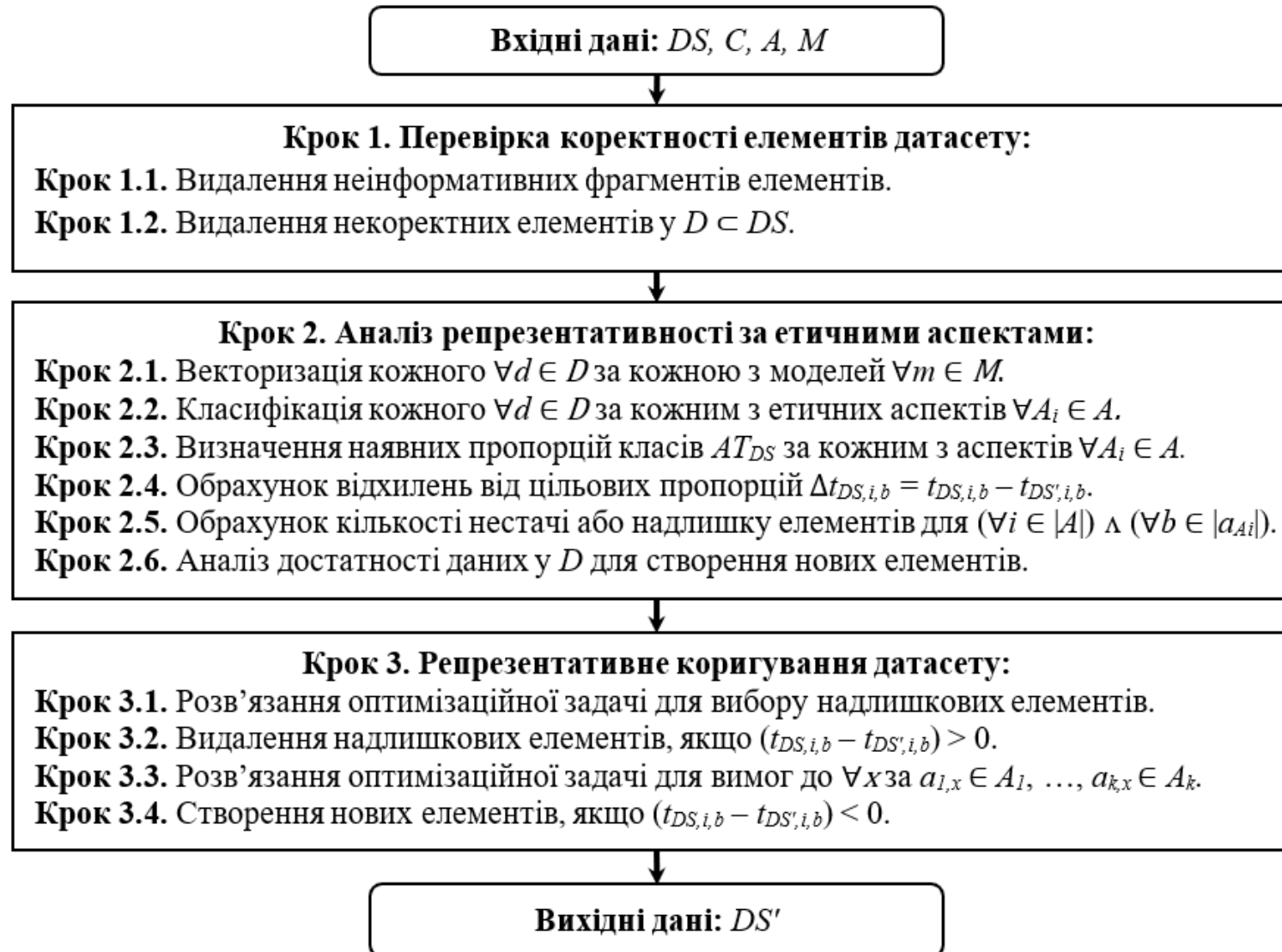
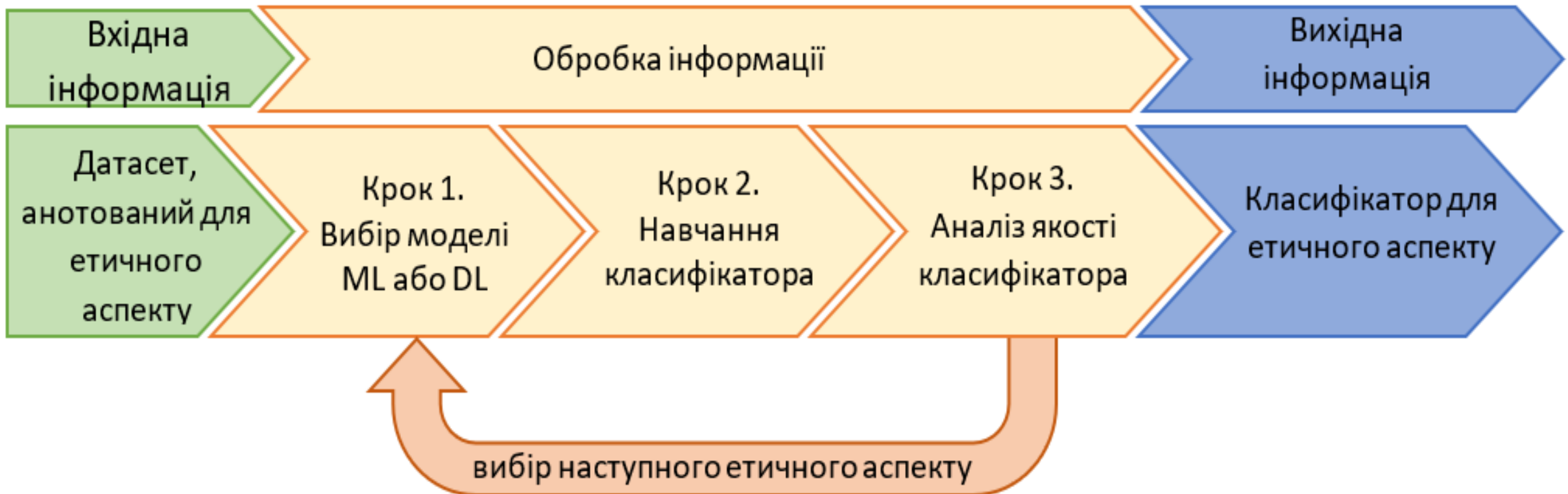


Схема та кроки
методу аналізу та
формування
репрезентативних
вбірок текстових
даних

М1. Метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості

Формування множини навчених моделей машинного навчання

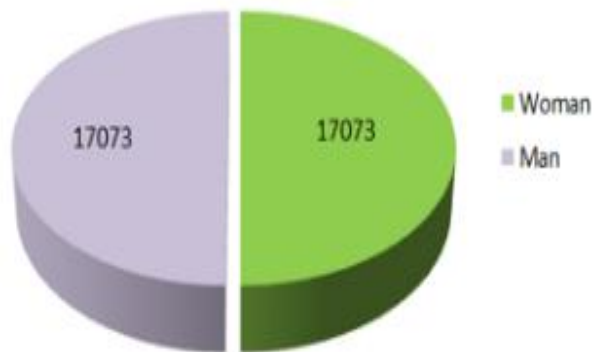


М1. Метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості

Перелік використаних датасетів

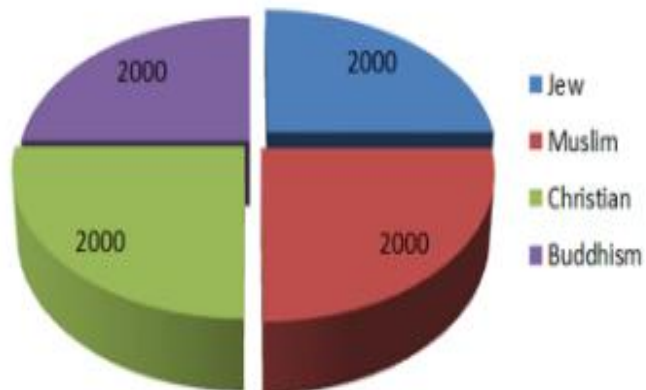
Датасети для навчання класифікаторів за етичними аспектами	Обсяг датасету	Розмічено за ознаками
«Tweet Files for Gender Guessing»	34146	Цисгендер
«TAG-it Dataset Distribution»	21948	Вік
«CyberBullying Detection Dataset»	20109	Релігія

CisGender DataSet



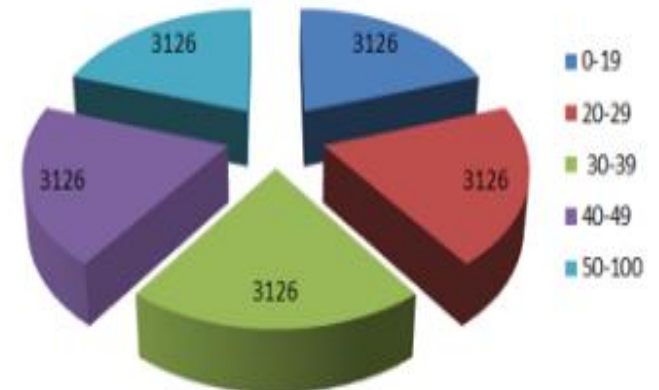
«Tweet Files for Gender Guessing»

Religion DataSet



«TAG-it Dataset Distribution»

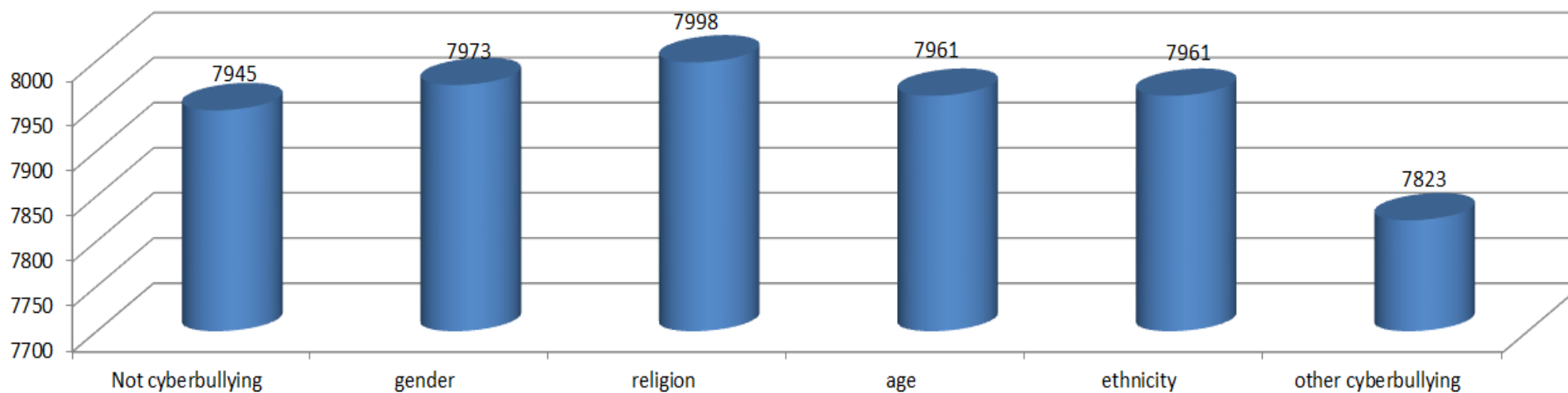
Age DataSet



«CyberBullying Detection Dataset»

M1. Метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості

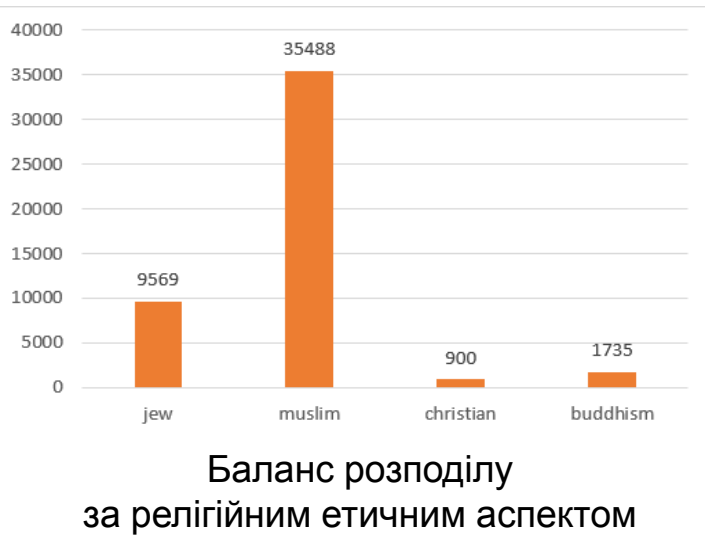
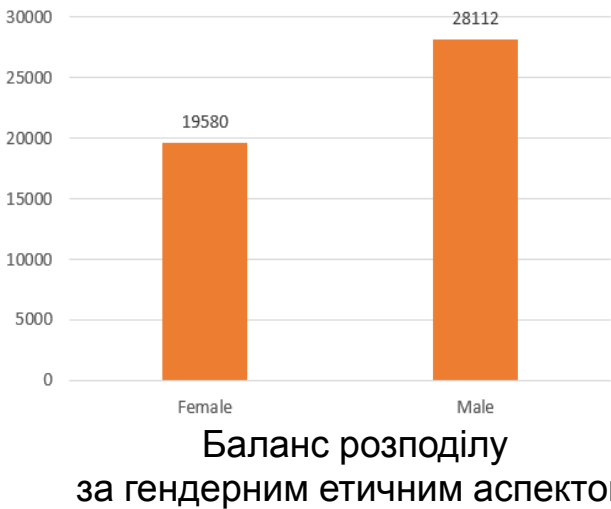
В якості даних для навчання класифікаторів та дослідження ефективності методу виявлення кіберзалякувань використано датасет *«Cyberbullying Tweets»* для бінарної класифікації та *«Cyberbullying Classification»* для мультилейблової.



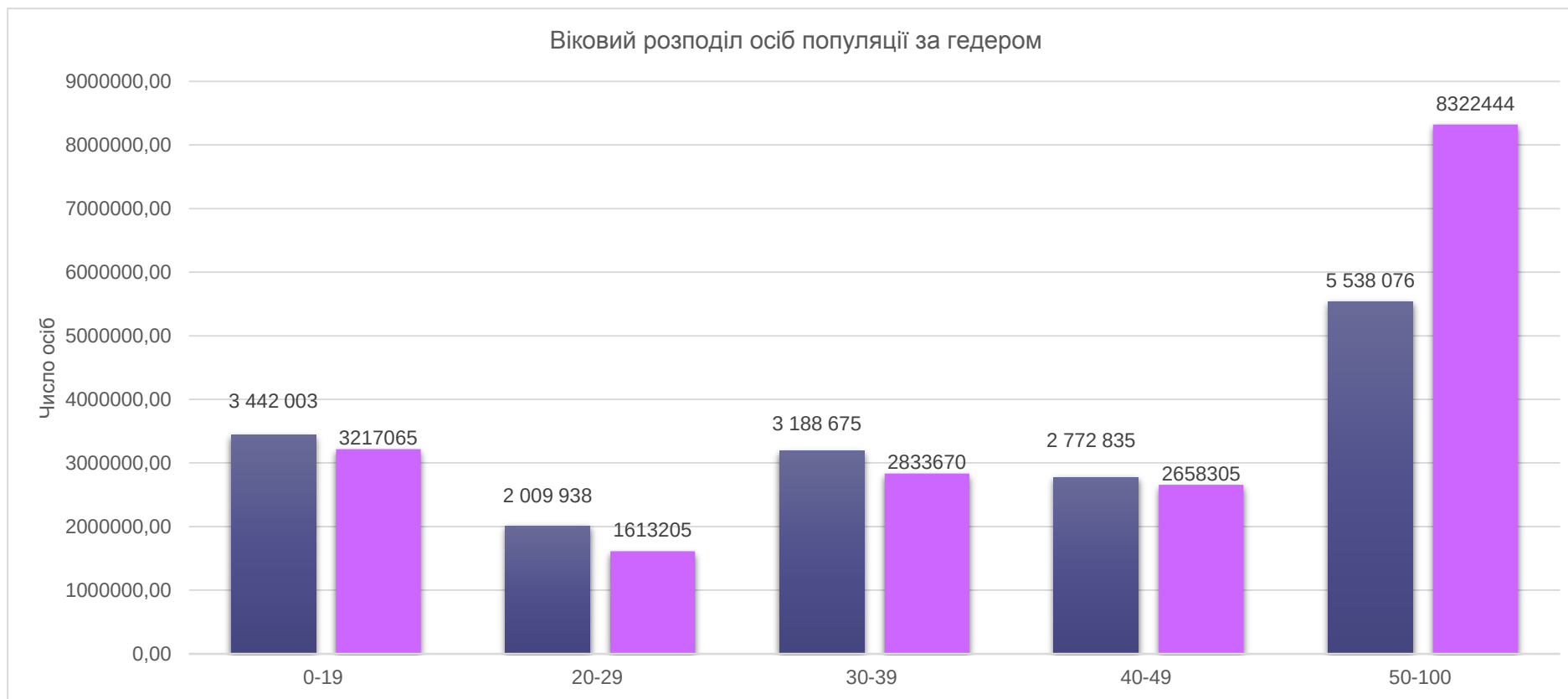
M1. Метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості

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

Модель	Accuracy, %	Precision, %	Recall, %	F1-score, %
Гендерний етичний аспект				
FastForest	55	56	52	54
SVM	50	50	50	50
LSTM	75	79	73	76
BERT	59	55	60	57
Віковий етичний аспект				
FastForest	46	46	43	43
SVM	79	68	59	63
LSTM	51	52	48	50
BERT	50	37	39	38
Релігійний етичний аспект				
FastForest	56	58	55	56
SVM	60	62	59	60
LSTM	61	68	61	62
BERT	85	87	71	78



М1. Метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості



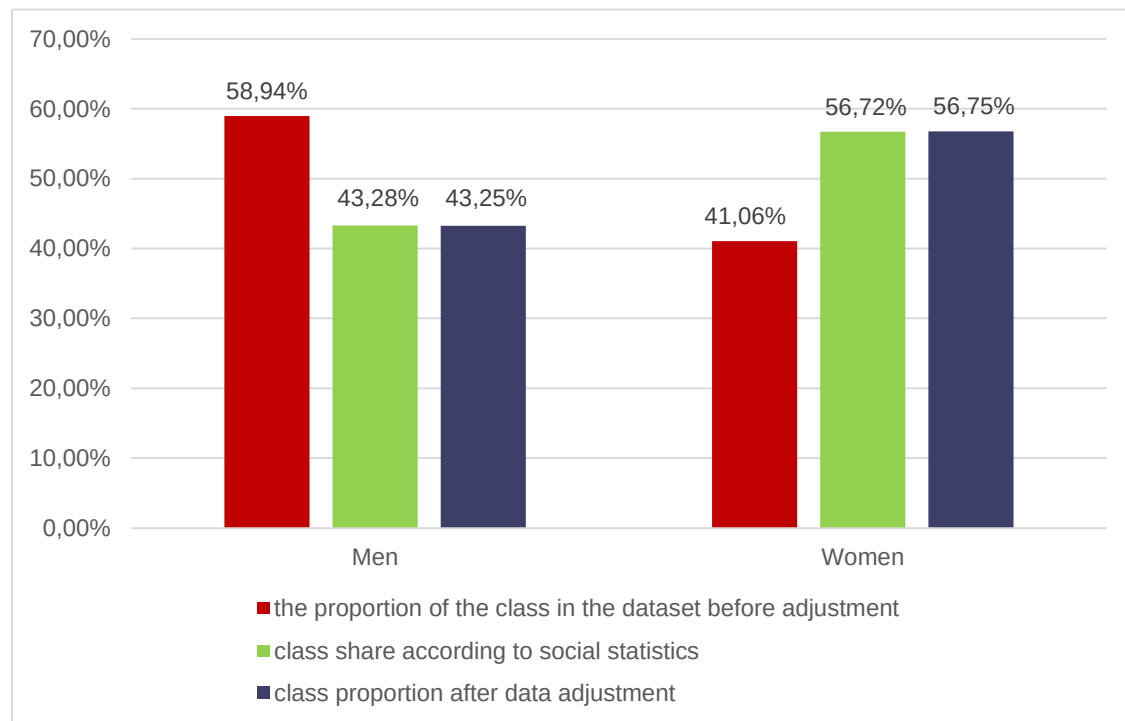
За даними Інституту демографії та соціальних досліджень імені М. В. Птуха НАН України станом на липень 2023 року загальна чисельність населення України становить 35 596 216 осіб.

M1. Метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості

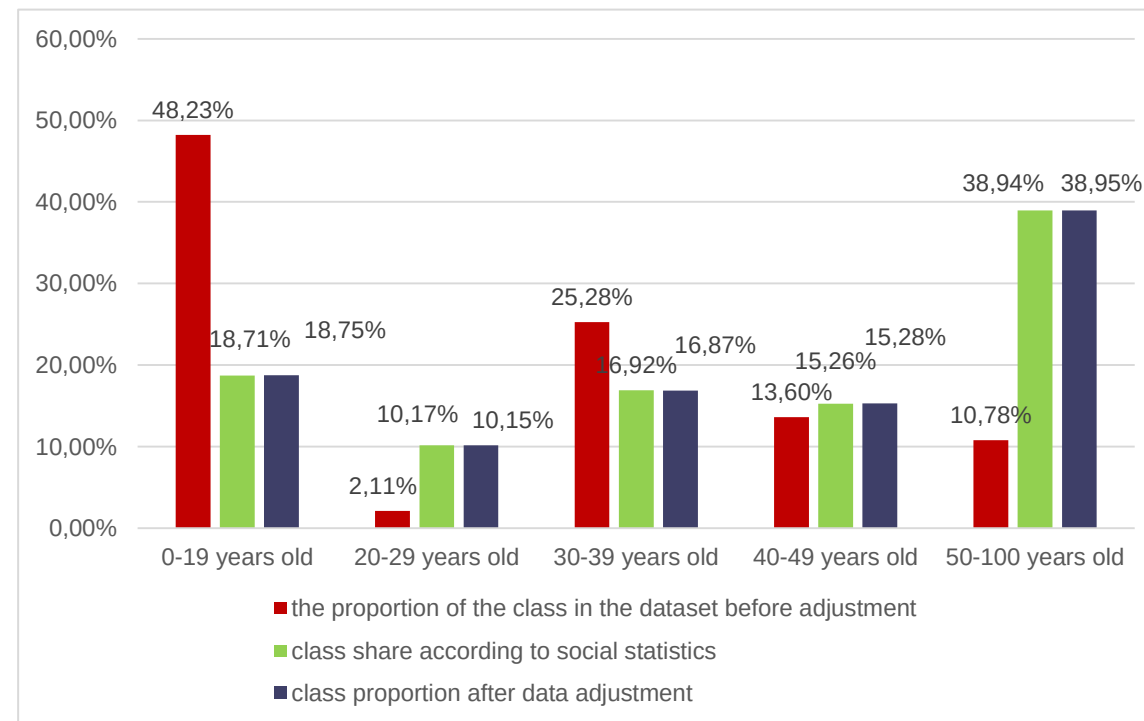
Розподіл зразків у сформованій репрезентативній вибірці після аугментації даних в результаті розв'язку багатокритеріальної оптимізаційної задачі

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 датасеті					
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 %

M1. Метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості



Баланс розподілу вхідного датасету за гендерним етичним аспектом FATE-принципу справедливості



Баланс розподілу вхідного датасету за віковим та гендерним етичними аспектом FATE-принципу справедливості

М2. Метод виявлення кіберзалякувань у текстовому контенті

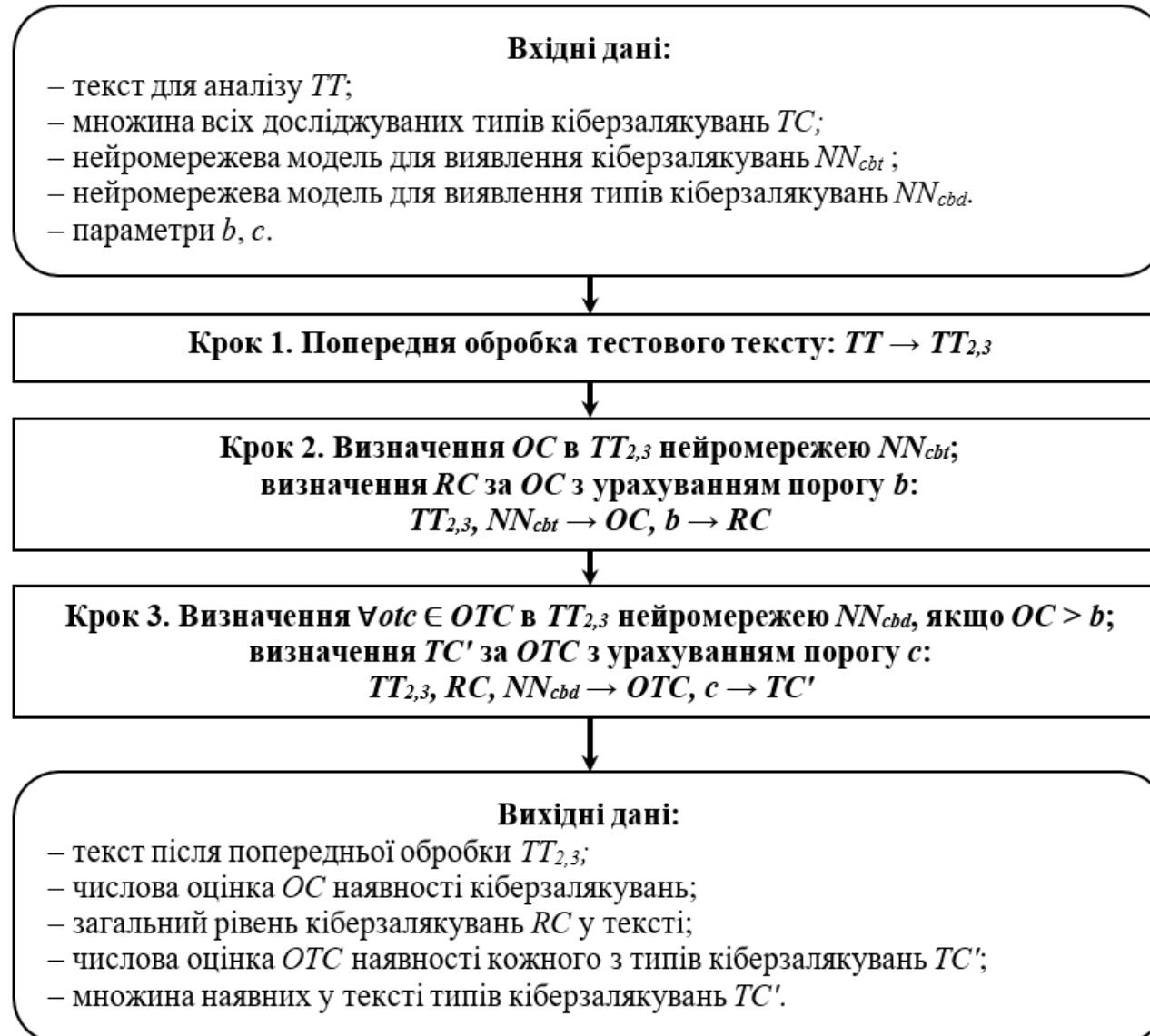


Схема методу
виявлення і
класифікації
кіберзалякувань
у текстовому
контенті

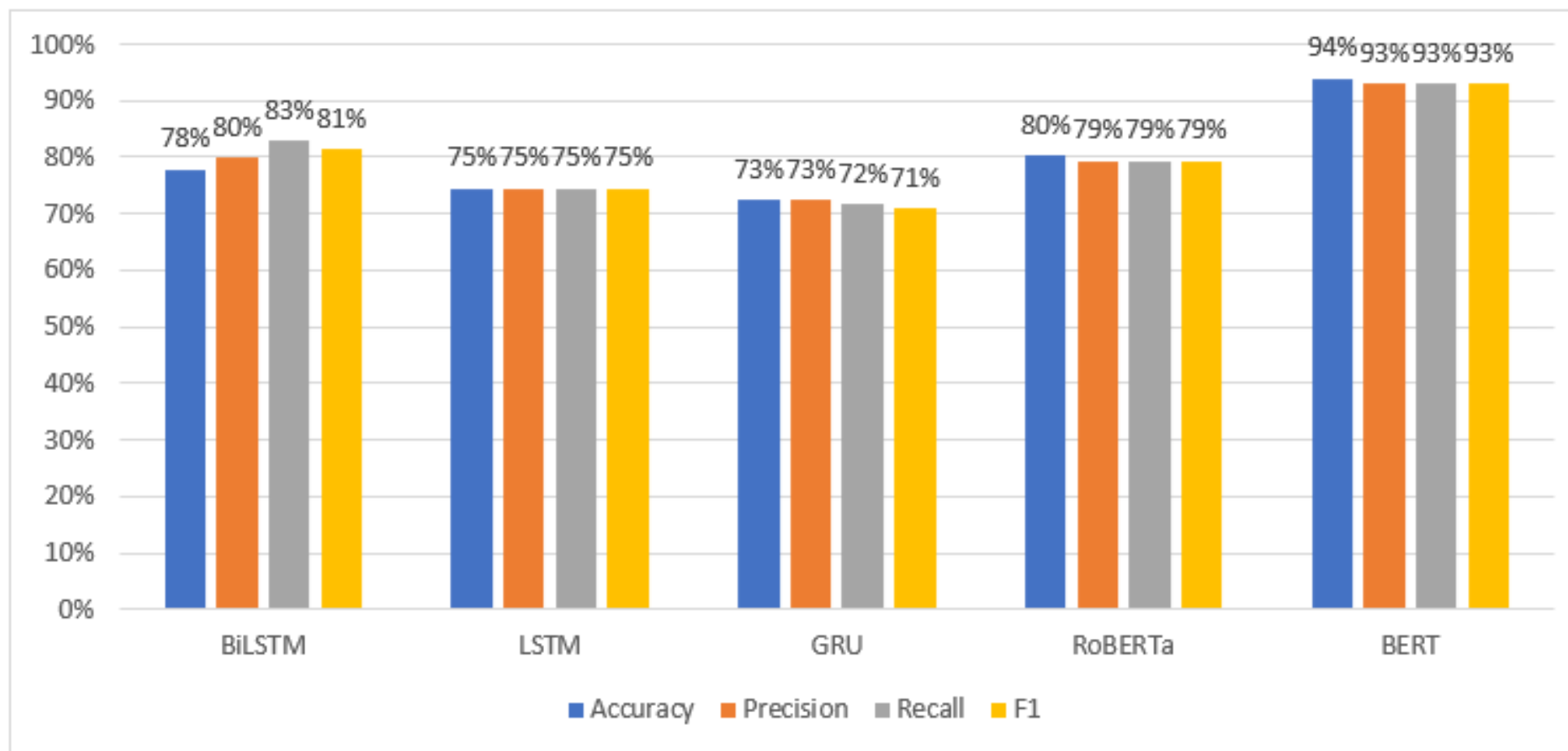
M2. Метод виявлення кіберзалякувань у текстовому контенті

Статистичні показники Accuracy, Precision, Recall, F1-score для бінарної класифікації згідно кроку 2 методу виявлення кіберзалякувань у текстовому контенті

Модель	Підхід до класифікації	Результати дослідження ефективності, %
BiLSTM	Два класи: кіберзалякування і не кіберзалякування	Accuracy: 94 Precision: 94 Recall: 94 F ₁ -score: 94
RNN з LSTM шаром	Словник + нейронна мережа для аналізу настрою	Accuracy: 94 Precision: 89,2 Recall: 91,2 F ₁ -score: 91,2

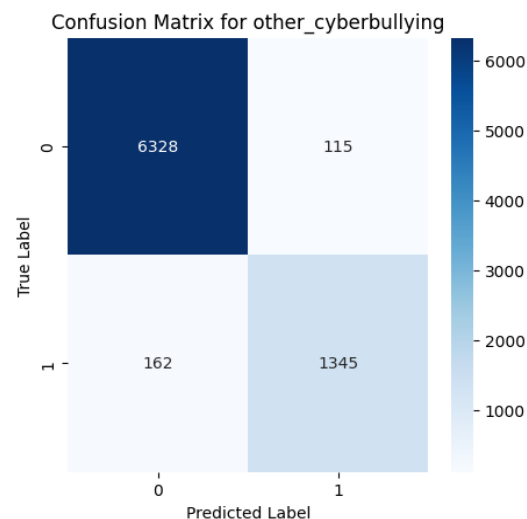
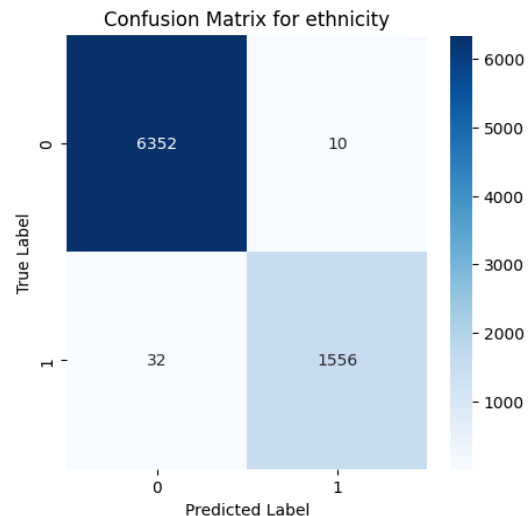
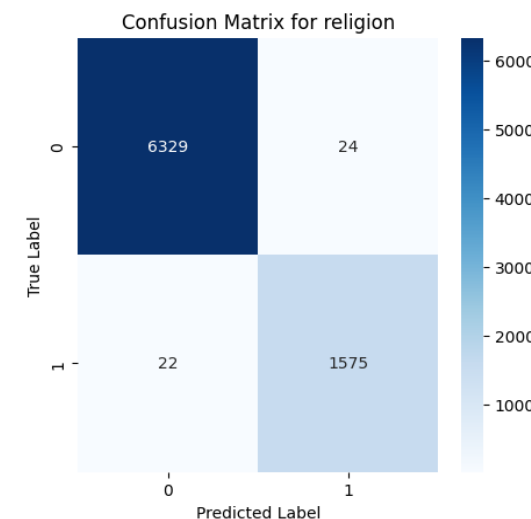
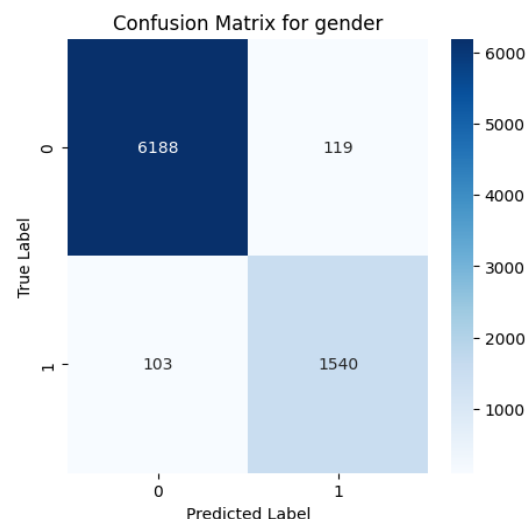
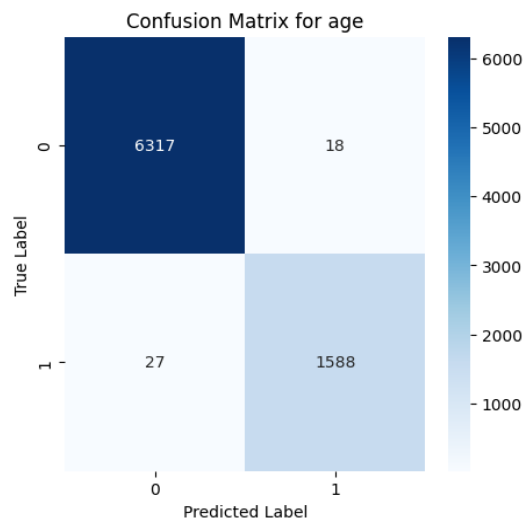
M2. Метод виявлення кіберзалякувань у текстовому контенті

Показники макрометрик моделей машинного та глибокого навчання для мультілейблової класифікації типів кіберзалякувань



M2. Метод виявлення кіберзалякувань у текстовому контенті

Матриці помилок для типів кіберзалякувань для HM BERT



Показники макрометрик навченої моделі BERT для мультілейблової класифікації типів кіберзалякувань отримали значення **Accuracy – 94 %**, **Precision – 93 %**, **Recall – 93 %**, **F1 Score – 93 %**.

М2. Метод виявлення кіберзалякувань у текстовому контенті

Показники метрик крос-валідації даних для моделі BERT

Вибірка	Accuracy, %	Precision, %	Recall, %	F1, %
Sample 1	93,8	93,4	93,2	93,3
Sample 2	94,0	93,5	93,6	93,5
Sample 3	94,2	93,6	93,4	93,5
Sample 4	94,1	93,8	93,7	93,7
Sample 5	94,3	93,7	93,8	93,8

Середні значення метрик після проведеної крос-валідації таким чином мають наступні значення: Accuracy 94.08%, Precision 93.6%, Recall 93.54%, F1-міра 93.56%. Відхилення від отриманих макрометрик отриманої моделі Accuracy 0.08%, Precision 0.6%, Recall: 0.54%, F1-міра 0.56%.

М3. Метод інтерпретації результатів виявлення кіберзалякувань

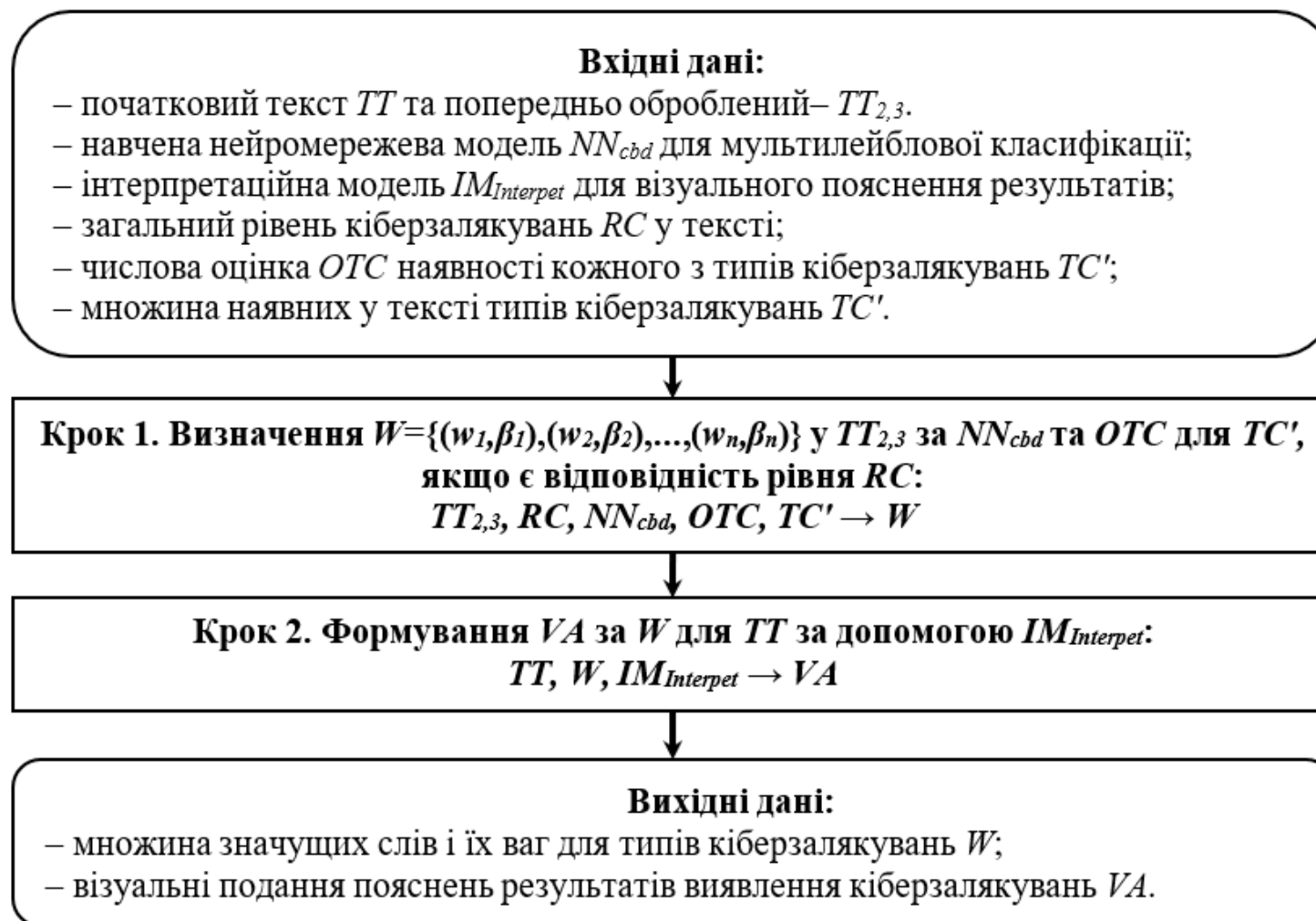


Схема методу
інтерпретації
результатів
виявлення
кіберзалякувань
у текстовому
контенті

М3. Метод інтерпретації результатів виявлення кіберзалякувань

«Ти нікчемна, як і всі твої однолітки, а те, що ти намагаєшся щось довести, тільки смішить – жінки взагалі не повинні займатися програмуванням, особливо такі, як ти»

Виявлення та класифікація кіберзалякувань

Ти нікчемна, як і всі твої однолітки, а те, що ти намагаєшся щось довести, тільки смішить – жінки взагалі не повинні займатися програмуванням, особливо такі, як ти.

Розпізнати

Завантажити файл

Метрики

Надати інтерпретацію результатів

Результати (мультилейблова класифікація):

- Кіберзалякування за віком: 0.67
- Кіберзалякування за гендером: 0.71
- Кіберзалякування за релігією: 0.12
- Кіберзалякування за етнічною ознакою: 0.02
- Не кіберзалякування: 0.1
- Інше: 0.22

Абсолютне значення ваг для інтерпретації результатів виявлення різних типів кіберзалякувань

Гендерне кіберзалякування:

Ти нікчемна, як і всі твої однолітки, а те, що ти намагаєшся щось довести (0.02), тільки смішить – жінки (0.12) взагалі не повинні (0.10) займатися програмуванням (0.09), особливо (0.07) такі, як ти.

Релігійне кіберзалякування:

Ти нікчемна (0.03), як і всі твої (0.005) однолітки, а те, що ти намагаєшся щось (0.007) довести, тільки смішить – жінки взагалі повинні (0.09) займатися програмуванням, особливо (0.05) такі (0.03), як ти.

Вікове кіберзалякування:

Ти нікчемна (0.08), як і всі твої однолітки (0.07), а те, що ти намагаєшся (0.04) щось (0.03) довести, тільки смішить (0.05) – жінки взагалі не повинні займатися програмуванням, особливо (0.01) такі, як ти.

Етнічне кіберзалякування:

Ти нікчемна, як і всі твої однолітки (0.04), а те, що ти намагаєшся (0.03) щось (0.015) довести, тільки смішить (0.05) – жінки взагалі не повинні займатися програмуванням, особливо (0.014) такі, як ти.

Інший тип кіберзалякування:

Ти нікчемна (0.09), як і всі твої однолітки (0.06), а те, що ти намагаєшся (0.04) щось (0.02) довести, тільки смішить (0.015) – жінки взагалі не повинні займатися програмуванням, особливо (0.01) такі, як ти.

М3. Метод інтерпретації результатів виявлення кіберзалякувань

Інтерпретація результатів виявлення різних типів кіберзалякувань

Виявлення та класифікація кіберзалякувань

Ти нікчемна, як і всі твої однолітки, а те, що ти намагаєшся щось довести, тільки смішить – жінки взагалі не повинні займатися програмуванням, особливо такі, як ти.

Розпізнати

Завантажити файл

Метрики

Надати інтерпретацію результатів

Результати (мультилейблова класифікація):

- Кіберзалякування за віком: 0.67
- Кіберзалякування за гендером: 0.71
- Кіберзалякування за релігією: 0.12
- Кіберзалякування за етнічною ознакою: 0.02
- Не кіберзалякування: 0.1
- Інше: 0.22

Абсолютне значення ваг для інтерпретації результатів виявлення різних типів кіберзалякувань

Гендерне кіберзалякування:

Ти нікчемна, як і всі твої однолітки, а те, що ти намагаєшся щось довести (0.02), тільки смішить – жінки (0.12) взагалі не повинні (0.10) займатися програмуванням (0.09), особливо (0.07) такі, як ти.

Релігійне кіберзалякування:

Ти нікчемна (0.03), як і всі твої (0.005) однолітки, а те, що ти намагаєшся щось (0.007) довести, тільки смішить – жінки взагалі повинні (0.09) займатися програмуванням, особливо (0.05) такі (0.03), як ти.

Вікове кіберзалякування:

Ти нікчемна (0.08), як і всі твої однолітки (0.07), а те, що ти намагаєшся (0.04) щось (0.03) довести, тільки смішить (0.05) – жінки взагалі не повинні займатися програмуванням, особливо (0.01) такі, як ти.

Етнічне кіберзалякування:

Ти нікчемна, як і всі твої однолітки (0.04), а те, що ти намагаєшся (0.03) щось (0.015) довести, тільки смішить (0.05) – жінки взагалі не повинні займатися програмуванням, особливо (0.014) такі, як ти.

Інший тип кіберзалякування:

Ти нікчемна (0.09), як і всі твої однолітки (0.06), а те, що ти намагаєшся (0.04) щось (0.02) довести, тільки смішить (0.015) – жінки взагалі не повинні займатися програмуванням, особливо (0.01) такі, як ти.

Виявлення та класифікація кіберзалякувань

Ти нікчемна, як і всі твої однолітки, а те, що ти намагаєшся щось довести, тільки смішить – жінки взагалі не повинні займатися програмуванням, особливо такі, як ти.

Розпізнати

Завантажити файл

Метрики

Надати інтерпретацію результатів

Результати (мультилейблова класифікація):

- Кіберзалякування за віком: 0.67
- Кіберзалякування за гендером: 0.71
- Кіберзалякування за релігією: 0.12
- Кіберзалякування за етнічною ознакою: 0.02
- Не кіберзалякування: 0.1
- Інше: 0.22

Інтерпретація результатів виявлення типів кіберзалякувань з урахуванням негативного чи позитивного типу впливу на результат

Гендерне кіберзалякування:

Ти нікчемна, як і всі твої однолітки, а те, що ти намагаєшся щось довести (-0.02), тільки смішить – жінки (0.12) взагалі не повинні (0.10) займатися програмуванням (0.09), особливо (0.07) такі, як ти.

Релігійне кіберзалякування:

Ти нікчемна (0.03), як і всі твої (-0.005) однолітки, а те, що ти намагаєшся щось (-0.007) довести, тільки смішить – жінки взагалі повинні (0.09) займатися програмуванням, особливо (0.05) такі (-0.03), як ти.

Вікове кіберзалякування:

Ти нікчемна (0.08), як і всі твої однолітки (0.07), а те, що ти намагаєшся (0.04) щось (-0.03) довести, тільки смішить (-0.05) – жінки взагалі не повинні займатися програмуванням, особливо (-0.01) такі, як ти.

Етнічне кіберзалякування:

Ти нікчемна, як і всі твої однолітки (0.04), а те, що ти намагаєшся (0.03) щось (-0.015) довести, тільки смішить (-0.05) – жінки взагалі не повинні займатися програмуванням, особливо (-0.014) такі, як ти.

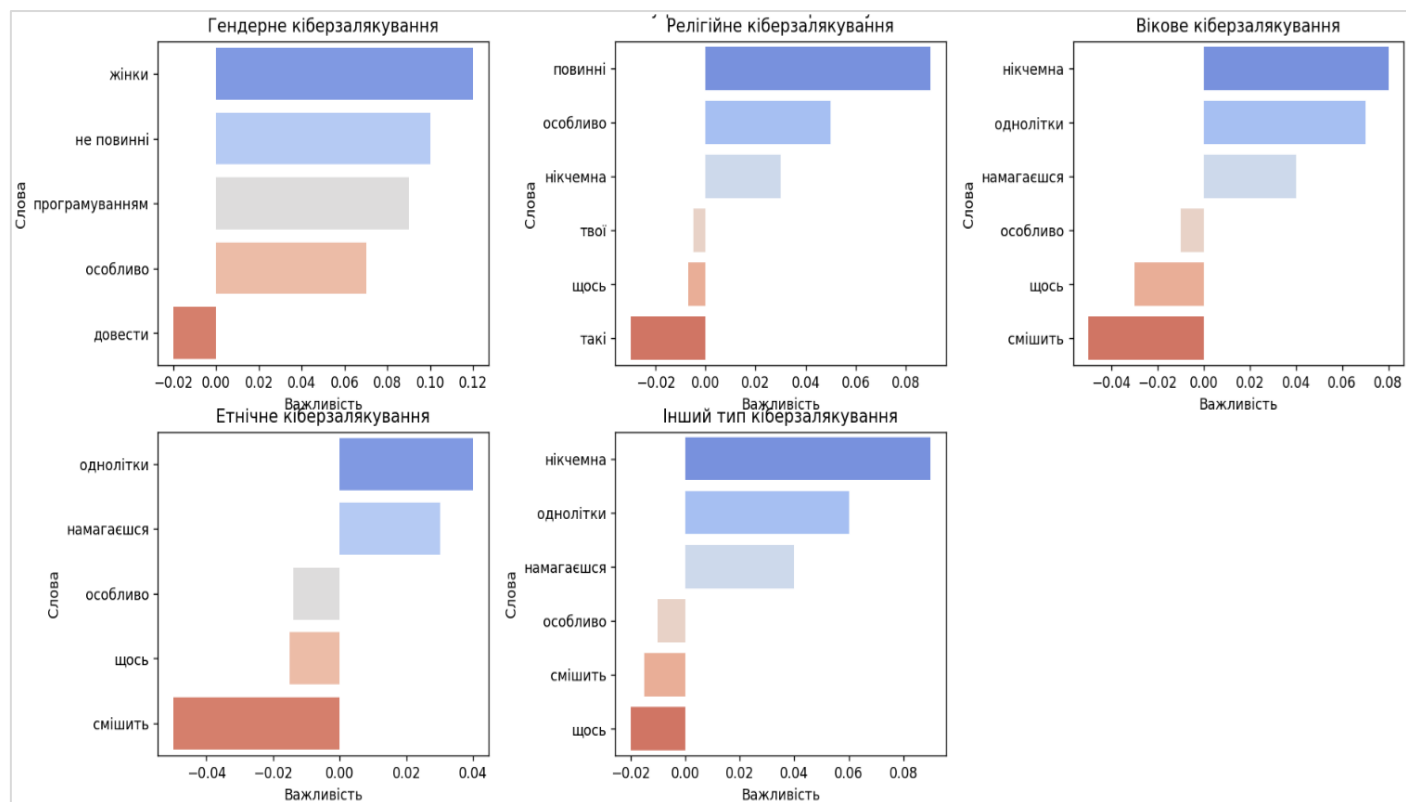
Інший тип кіберзалякування:

Ти нікчемна (0.09), як і всі твої однолітки (0.06), а те, що ти намагаєшся (0.04) щось (-0.02) довести, тільки смішить (-0.015) – жінки взагалі не повинні займатися програмуванням, особливо (-0.01) такі, як ти.

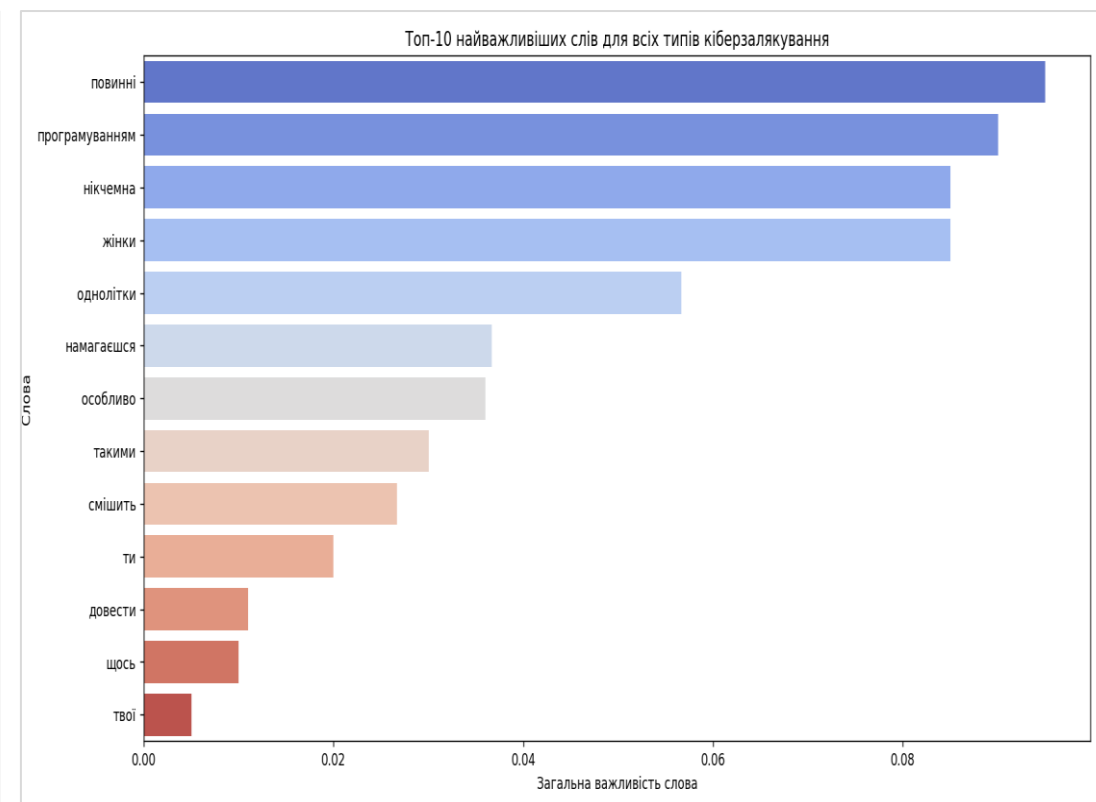
Використання абсолютного значення ваги для визначення яскравості кольору для інтерпретації результатів виявлення різних типів кіберзалякувань

Визначення кольору та його яскравості для інтерпретації результатів виявлення типів кіберзалякувань з урахуванням негативного чи позитивного типу впливу на результат

М3. Метод інтерпретації результатів виявлення кіберзалякувань



Діаграми для графічної інтерпретації впливу окремих слів тексту на ймовірність віднесення цього тексту до конкретного типу кіберзалякування



Діаграма з відображенням середнього значення важливості топ 10 слів для всіх класів

М3. Метод інтерпретації результатів виявлення кіберзалякувань

Виявлення та класифікація кіберзалякувань

Оце так новина, чого сюди приперся? Невже дома не сиділось?

Розпізнати

Завантажити файл

Метрики

Надати інтерпретацію результатів

Результати (мультилейблова класифікація):

- Кіберзалякування за віком: 0.012
- Кіберзалякування за гендером: 0.026
- Кіберзалякування за релігією: 0.127
- Кіберзалякування за етнічною ознакою: 0.366
- Не кіберзалякування: 0.234
- Інше: 0.548

Абсолютне значення ваг для інтерпретації результатів виявлення різних типів кіберзалякувань

Гендерне кіберзалякування:
Оце так новина, чого сюди приперся (0.005)? Невже дома (0.012) не сиділось?

Релігійне кіберзалякування:
Оце так новина, чого сюди(0.005) приперся? Невже дома не сиділось? (0.003)

Вікове кіберзалякування:
(0.001) Оце так новина, чого сюди (0.0006) приперся? Невже дома не сиділось?

Етнічне кіберзалякування:
Оце так новина, чого(0.005) сюди приперся? Невже дома не (0.0001)сиділось?

Інший тип кіберзалякування:
Оце (0.08) так новина(0.035) , чого сюди приперся?(0.089) Невже дома не (0.01) сиділось?

Результати аналізу текстового зразка до 100 символів

Виявлення та класифікація кіберзалякувань

Діду, що ти там опять пишеш? Вже б давно на пенсії відпочивав, а не позорився тут. Та кому цікаві твої допотопні думки? Всі нормальні люди вже давно пішли вперед, а ти як той динозавр, що вимерти забув!

Розпізнати

Завантажити файл

Метрики

Надати інтерпретацію результатів

Результати (мультилейблова класифікація):

- Кіберзалякування за віком: 0.321
- Кіберзалякування за гендером: 0.483
- Кіберзалякування за релігією: 0.201
- Кіберзалякування за етнічною ознакою: 0.233
- Не кіберзалякування: 0.326
- Інше: 0.689

Абсолютне значення ваг для інтерпретації результатів виявлення різних типів кіберзалякувань

Гендерне кіберзалякування:
Діду (0.30), що ти там опять пишеш? Вже б давно на пенсії відпочивав, а не позорився (0.25) тут. Та кому цікаві твої допотопні думки? Всі нормальні люди вже давно пішли вперед, а ти як той динозавр, що вимерти (0.30) забув!

Вікове кіберзалякування:
Діду (0.35), що ти там опять пишеш? Вже б давно на пенсії (0.25) відпочивав, а не позорився тут. Та кому цікаві твої допотопні думки? Всі нормальні люди вже давно пішли вперед, а ти як той динозавр, що забув (0.169)!

Релігійне кіберзалякування:
Діду, що ти там (0.20) опять пишеш? Вже б давно на пенсії відпочивав, а не позорився (0.10) тут. Та кому цікаві твої допотопні думки? Всі нормальні люди вже давно пішли вперед, а ти як той динозавр, що вимерти забув(0.056)!

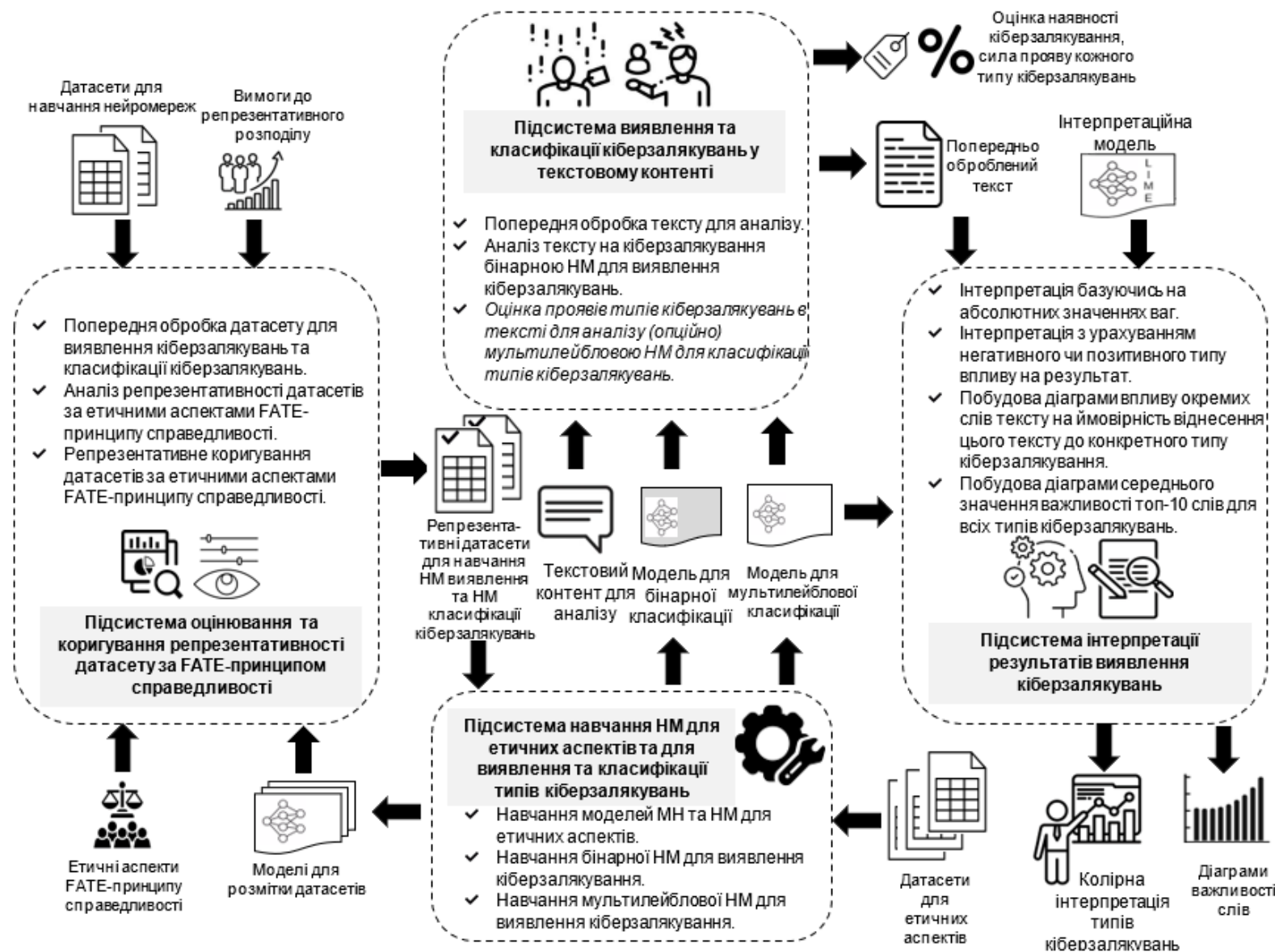
Вікове кіберзалякування:
Діду, що ти (0.046) там опять пишеш? Вже б давно на пенсії (0.118) відпочивав, а не позорився тут. Та кому цікаві твої допотопні думки (0.026)? Всі нормальні люди вже давно пішли вперед, а ти як той динозавр, що вимерти забув!

Етнічне кіберзалякування:
Діду, що ти там (0.015) опять пишеш? Вже б давно (0.062) на пенсії відпочивав, а не позорився (0.002) тут. Та кому цікаві твої допотопні думки? Всі нормальні (0.179) люди вже давно пішли вперед, а ти як той динозавр, що вимерти забув!

Інший тип кіберзалякування:
Діду, що ти там опять пишеш (0.390)? Вже (0.095) б давно на пенсії відпочивав (0.261) , а не позорився тут. Та кому цікаві твої допотопні думки? Всі забияки(0.236)!

Результати аналізу текстового зразка, що містить суржик та сленг

Процес перетворення інформації в прикладній інтелектуальній системі для виявлення та класифікації кіберзалякувань



Авторські свідоцтва



1. а. с. № 132920 Україна. Комп'ютерна програма «Інтелектуальна інформаційна система для оцінювання та коригування репрезентативності текстових датасетів» / О.В. Собко. 2025.
2. а. с. № 132921 Україна. Комп'ютерна програма «Інтелектуальна інформаційна система для виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту» / О.В. Собко. 2025.



3. а. с. № 132042 Україна. «Метод аналізу репрезентативності та формування навчальних вибірок» / О.В. Собко. 2024.
4. а. с. № 132726 Україна. «Метод виявлення кібербулінгу у текстовому контенті» / О.В. Собко. 2025.
5. а. с. № 132727 Україна. «Метод інтерпретації результатів нейромережевого виявлення кібербулінгу у текстовому контенті» / О.В. Собко. 2025.