УДК 004.8
DOI https://doi.org/10.32782/IT/2024-4-23

Олена СОБКО
викладач кафедри комп'ютерних наук, Хмельницький національний університет, вул. Інститутська, 11, Хмельницький, Україна, 29016
ORCID: *0000-0001-5371-5788*
Scopus Author ID: *57456803900*

Бібліографічний опис статті: Собко, О. (2024). Нейромережевий пошук і класифікація кіберзалякувань у текстових повідомленнях. *Information Technology: Computer Science, Software Engineering and Cyber Security,* 4, 197–205, doi: https://doi.org/10.32782/IT/2024-4-23

НЕЙРОМЕРЕЖЕВИЙ ПОШУК І КЛАСИФІКАЦІЯ КІБЕРЗАЛЯКУВАНЬ У ТЕКСТОВИХ ПОВІДОМЛЕННЯХ

У статті висвітлено проблему пошуку і класифікації кіберзалякувань у текстових повідомленнях, що є одним із ключових викликів сучасного інформаційного суспільства. Актуальність дослідження зумовлена необхідністю створення ефективних систем, здатних забезпечувати точне, етичне та прозоре нейромережеве виявлення кіберзалякувань. Особливе значення приділяється адаптації таких систем до чутливих тем, як дискримінація за віковими, етнічними, гендерними та релігійними ознаками.

Мета роботи *полягає у створенні комплексного методу до нейромережевих пошуку і класифікації кіберзалякувань у текстових повідомленнях, який передбачає забезпечення репрезентативності даних у датасеті, що використовується для навчання моделі, дотримання етичного принципу справедливості в розробці моделі та можливість інтерпретації результатів моделі щодо типів виявлених кіберзалякувань.*

Новизна *запропонованого підходу полягає у створенні нового методу, який дозволяє не тільки оцінювати наявність кібербулінгу в текстових повідомленнях, але й визначати з високою точністю прояв кожного з видів кібербулінгу, забезпечуючи формування репрезентативно збалансованих датасетів для навчання нейромережевих моделей, що виконується у три етапи. На першому етапі проводиться оцінка репрезентативності датасету для навчання нейромережевих моделей для задачі виявлення та класифікації кіберзалякувань. Зокрема, метод дозволяє мінімізувати відхилення у розподілі даних за класами, що досягає максимуму лише 0,04%. На другому етапі використовуються моделі нейромережевої класифікації: BiLSTM для бінарної класифікації кіберзалякувань, яка демонструє точність 96%, та BERT для мультилейблової класифікації за різними типами кіберзалякувань з точністю 94%. Третій етап передбачає застосування моделі LIME, яка забезпечує візуальну інтерпретацію рішень нейромережі, дозволяючи користувачам отримати пояснення кожного виявленого типу кіберзалякувань.*

Методологія *дослідження базується на поєднанні сучасних підходів до машинного навчання, якісного аналізу репрезентативності даних та використання інтерпретаційних моделей. Інтеграція цих підходів спрямована на створення прозорих і довірених систем виявлення кіберзалякувань, що можуть бути застосовані у реальних умовах.*

Результати *демонструють ефективність запропонованого методу, який не лише підвищує точність і прозорість процесу виявлення та класифікації кіберзалякувань, але й відповідає Цілям Сталого Розвитку № 5, № 10 та № 16, що дозволяє запропонованому комплексному методу бути релевантним для використання в системах, де етичність і точність є важливими.*

Ключові слова: *кіберзалякування, репрезентативність, інтерпретація результатів, BERT, LIME.*

Olena SOBKO
Teacher at the Department of Computer Science, Khmelnytskyi National University, 11, Instytutska Str., Khmelnytskyi, Ukraine, 29016, olenasobko.ua@gmail.com
ORCID: *0000-0001-5371-5788*
Scopus Author ID: *57456803900*

To cite this article: Sobko, O. (2024). Neiromerezhevyi poshuk i klasyfikatsiia kuberzaliakuvan u tekstovykh povidomlenniakh [Neural network search and classification of cyberbullying in text messages]. *Information Technology: Computer Science, Software Engineering and Cyber Security,* 4, 197–205, doi: https://doi.org/10.32782/IT/2024-4-23

NEURAL NETWORK SEARCH AND CLASSIFICATION OF CYBERBULLYING IN TEXT MESSAGES

The article highlights the problem of searching and classifying cyberbullying in text messages, which is one of the key challenges of the modern information society. The relevance of the study is due to the need to create effective systems capable of ensuring accurate, ethical and transparent neural network detection of cyberbullying. Particular importance is given to adapting such systems to sensitive topics such as discrimination on the basis of age, ethnicity, gender and religion.

***The purpose** of the work is to create a comprehensive method for neural network search and classification of cyberbullying in text messages, which involves ensuring the representativeness of the data in the dataset used to train the model, adhering to the ethical principle of fairness in the development of the model and the ability to interpret the results of the model regarding the types of detected cyberbullying.*

***The novelty** of the proposed approach lies in the creation of a new method that allows not only to assess the presence of cyberbullying in text messages, but also to determine with high accuracy the manifestation of each type of cyberbullying, ensuring the formation of representatively balanced datasets for training neural network models, which is performed in three stages. At the first stage, the representativeness of the dataset for training neural network models for the task of detecting and classifying cyberbullying is assessed. In particular, the method allows minimizing deviations in the distribution of data by classes, which reaches a maximum of only 0.04%. At the second stage, neural network classification models are used: BiLSTM for binary classification of cyberbullying, which demonstrates an accuracy of 96%, and BERT for multi-label classification by different types of cyberbullying with an accuracy of 94%. The third stage involves the application of the LIME model, which provides a visual interpretation of the neural network solutions, allowing users to obtain an explanation for each detected type of cyberbullying.*

***The research** methodology is based on a combination of modern approaches to machine learning, qualitative analysis of data representativeness and the use of interpretive models. The integration of these approaches is aimed at creating transparent and trusted cyberbullying detection systems that can be applied in real-world conditions.*

***The results** demonstrate the effectiveness of the proposed method, which not only increases the accuracy and transparency of the cyberbullying detection and classification process, but also meets Sustainable Development Goals No. 5, No. 10 and No. 16, which allows the proposed comprehensive method to be relevant for use in systems where ethics and accuracy are important.*

***Key words:** cyberbullying, neural networks, interpretation of results, BERT, LIME.*

Актуальність проблеми. Кіберзалякування, або булінг у інтернет-середовищі, є серйозною соціальною проблемою сучасності, яка набуває дедалі більшої актуальності з розвитком інтернет-комунікацій. Явище кіберзалякування представляє собою агресивну, образливу або маніпулятивну поведінку в онлайн-просторі, що завдає значної психологічної шкоди жертві (Teng T. H., 2023, C. 55533-55560). Особливо вразливими до кіберзалякувань є підлітки та молодь, які активно користуються соціальними мережами та месенджерами для спілкування, що може призвести до втрати впевненості в собі, депресії чи навіть суїцидальних думок.

Складність пошуку кіберзалякування обумовлена його багатогранністю та контекстуальною природою. Агресивні дії часто приховані за сарказмом, жартами чи культурними особливостями спілкування, що робить їх важкими для однозначної ідентифікації (Unnava S., 2024, C. 200–206). До того ж, великий обсяг текстової інформації, яка щодня створюється в мережі, значно ускладнює процес аналізу. Нерепрезентативні дані можуть призводити до упередженості моделі (систематичної помилки в прогнозах, яка призводить до несправедливих або некоректних результатів), що ставить під сумнів її ефективність та етичність. Використання традиційних методів є малоефективним, тому виникає потреба у застосуванні нейромережевих технологій, здатних аналізувати тексти з урахуванням їхнього змісту, стилю і контексту, допомагаючи у вирішенні цієї проблеми (Pagano T.P., 2024, C. 15).

Аналіз останніх досліджень і публікацій. Проблема пошуку кіберзалякувань у текстових повідомленнях ускладнюється через різноманіття його форм і проявів: від прямих образ до прихованих дій, таких як сарказм або маніпуляції (Собко О.В., 2024, С. 262–265). Сучасні дослідження часто стикаються з обмеженнями, пов’язаними з нерівномірністю даних і недостатньою увагою до локальних особливостей кіберзалякувань, що підкреслює потребу в нових підходах до їх пошуку та класифікації (Krak I., 2024, C. 16–28).

Наприклад, у (Harish D., 2023, C. 1-6) автори розглядаються алгоритми машинного навчання для автоматизації процесу пошуку кіберзалякувань у соціальних мережах. У дослідженні використано різні моделі машинного навчання, такі як згорткові нейронні мережі, підтримуючі векторні машини та логістична регресія для пошуку кіберзалякувань у текстах. Результати

показали, що модель логістичної регресії з оптимізацією за допомогою Stochastic Average Gradient перевершує інші методи в точності визначення кіберзалякувань.

Стаття (Orrù G., 2023, С. 430) описує розробку технологій для пошуку кіберзалякувань у рамках проєкту BullyBuster. Зокрема, вона зосереджується на застосуванні машинного навчання та обробки природної мови для пошуку та класифікації агресивної поведінки в онлайн-спілкуванні. Підхід поєднує методи аналізу настроїв і контексту для підвищення точності пошуку кіберзалякувань, враховуючи лінгвістичні та поведінкові аспекти.

У статті (Samee, N. A., 2023) автори пропонують комбінувати векторні уявлення слів, емоційні характеристики та федеративне навчання для покращення пошуку кіберзалякувань. Автори стверджують, що векторні уявлення слів покращують розуміння контексту, а емоційні характеристики дозволяють виявляти афективні аспекти тексту. Федеративне навчання дозволяє зберігати конфіденційність даних, навчаючи моделі на розподілених пристроях. Використання моделей BERT, CNN, DNN та LSTM показує, що цей підхід перевищує традиційні методи в точності пошуку кіберзалякувань.

Попри значний прогрес у дослідженнях кіберзалякувань, проведений аналіз показує, що існуючі підходи часто мають суттєві недоліки: нерепрезентативні дані, недостатнє врахування контексту повідомлень і складнощі з ідентифікацією типів кіберзалякувань, що створює потребу у розробці нового, відмінного від існуючих, підходу, який враховуватиме репрезентативне формування датасету з метою навчання нейромережевих моделей для пошуку і класифікації кіберзалякувань у текстових повідомленнях з подальшою інтерпретацією отриманих результатів.

Мета роботи полягає у створенні комплексного методу до нейромережевих пошуку і класифікації кіберзалякувань у текстових повідомленнях, який передбачає забезпечення репрезентативності даних у датасеті, що використовується для навчання моделі, дотримання етичного принципу справедливості в розробці моделі та можливість інтерпретації результатів моделі щодо типів виявлених кіберзалякувань.

Виклад основного матеріалу дослідження. Запропонований у статті комплексний метод до нейромережевих пошуку і класифікації кіберзалякувань у текстових повідомленнях складається з трьох етапів, що наведені на рисунку 1. Наведені етапи забезпечують не лише точний пошук та класифікацію кіберзалякувань, а й враховує етичні принципи справедливості для формування датасету, що призначений для навчання нейромереж з метою пошуку кіберзалякувань. Також важливою є інтерпретація результатів, яка надається для поясненності прийнятих нейромережевою моделюю рішень.

На першому етапі відбувається оцінка та коригування репрезентативності датасету для пошуку та класифікації кіберзалякувань, схема та кроки якого наведені на рисунку 2.

Вхідними даними цього етапу є вибірка текстових даних, яка містить певну кількість елементів, що відповідають визначеним цільовим пропорціям за етичними аспектами. Для

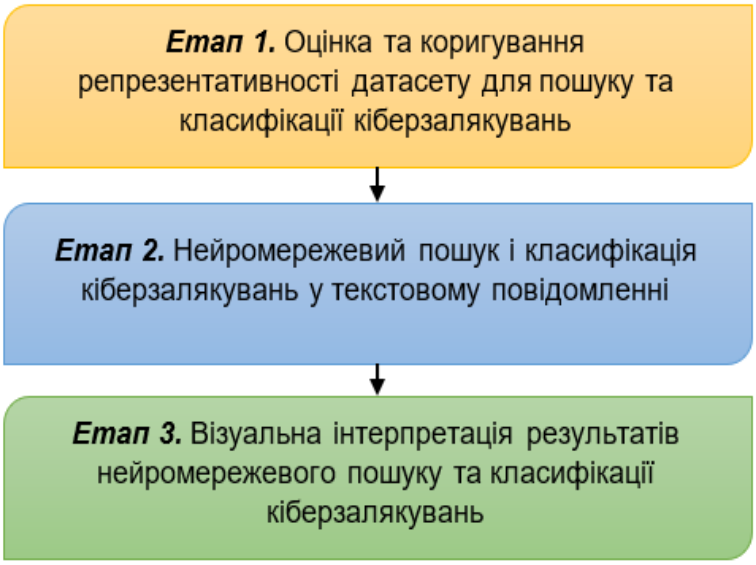

Рис. 1. Схема етапів нейромережевих пошуку і класифікації кіберзалякувань у текстових повідомленнях

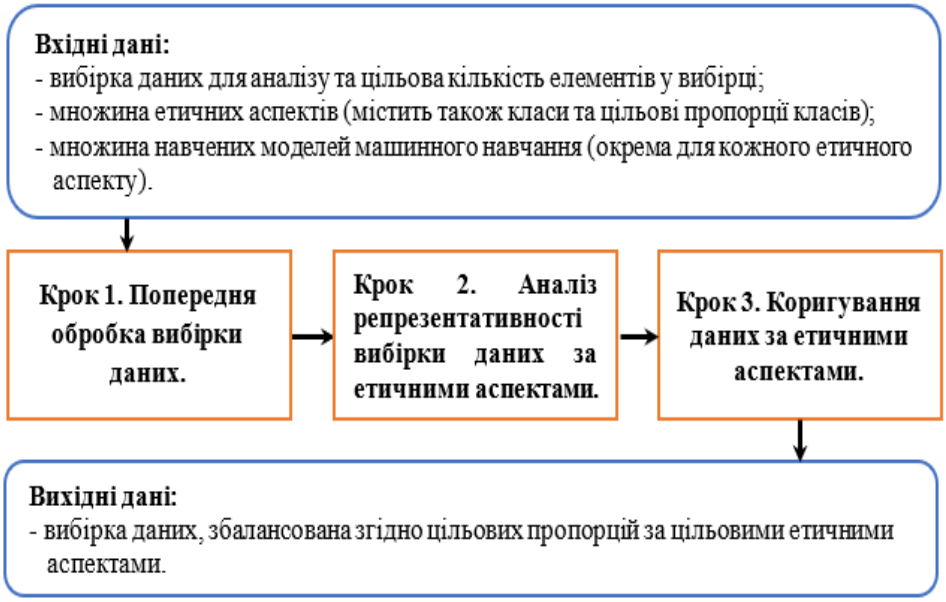

Рис. 2. Схема оцінювання та коригування репрезентативності датасету

кожного аспекту використовується попередньо навчена модель машинного навчання, що виконує аналіз конкретного етичного критерію.

На першому кроці здійснюється попередня обробка даних, яка включає видалення непотрібних елементів (знаків пунктуації, цифр) та некоректних записів (порожніх або неінформативних текстів).

Другий крок полягає в аналізі репрезентативності вибірки, що включає векторизацію елементів, класифікацію за етичними аспектами, оцінку пропорцій класів та пошуку відхилень від заданих пропорцій. Також перевіряється, чи є достатнім обсяг даних для представлення кожного класу.

На третьому кроці здійснюється коригування вибірки, яке включає видалення надлишкових елементів, аугментацію менших за кількість зразків у класах та формування збалансованої вибірки згідно з етичними вимогами. Результатом етапу є вибірка, що відповідає встановленим пропорціям та етичним критеріям.

На другому етапі відбувається нейромережевий пошук і класифікація кіберзалякувань у текстовому повідомленні, схема наведена на рисунку 3.

Перший крок етапу нейромережевого пошуку та класифікації кіберзалякувань у текстовому повідомленні включає попередню обробку вхідного тексту. На цьому кроці текст очищується від зайвих символів, після чого перетворюється у векторне представлення, що підготовлене для обробки нейромережею.

На другому кроці нейромережева модель, яка пройшла навчання для бінарної класифікації кіберзалякувань, аналізує текст на наявність таких ознак, визначаючи інтенсивність проявів кіберзалякування. Якщо рівень прояву кіберзалякування перевищує 50%, текстове повідомлення вважається таким, що містить кіберзалякування, і передається на наступний етап для визначення типу.

Третій крок етапу передбачає використання нейромережевої моделі, яка спеціалізується на визначенні типів кіберзалякувань. Ця модель проводить аналіз текстового повідомлення і визначає відсоткове співвідношення кожного з типів кіберзалякувань.

Результатом цього етапу є загальна оцінка рівня кіберзалякувань у текстовому повідомленні, а також окремі оцінки для кожного типу кіберзалякувань.

На третьому етапі відбувається візуальна інтерпретація результатів нейромережевого пошуку та класифікації кіберзалякувань, схема кроків якого подана на рисунку 4.

Вхідними даними цього етапу є модель трансформерної архітектури, попередньо навчена для мультилейблової класифікації кіберзалякувань, що здатна ідентифікувати різні типи кіберзалякувань, включаючи вікові, гендерні, етнічні, релігійні та загальні категорії. Крім того, використовується інтерпретаційна модель, яка дозволяє пояснити, як окремі слова чи фрази впливають на результати класифікації.

На першому кроці текстове повідомлення проходить через процес токенізації.

Після чого на другому кроці здійснюється прогнозування ймовірностей, що вказують на належність текстового повідомлення до кожного із знайдених типів кіберзалякування.

Рис. 3. Схема нейромережевого пошуку і класифікації кіберзалякувань у текстовому повідомленні

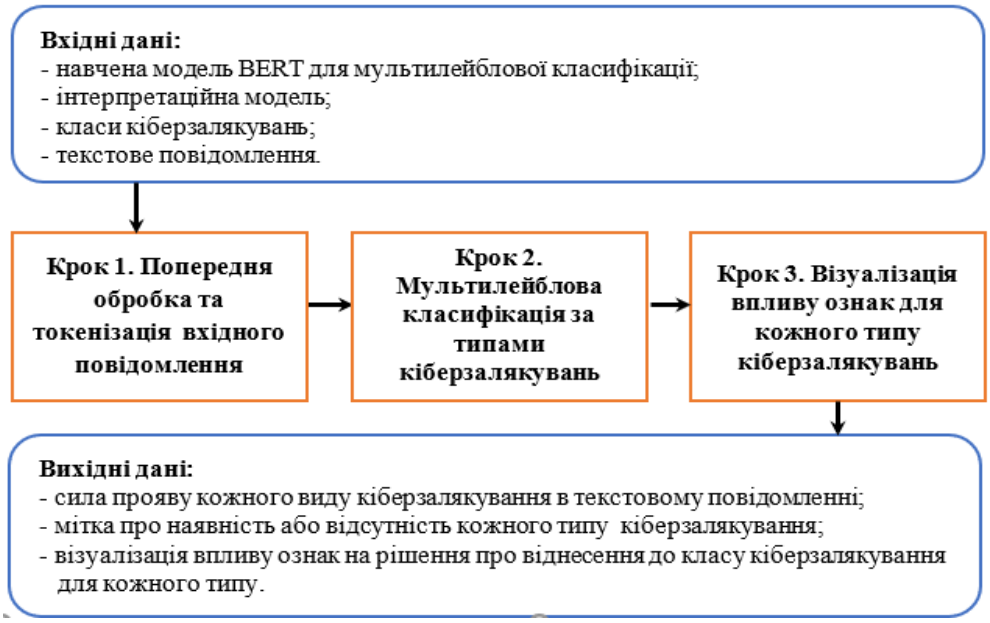

Рис. 4. Схема візуальної інтерпретації результатів нейромережевого пошуку та класифікації кіберзалякувань

Третій крок полягає в інтерпретації результатів, що проводиться за допомогою спеціальної інтерпретаційної моделі, яка візуалізує вплив конкретних слів чи фраз на класифікацію тексту.

Кінцевими результатами цього етапу є ймовірності для кожного типу кіберзалякування та візуальне представлення, що підкреслюють слова, які мали найбільший вплив на класифікацію текстового повідомлення за типами кіберзалякувань.

Таким чином кожен з описаних етапів є частиною запропонованого комплексного методу нейромережевих пошуку і класифікації кіберзалякувань у текстових повідомленнях, що враховує різні групи людей, як за віковими, так і за етнічними чи релігійними ознаками для завдання пошуку типів кіберзалякувань у текстовому повідомленні, а також надає пояснення рішень моделі щодо визначених у текстовому повідомленні типів кіберзалякувань шляхом візуальної інтерпретації.

Експеримент, результати та дискусія. З метою оцінки ефективності запропонованого підходу до нейромережевого визначення цілей пропаганди у текстовому контенті з візуальною аналітикою було розроблено програмне забезпечення. Для апробації методу були використанні набори «Cyberbullying Classification» (Kaggle. Cyberbullying Classification) та «Cyberbullying Detection Dataset» (Kaggle. Cyberbullying Detection Dataset). Однак ці датасети не мають міток щодо статі, віку, релігії та етнічності авторів повідомлень, тому для навчання моделей машинного навчання, що розмічають вхідні дані, використовувались додаткові датасети, які враховують три етичні аспекти принципу справедливості: гендер, вік та релігію (Kaggle.com. Tweet Files for Gender Guessing, Live.european-language-grid.eu. TAG-it Dataset Distribution, Kaggle.com. Cyberbullying Tweets). Оскільки класи в цих датасетах були нерівномірно представлені, вони були збалансовані за чисельністю, щоб уникнути негативного впливу на якість навчання моделей. Кількість зразків для кожного класу в навчальних вибірках для EML за етичними аспектами показана на рисунку 5.

У якості моделей машинного навчання, які б здійснювали розмітку вхідного датасету за етичними аспектами обрано різні архітектури: класифікатор SVM та моделі глибокого навчання BERT, LSTM. Результати обчислення статичних метрик кращих моделей машинного навчання для вікового, гендерного та релігійно аспектів наведено в таблиці 1

Для різних класів було досягнуто різних рівнів роздільної здатності: для релігійної ознаки, при використанні класифікатора BERT, який показав найкращі результати серед навчених моделей для класифікації текстів за релігійним аспектом, дані виявились добре роздільними. За гендерною ознакою, де застосовувався класифікатор LSTM, що продемонстрував кращу ефективність у порівнянні з іншими моделями, дані показали середню роздільність. Водночас, за віковою ознакою, використовуючи класифікатор SVM, дані виявились погано роздільними.

У процесі формування репрезентативного датасету з урахуванням вікового та гендерного етичних аспектів, на основі демографічних підгруп населення України (Idss. Національні демографічні прогнози), було застосовано

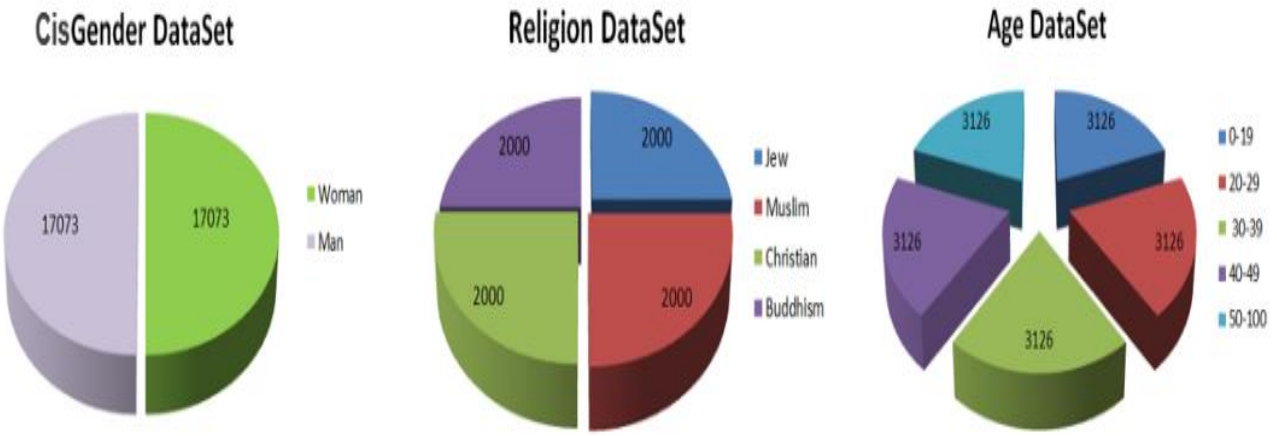

Рис. 5. Класи та кількість зразків у датасетах для навчання моделей машинного навчання

Таблиця 1

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

Модель машинного навчання	*Accuracy*	*Precision*	*Recall*	*F1-score*
Релігійний етичний аспект				
BERT	0.910	0.980	0.74 0	0.840
Гендерний етичний аспект				
LSTM	0.70	0.770	0.670	0.720
Віковий етичний аспект				
SVM	0.815	0.770	0.779	0.770

метод аугментації для створення збалансованої вибірки текстових даних (таблиця 2).

Відхилення розподілу зразків за віковими та гендерними класами в коригованому датасеті були мінімальними. Зокрема, мінімальне відхилення становило 0.00%, максимальне – 0.04%, а середнє – 0.02%. Це свідчить про високу точність коригування вибірки, що дозволяє забезпечити достатнє представлення кожної категорії без значних диспропорцій.

Для нейромережевих пошуку та класифікації кіберзалякувань у текстових повідомленнях сформований датасет використовується в кілька етапів. Спочатку застосовується модель для бінарної класифікації, яка визначає загальний рівень кіберзалякувань у текстовому повідомленні. Для цієї задачі використано нейромережу BiLSTM, яка досягла високих показників ефективності: Accuracy 0.96, Precision 0.91, Recall 0.93 та F1 Score 0.92. Після цього здійснюється мультилейблова класифікація для визначення конкретних типів кіберзалякувань. Для цього обрано модель BERT, що показала також відмінні результати: Accuracy 0.94, Precision 0.93, Recall 0.93 та F1 Score 0.93. Модель була навчена для пошуку таких типів кіберзалякувань, як вікове, гендерне, релігійне, етнічне та інші.

У дослідженні запропонованого методу було використано текстовий зразок «Your God has no place here. Stick to your country and stop dragging your outdated traditions and religions into ours». Модель BERT визначила ймовірності наявності вікових, етнічних, гендерних, релігійних та інших типів кіберзалякувань у текстовому зразку, зокрема релігійне кіберзалякування мало ймовірність 99.86%, тоді як інші види були значно менш представлені.

Результати мультилейблової класифікації типів кіберзалякувань, отримані за допомогою моделі BERT, використовуються на етапі візуальної інтерпретації за допомогою LIME. Це дозволяє отримати візуальне представлення, де слова виділяються кольорами в залежності від їх впливу на класифікацію, з більш насиченими кольорами для слів із більшим впливом. Додатково, яскравість слів коригується для відображення позитивного або негативного впливу, що дозволяє точніше пояснити прийняті моделі рішення (рисунок 6).

Запропонована інтерпретація результатів нейромережевих пошуку і класифікації кіберзалякувань є важливим інструментом візуальної аналітики, який сприяє забезпеченню прозорості та етичності в рішеннях штучного інтелекту. Вона дозволяє краще зрозуміти, чому і як були прийняті ті чи інші рішення моделлю, що є важливим кроком для запобігання упередженості та неетичних впливів у таких системах.

Висновки. Розроблено комплексний метод нейромережевих пошуку і класифікації кіберзалякувань у текстових повідомленнях, який складається з трьох етапів, спрямованих на аналіз текстового повідомлення з урахуванням етичних аспектів, таких як вікові та етнічні ознаки.

Перший етап передбачає оцінку та коригування репрезентативності датасету, що забезпечує відповідність вибірки етичним вимогам, з мінімальними відхиленнями від ідеального розподілу (0.00%-0.04%). Це дозволяє адаптувати датасет до репрезентативного для подальшого аналізу.

Другий етап передбачає нейромережевий пошук і класифікацію кіберзалякувань у текстовому повідомленні, що забезпечує не лише загальну оцінку рівня кіберзалякувань, але й мультилейблову класифікацію за різними типами, такими як вікові, релігійні та етнічні. Моделі BiLSTM для бінарної класифікації та BERT для мультилейблової класифікації показали високі результати: точність 96% та 94% відповідно.

Таблиця 2

Розподіл зразків у створеній репрезентативній вибірці після проведення аугментації

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%

Рис. 6. Інтерпретація результатів нейромережевих пошуку і класифікації типів кіберзалякувань у текстових повідомленнях

Третій етап надає візуальну інтерпретацію результатів пошуку, що забезпечує пояснення рішень моделі через візуальні інструменти. Такий підхід сприяє підвищенню прозорості та довіри до систем штучного інтелекту, особливо для чутливих тем, як кіберзалякування.

Отже, запропонований комплексний метод нейромережевих пошуку і класифікації кіберзалякувань у текстових повідомленнях враховує соціальні фактори, що дозволяє ефективно здійснювати пошук та класифікацію кіберзалякувань та забезпечує поясненність рішень моделі.

ЛІТЕРАТУРА:

1. Teng T. H., Varathan, K. D. Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. *IEEE Access,* vol. 11*,* 2023. С. 55533–55560.

2. Unnava S., Parasana S. R. A Study of Cyberbullying Detection and Classification Techniques: A Machine Learning Approach. *Engineering, Technology & Applied Science Research,* 14(4), 2024. P. 15607–15613.

3. Pagano T. P., Loureiro R. B., Lisboa F.V.N., Peixoto R. M., Guimarães G.A.S., Cruz G.O.R., Araujo M. M., Santos L. L., Cruz M.A.S., Oliveira E.L.S. Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. *Big Data Cogn. Comput.,* 7(1), 2023. P. 15.

4. Собко О. В. Метод інтелектуального пошуку кіберзалякувань у текстовому контенті. *Розвитки інформаційно-керуючих систем та технологій: монографія. Львів-Торунь: Lina-Pres,* 2024. С. 267–287.

5. 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.* Vol. 3688, 2024. С. 16–28.

6. Harish D., Alamelu M., Manimaran M. Automatic Detection of Cyberbullying on Social Media Using Machine Learning. *In 2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, 2023. С. 1–6.

7. Orrù G., Galli A., Gattulli V., Gravina M., Micheletto M., Marrone S., Sansone C. Development of Technologies for the Detection of (Cyber) Bullying Actions: The BullyBuster Project*. Information,* 14(8), 430, 2023.

8. Samee N. A., Khan U., Khan S., Jamjoom M. M., Sharif M., Kim D. H. Safeguarding Online Spaces: A Powerful Fusion of Federated Learning, *Word Embeddings, and Emotional Features for Cyberbullying Detection. IEEE Access, vol. 11,* 2023. С. 124524–124541.

9. Kaggle.com. Cyberbullying Classification, 2021. URL: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification?resource=download. Дата останнього звернення: 2024/12/02.

10. Kaggle.com. CyberBullying Detection Dataset, 2024. URL: https://www.kaggle.com/datasets/sayankr007/cyber-bullying-data-for-multi-label-classification. Дата останнього звернення: 2024/12/02.

11. Kaggle.com. Tweet Files for Gender Guessing, 2019. URL: https://www.kaggle.com/datasets/aharless/tweet-files-for-gender-guessing. Дата останнього звернення: 2024/12/02.

12. Live.european-language-grid.eu. TAG-it Dataset Distribution, 2024. URL: https://live.european-language-grid.eu/catalogue/corpus/8112/download. Дата останнього звернення: 2024/12/02.

13. Cyberbullying Tweets. URL: https://www.kaggle.com/datasets/soorajtomar/cyberbullying-tweets. Last accessed: 2024/10/27. Дата останнього звернення: 2024/12/02.

14. Idss.org.ua. Національні демографічні прогнози 2023. URL: https://idss.org.ua/forecasts/nation_pop_proj. Дата останнього звернення: 2024/12/02.

REFERENCES:

1. Teng, T. H., Varathan, K. D. (2023). Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. *IEEE Access*, vol. 11, C. 55533–55560.

2. Unnava, S., Parasana, S. R. (2024). A Study of Cyberbullying Detection and Classification Techniques: A Machine Learning Approach. Engineering*, Technology & Applied Science Research,* 14(4), P. 15607–15613.

3. Pagano, T. P., Loureiro, R. B., Lisboa, F.V.N., Peixoto, R. M., Guimarães, G.A.S., Cruz, G.O.R., Araujo, M. M., Santos, L. L., Cruz, M.A.S., Oliveira, E.L.S. (2023). Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. *Big Data Cogn. Comput.,* 7(1), P. 15.

4. Sobko, O. V. (2024). Doslidzhenia efektyvnosti metodu otsiniuvannia ta koryhulyzatsii reprezentatyvnosti datasetu za FATE-pryntsypom spravedlyvosti [Research on the effectiveness of the method for assessing and adjusting the representativeness of a dataset according to the FATE principle of fairness]. *Perspektyvy suchasnoi nauky: teoriia i praktyka: materialy VIII Mizhnarodnoyi naukovo-praktychnoi konferentsii,* 217–221 [in Ukrainian].

5. Sobko, O. V. (2024). Metod intelektualnoho vyiavlennia kyberzaliakuvan v tekstovomu kontenti [A method for intelligently detecting cyberbullying in text content]. *Rozvytok informatsiino-kervuiuchykh system ta tekhnolohii: monohrafiia. L'viv-Torun: Lina-Pres,* 267–287 [in Ukrainian].

6. Krak, I., Zalutska, O., Molchanova, M., Mazurets, O., Bahrii, R., Sobko, O., & Barmak, O. (2024). Abusive speech detection method for Ukrainian language using recurrent neural network. *CEUR Workshop Proceedings, 3688*, 16–28.

7. Harish, D., Alamelu, M., & Manimaran, M. (2023). Automatic detection of cyberbullying on social media using machine learning. *In 2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA),* 1–6.

8. Orrù, G., Galli, A., Gattulli, V., Gravina, M., Micheletto, M., Marrone, S. & Sansone, C. (2023). Development of technologies for the detection of (cyber) bullying actions: The BullyBuster project. *Information,* 14(8), 430.

9. Samee, N. A., Khan, U., Khan, S., Jamjoom, M. M., Sharif, M., & Kim, D. H. (2023). Safeguarding online spaces: A powerful fusion of federated learning, word embeddings, and emotional features for cyberbullying detection. *IEEE Access,* 11, 124524–124541.

10. Kaggle.com. Cyberbullying Classification, 2021. Retrieved from: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification?resource=download. Last accessed: 2024/12/02.

11. Kaggle.com. CyberBullying Detection Dataset, 2024. Retrieved from: https://www.kaggle.com/datasets/sayankr007/cyber-bullying-data-for-multi-label-classification. Last accessed: 2024/12/02.

12. Kaggle.com. Tweet Files for Gender Guessing, 2019. Retrieved from: https://www.kaggle.com/datasets/aharless/tweet-files-for-gender-guessing.

13. Live.european-language-grid.eu. TAG-it Dataset Distribution, 2024. Retrieved from: https://live.european-language-grid.eu/catalogue/corpus/8112/download. Last accessed: 2024/12/02.

14. Cyberbullying Tweets. Retrieved from: https://www.kaggle.com/datasets/soorajtomar/cyberbullying-tweets. Last accessed: 2024/10/27. Last accessed: 2024/12/02.

15. Idss.org.ua. Natsionalni demohrafichni prohnozy 2023. [National demographic projections 2023]. Retrieved from: https://idss.org.ua/forecasts/nation_pop_proj. Last accessed 2024/10/27 [in Ukrainian].