https://doi.org/10.31891/2219-9365-2024-80-18
УДК 004.8

СОБКО Олена
Хмельницький національний університет
https://orcid.org/0000-0001-5371-5788
e-mail: olenasobko.ua@gmail

ВИЯВЛЕННЯ ТА КЛАСИФІКАЦІЯ КІБЕРЗАЛЯКУВАНЬ У ЦИФРОВИХ ТЕКСТАХ ЗАСОБАМИ ШТУЧНОГО ІНТЕЛЕКТУ

У статті запропоновано комплексний підхід до виявлення та класифікації кіберзалякувань у цифрових текстах за допомогою штучного інтелекту. Підхід складається з трьох етапів: оцінювання та коригування репрезентативності датасету з урахуванням етичних критеріїв, нейромережевого виявлення та класифікації кіберзалякувань за різними типами (віковими, релігійними, етнічними, гендерними тощо), а також візуальної інтерпретації результатів моделі. Підхід дозволяє забезпечити неупередженість та відповідність етичним вимогам, а також надає пояснення рішень моделі щодо кожного виявленого типу кіберзалякування, що підвищує прозорість і довіру до систем штучного інтелекту. Результати дослідження підтверджують ефективність підходу, зокрема точність не нижче 94% для моделей BiLSTM і BERT для виявлення та класифікації кіберзалякувань, а також успішну адаптацію текстових датасетів до репрезентативних розподілів.

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

SOBKO Olena
Khmelnytskyi National University

DETECTION AND CLASSIFICATION OF CYBERBULLYING IN DIGITAL TEXTS USING ARTIFICIAL INTELLIGENCE

A comprehensive approach to detecting and classifying cyberbullying in digital text using artificial intelligence has been developed. This approach consists of three key stages, each addressing specific challenges in cyberbullying detection. The first stage involves evaluating and adjusting the representativeness of the dataset to ensure ethical balance with respect to age, ethnicity, and gender. This is essential to prevent bias in the model's decision-making process. The analysis of dataset representativeness showed minimal deviations from the ideal distribution, with a maximum deviation of 0.04%, confirming the effectiveness of data preprocessing. The second stage uses a neural network-based model for detecting and classifying cyberbullying. This multi-label classification model evaluates the overall level of cyberbullying in a text and classifies it into different types, such as age-related, religious, ethnic, and gender-based bullying. The BiLSTM model, used for binary classification, achieved an accuracy of 0.96, precision of 0.96, recall of 0.96, and an F1 score of 0.96. The BERT model, used for multi-label classification, demonstrated an accuracy of 0.94, precision of 0.93, recall of 0.93, and an F1 score of 0.93. The third stage focuses on the visual interpretation of the model's decisions, providing detailed explanations for each detected type of cyberbullying. This is achieved by combining a transformer-based multi-label classifier with an interpretative machine learning model, enhancing transparency and understanding of the model's reasoning. The proposed approach not only improves the detection of cyberbullying but also ensures fairness and transparency in AI systems. By addressing ethical concerns and providing interpretable results, this approach contributes to building trust in AI systems, especially in sensitive areas like cyberbullying detection. It provides a robust framework for ethical and effective cyberbullying detection in textual data.

Keywords: cyberbullying, representativeness, interpretation of results, BERT, LIME.

ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ'ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ЧИ ПРАКТИЧНИМИ ЗАВДАННЯМИ

Кіберзалякування в цифрових текстах є однією з найактуальніших проблем сучасного інформаційного суспільства, що впливає на емоційний, психологічний і соціальний стан користувачів цифрових платформ [1]. Масштабне використання соціальних мереж, форумів та інших комунікаційних ресурсів значно ускладнило моніторинг і контроль над поширенням образливого чи дискримінаційного контенту. Також, варто зазначити, що люди можуть одночасно піддаватися різним типам кіберзалякувань, наприклад, дискримінації за етнічною приналежністю, гендерною ознакою чи віком, що значно посилює негативний вплив на їхній психологічний стан та соціальну взаємодію [2].

Така ситуація породжує потребу у впровадженні ефективних систем для автоматичного виявлення та класифікації кіберзалякувань. Зокрема, мова йде про створення інструментів, здатних розпізнавати різноманітні типи кіберзалякувань в текстових повідомленнях, таких як залякування на основі вікових, етнічних, релігійних чи гендерних ознак. Виявлення кіберзалякувань є завданням у галузі обробки природної мови, яке передбачає розробку моделей машинного навчання, здатних аналізувати контекст тексту, розпізнавати приховану агресію та враховувати культурні й мовні особливості, які можуть вказувати на наявність кіберзалякувань [3].

Проте досягнення високої точності моделей часто супроводжується проблемами інтерпретації їхніх результатів, що викликає сумніви щодо їхнього застосування у чутливих і соціально значущих сферах, таких як виявлення кіберзалякувань. У цьому контексті особливе значення має здатність моделі пояснювати свої

висновки, оскільки прозорість роботи алгоритмів є ключовою умовою для формування довіри користувачів до рішень, запропонованих штучним інтелектом [4].

Важливим є і той факт, що погляди молоді, жінок та представників етнічних меншин нерідко ігноруються у розробці моделей машинного навчання, що призначені для виявлення кіберзалякувань, що не забезпечує етичний та справедливий підхід до розробки систем штучного інтелекту, а отже призводить до упередженості. Ці групи частіше зазнають кіберзалякувань, проте їхній досвід залишається недостатньо вивченим. Останні дослідження свідчать, що кіберзалякування у цих груп може мати специфічні прояви, а його вплив виявляється більш руйнівним порівняно з іншими соціальними категоріями [5].

Отже, виявлення та класифікація кіберзалякувань у цифрових текстах має ґрунтуватися на комплексному підході, який враховує різні групи людей, як за віковими, так і за етнічними чи віковими ознаками, а також надаватиметься пояснення рішень моделі щодо визначених у текстовому контенті типів кіберзалякувань.

АНАЛІЗ ДОСЛІДЖЕНЬ ТА ПУБЛІКАЦІЙ

На сьогодні для виявлення кіберзалякувань використовуються два основні підходи. Перший – це бінарна класифікація, яка спрямована на розподіл тексту на два класи: «кіберзалякування» та «не кіберзалякування» [6]. Другий – мультикласовий підхід, що передбачає класифікацію тексту за різними типами кіберзалякувань. Цей підхід дозволяє деталізувати аналіз, виділяючи класи, наприклад, залякування за ознаками релігії, віку, статі чи етнічної приналежності. У зв'язку з цим багато дослідників використовують моделі машинного та глибокого навчання для вирішення завдань виявлення та класифікації кіберзалякувань у текстових повідомленнях.

Дослідження [7] аналізує ефективність методів машинного і глибокого навчання для цієї задачі. Серед протестованих моделей, таких як Random Forest, XgBoost, Naive Bayes, SVM, CNN, RNN та BERT, найвищі результати продемонструвала модель BERT. Її точність досягла 88,8% у задачі бінарної класифікації та 86,6% у мультикласовій класифікації.

Інше дослідження [8] акцентує увагу на проблемі дисбалансу класів у наборах даних для виявлення кіберзалякувань, що створює значні труднощі для алгоритмів машинного навчання, особливо в задачах бінарної класифікації. Для вирішення цієї проблеми застосовуються методи корекції дисбалансу, зокрема SMOTE та випадкове надсемплювання. Результати показують, що ефективність цих технік залежить від характеристик датасету, зокрема рівня дисбалансу та обсягу даних, а також від обраного класифікатора. Використання балансування класів із застосуванням класифікатора SVC дозволило досягти точності 99%.

У дослідженні [9] була представлена нова теорія для виявлення кіберзалякувань, у межах якої протестували моделі Support Vector Machine, Naive Bayes і Logistic Regression у поєднанні з різними техніками обробки природної мови. Згідно з результатами, точність виявлення кіберзалякувань значно покращується завдяки використанню аналізу настроїв, аналізу N-грам та альтернативних методів виділення ознак, таких як TF-IDF і ідентифікація ненормативної лексики. Запропонований комбінований підхід дозволив досягти точності 75,17%.

У дослідженні [10] було розроблено новий підхід до виявлення та класифікації кіберзалякувань у текстах із соціальних мереж, який базується на поєднанні моделей BERT і SVM з використанням пошуку по сітці для задачі багатокласової класифікації. Порівняльний аналіз із іншими методами машинного та глибокого навчання показав, що запропонована модель забезпечує точність 90% на тестових даних, перевершуючи альтернативні підходи. Для пояснення прогнозів моделі застосовано метод SHAP.

На основі проведеного аналізу можна зробити висновок, що сучасні підходи до виявлення кіберзалякувань орієнтуються на інтеграцію різних методів аналізу, що включає поєднання класичних технік машинного навчання, таких як SVM і Naive Bayes, з методами глибокого навчання [11].

У статті пропонується підхід до виявлення та класифікації кіберзалякувань у цифрових текстах засобами штучного інтелекту шляхом з врахуванням репрезентативного представлення популяції населення у датасеті для навчання моделі з метою забезпечення етичного принципу справедливості розроблюваної моделі та інтерпретацією отриманих моделлю результатів щодо виявлених типів кіберзалякувань, що є відмінним підходом від описаних вище.

ФОРМУЛЮВАННЯ ЦІЛЕЙ СТАТТІ

Метою роботи є: розробка підходу до виявлення та класифікації кіберзалякувань у цифрових текстах засобами штучного інтелекту, що забезпечує репрезентативне представлення даних у датасеті для навчання моделі з метою забезпечення етичного принципу справедливості розроблюваної моделі та інтерпретацією отриманих моделлю результатів щодо виявлених типів кіберзалякувань.

ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ

Запропонований у статті підхід до виявлення та класифікації кіберзалякувань у цифрових текстах засобами штучного інтелекту складається з послідовних етапів, які забезпечують виявлення типів

кіберзалякувань з врахуванням етичних принципів справедливості та інтерпретації наданих нейромережевою моделлю результатів щодо виявлених типів кіберзалякувань. Схема запропонованого підходу подана на рисунку 1.

Вхідними даними першого етапу оцінювання та коригування репрезентативності датасету для виявлення кіберзалякувань є текстовий датасет для виявлення та класифікації типів кіберзалякувань, що підлягає оцінюванню на репрезентативність та корекцію відповідно до етичних аспектів, як от віковий аспект, гендерний, етнічний, релігійний, тощо. Крім цього, вхідними є попередньо навчені моделі машинного навчання, що навчені для оцінки етичних аспектів вхідного датасету з кіберзалякувань, що використовується для навчання нейромережевої моделі для виявлення та класифікації типів кіберзалякувань. Також важливими є вимоги, до яких має бути приведений вхідний датасет. У якості таких вимог виступає пропорційне співвідношення популяції за обраними критеріями (наприклад вік, стать, тощо) та пропорціями зразків за наведеними критеріями. Усі ці дані використовуються для створення репрезентативного, неупередженого текстового датасету, який використовується на етапі 2 з метою навчання нейромережевої моделі для виявлення та класифікації типів кіберзалякувань.

Рис. 1. Схема підходу до виявлення та класифікації кіберзалякувань у цифрових текстах

На другому етапі нейромережевого виявлення і класифікації кіберзалякувань у текстовому контенті вхідними даними є репрезентативний, неупереджений за етичними аспектами текстовий датасет, створений на попередньому етапі, текстовий зразок для дослідження, а також нейромережева модель, яка призначенна для бінарної класифікації на наявність або відсутність ознак кіберзалякувань та нейромережева модель для класифікації текстових зразків, у яких виявлено ознаки кіберзалякувань за типами кіберзалякувань. Ці моделі дозволяють ефективніше аналізувати текстовий контент та визначати, до якого з типів кіберзалякувань належать певні висловлювання. Вихідними даними даного етапу є загальна оцінка прояву кіберзалякувань в тексті, а також оцінки наявності кожного з типів кіберзалякувань у вхідному текстовому зразку

На третьому етапі візуальної інтерпретації результатів нейромережевого виявлення кіберзалякувань отримані з попереднього етапу оцінки прояву кожного з типів кіберзалякувань використовуються для інтерпретації висновків, що були надані нейромережевою моделлю для виявлення типів кіберзалякувань.

Таким чином кожен з описаних етапів є частиною запропонованого підходу для виявлення та класифікації кіберзалякувань у цифрових текстах засобами, який є комплексним підходом, що враховує різні групи людей, як за віковими, так і за етнічними чи віковими ознаками для завдання виявлення типів кіберзалякувань у текстовому контенті, а також надає пояснення рішень моделі щодо визначених у текстовому контенті типів кіберзалякувань шляхом візуальної інтерпретації.

Етап оцінювання та коригування репрезентативності датасету для виявлення кіберзалякувань призначений для аналізу та створення репрезентативних вибірок текстових даних, орієнтованих на принцип справедливості FATE у різних предметних областях. Він дозволяє оцінювати репрезентативність вибірок з точки зору етичних критеріїв та здійснювати відповідне коригування датасету, забезпечуючи його відповідність етичним вимогам. У процесі коригування вирішується оптимізаційна задача, що включає ідентифікацію надлишкових елементів для видалення, а також формування специфічних вимог щодо етичної приналежності кожного елемента для подальшої аугментації даних. Схема етапу наведена на рисунку 2.

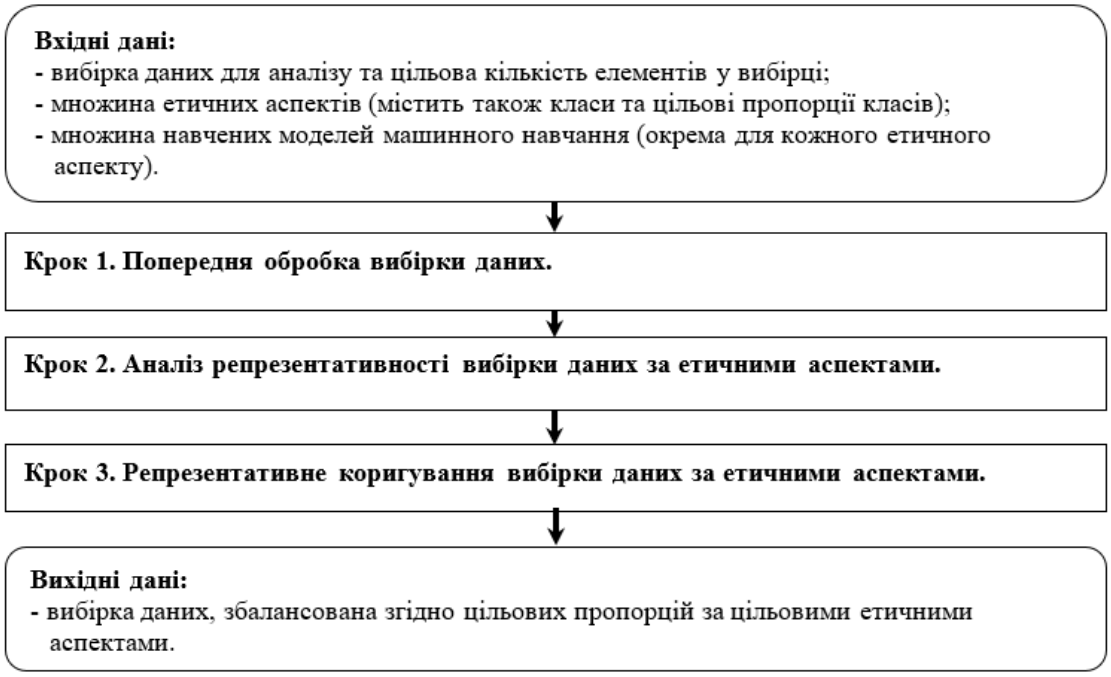

Рис. 2. Кроки оцінювання та коригування репрезентативності датасету для виявлення кіберзалякувань

Вхідними даними для етапу є вибірка текстових даних, що містить визначену цільову кількість елементів. Ця вибірка включає множину етичних аспектів, які представляють класи з відповідними цільовими пропорціями. Крім того, використовуються попередньо навчені моделі машинного навчання, кожна з яких відповідає за аналіз окремого етичного аспекту.

На першому кроці відбувається попередня обробка даних, яка включає видалення неінформативних фрагментів (наприклад, знаків пунктуації, цифр) та некоректних записів (порожніх або беззмістовних).

Другий крок передбачає аналіз репрезентативності вибірки. Здійснюється векторизація кожного елемента, класифікація даних за етичними аспектами, оцінка пропорцій класів і виявлення відхилень від цільових пропорцій. Також перевіряється достатність даних для мінімальної представленості кожного класу.

Третій крок включає репрезентативне коригування вибірки, яке передбачає видалення надлишкових елементів, аугментацію недостатніх класів та формування збалансованої вибірки відповідно до цільових вимог.

Результатом є вибірка даних, яка відповідає цільовим пропорціям та етичним критеріям.

Отже, в результаті виконання кроків етапу оцінювання та коригування репрезентативності датасету для виявлення кіберзалякувань буде сформовано датасет, який є відображає пропорційне до реальних демографічних підгруп популяції та буде недискримінаційним і неупередженим.

Етап нейромережевого виявлення і класифікації кіберзалякувань у текстовому контенті надає змогу аналізувати текстове повідомлення для визначення узагальненого рівня прояву кіберзалякувань, а також виконувати мультилейблову класифікацію, забезпечуючи окремі показники для кожного типу кіберзалякувань, що дозволяє оцінювати рівень прояву таких типів, як вікові, релігійні, етнічні, гендерні та інші форми кіберзалякувань [12]. Схема та кроки етапу наведена на рисунку 3.

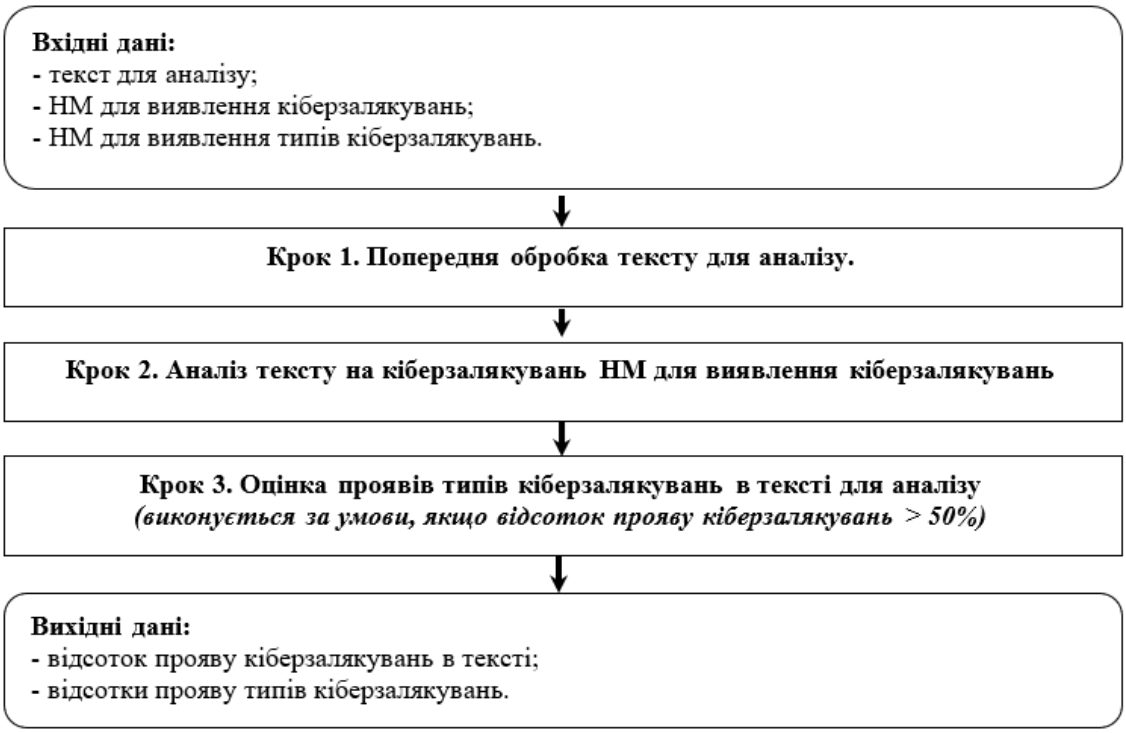

Рис. 3. Кроки нейромережевого виявлення і класифікації кіберзалякувань у текстовому контенті

Перший крок етапу нейромережевого виявлення і класифікації кіберзалякувань у текстовому контенті полягає у попередній обробці тексту. Вхідний текстовий зразок очищується від зайвих символів, таких як пунктуація, смайли та пробіли, після чого перетворюється у векторне представлення для роботи з нейромережею.

На другому кроці нейромережева модель, що навчена бінарній класифікації кіберзалякувань аналізує текст на наявність кіберзалякувань, визначаючи їхній рівень прояву. Якщо цей рівень перевищує 50%, текст вважається таким, що містить ознаки кіберзалякувань, і передається для визначення їх типів на крок 3.

Третій крок передбачає використання нейромережевої моделі, що навчена для виявлення типів кіберзалякувань. Модель аналізує текст, визначаючи частку кожного типу кіберзалякувань.

На виході етапу формується узагальнена оцінка рівня кіберзалякувань у тексті, а також оцінки для кожного типу кіберзалякувань.

Отже, наведені кроки етапу нейромережевого виявлення і класифікації кіберзалякувань у текстовому контенті дозволять ідентифікувати ознаки кіберзалякувань в тексті, за також провести мультилейблову класифікацію, яка вказує на оцінку наявності різних типів кіберзалякувань у одному текстовому зразку.

Етап візуальної інтерпретації результатів нейромережевого виявлення кіберзалякувань спрямований на пояснення рішень нейромережевої моделі щодо виявлених типів кіберзалякувань у тексті, унікальність якого полягає в інтерпретації результатів окремо для кожного виявленого типу кіберзалякування. Забезпечується це завдяки використанню мультилейблового класифікатора на базі трансформерної архітектури та спеціалізованої інтерпретаційної моделі машинного навчання. Схема та кроки етапу подана на рисунку 4.

Вхідними даними цього етапу є навчена модель трансформерної архітектури для мультилейблової класифікації, здатна розпізнавати різні типи кіберзалякувань, такі як вікові, етнічні, гендерні, релігійні та узагальнений тип, що охоплює інші види кіберзалякувань. Також використовується інтерпретаційна модель, яка пояснює вплив окремих слів чи фраз на результати класифікації. Вхідний текст аналізується на наявність ознак кіберзалякувань, з подальшою інтерпретацією результатів.

На першому кроці текст розбивається токенізатором на окремі елементи, які перетворюються у числові послідовності для роботи нейромережевої моделі [13].

На другому кроці модель прогнозує ймовірності належності тексту до кожного класу кіберзалякувань, оцінюючи наявність ознак типів, як-от вікові, етнічні чи гендерні.

На третьому кроці результати класифікації пояснюються та візуалізуються за допомогою інтерпретаційної моделі, що виявляє вплив окремих слів або фраз на ідентифікацію ознак кіберзалякувань.

Вихідними даними є ймовірності прояву кожного виду кіберзалякувань у тексті, а також графічна візуалізація, яка підсвічує важливі слова, що вплинули на рішення моделі для кожного типу кіберзалякування.

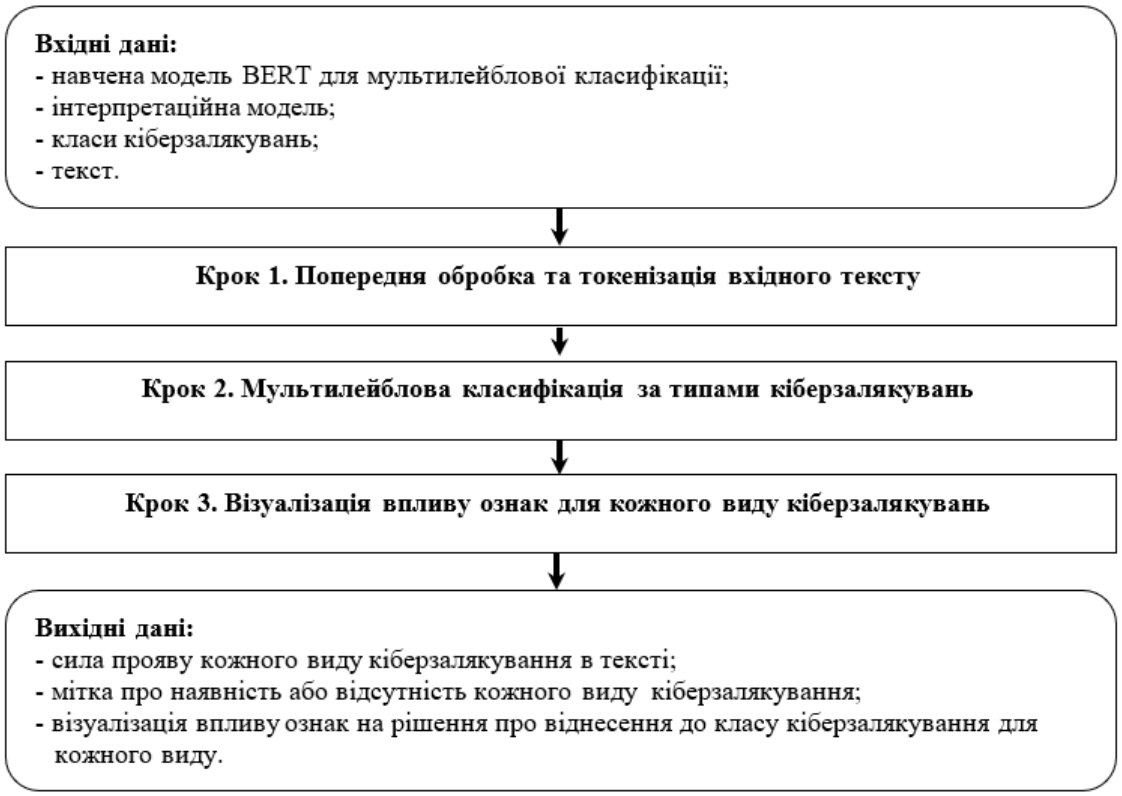

Рис. 4. Кроки етапу візуальної інтерпретації результатів виявлення кіберзалякувань

Отже, наведений етап візуальної інтерпретації результатів нейромережевого виявлення кіберзалякувань сприятиме зрозумінню та поясненню рішень, ухвалених моделлю для мультилейблової класифікації текстового контенту щодо виявлених типів кіберзалякувань.

ЕКСПЕРЕМЕНТИ ТА РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Першогочергово необхідно виконати оцінювання та коригування репрезентативності датасету для виявлення кіберзалякувань. Тому для навчання моделей машинного навчання, призначених для розмітки вхідного датасету, були використані набори даних, що ілюструють два етичні аспекти принципу справедливості: гендер [11] (34146 унікальних текстових записів), вік [12] (20109 унікальних текстових записів). Оскільки класи в цих датасетах були нерівномірно представлені та містили різну кількість зразків, що могло б негативно вплинути на якість навчання моделей, усі класи були збалансовані за чисельністю.

Для оцінювання та коригування репрезентативності датасету для виявлення кіберзалякувань із цільовими пропорціями класів за віком і статтю використано демографічні дані населення України. Згідно з оцінками Інституту демографії та соціальних досліджень імені М. В. Птухи НАН України (https://idss.org.ua/forecasts/nation_pop_proj), станом на липень 2023 року чисельність населення України становила 35 596 216 осіб.

У процесі формування репрезентативного датасету з урахуванням вікового та гендерного етичних аспектів на основі демографічних підгруп населення України було створено репрезентативну вибірку текстових даних шляхом аугментації. Баланс класів цієї вибірки наведено в таблиці 1.

Відхилення розподілів зразків за класами вікового та гендерного етичних аспектів у датасеті, скоригованому за запропонованим підходом, від ідеального репрезентативного розподілу становили: мінімальне – 0.00%, максимальне – 0.04%, середнє – 0.02%.

Далі сформований датасет використовується на етапі нейромережевого виявлення і класифікації кіберзалякувань у текстовому контенті. Згідно крокам етапу спочатку для бінарної класифікації відбувається виявлення загальної оцінки кіберзалякувань у текстовому контенті використовується навчена на датасеті, що є результуючим етапу 1 нейромережева модель. Для бінарної класифікації використано

нейромережу BiLSTM, що отримала такі показники ефективності: Accuracy 0.96, Precision 0.91, Recall 0.93, F1 Score 0.92.

Таблиця 1

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%

Наступним кроком проведено мультилейблову класифікацію текстового зразка. В якості нейромережевої моделі для завдання мультилейблової класифікації обрано BERT [14], яка отримала наступні показники ефективності: Accuracy 0.94, Precision 0.93, Recall 0.93, F1 Score 0.93. Нейромережу було навчено для виявлення таких типів кіберзалякувань, як вікове, гендерне, релігійне, етнічне, а також окремий узагальнений тип – інші кіберзалякування.

Для дослідження запропонованого підходу 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 такі ймовірності наявності різних видів кіберзалякувань у цьому текстовому зразку:

– вікове: 0.06%
– етнічне: 0.08%;
– гендерне: 0.10%;
– релігійне: 99.86%;
– інший тип: 0.09%.

Надані результати нейромережевою моделлю BERT для мультилейблової класифікації типів кіберзалякувань використовуються на третьому етапі візуальної інтерпретації [14] результатів нейромережевого виявлення кіберзалякувань. Застосування моделі LIME для інтерпретації результатів моделі BERT у задачі мультилейблової класифікації типів кіберзалякувань у текстовому зразку дозволило отримати візуальне представлення виявлених типів кіберзалякувань, базуючись на абсолютних значеннях ваг, що наведені на рисунку 5. Для пояснення рішень, прийнятих моделлю BERT, слова було виділено кольорами: найбільш насичений колір вказує на найвище значення ваги слова, тобто це слово мало найбільший вплив, тоді як найсвітліший колір відображає найменший вплив.

Рис. 5. Інтерпретація результатів виявлення різних типів кіберзалякувань

У випадку з LIME важливо не лише продемонструвати силу впливу слова, але й вказати, чи є цей вплив позитивним (збільшує ймовірність) або негативним (зменшує ймовірність). Тому був реалізований додатковий підхід до коригування яскравості, щоб від'ємні значення мали меншу яскравість і використовували інший колірний відтінок для негативних та позитивних значень. Результат такої візуальної інтерпретації показано на рисунку 6.

Рис. 6. Інтерпретація результатів виявлення різних типів кіберзалякувань з урахуванням негативного чи позитивного типу впливу на результат

Таким чином, запропоновані візуальні інтерпретації результатів виявлення кіберзалякувань у текстовому контенті дозволяють оцінити, чи модель використовує відповідні ознаки для ухвалення рішень, або ж її поведінка може бути спричинена випадковими чи нерелевантними факторами [16, 17]. Наприклад, якщо в тексті зустрічаються слова, що не мають значення для вікового кіберзалякування, але при цьому мають високий вплив, це може вказувати на наявність помилки чи упередженості в моделі.

ВИСНОВКИ З ДАНОГО ДОСЛІДЖЕННЯ І ПЕРСПЕКТИВИ ПОДАЛЬШИХ РОЗВІДОК У ДАНОМУ НАПРЯМІ

Розроблено підхід до виявлення та класифікації кіберзалякувань у цифрових текстах засобами штучного інтелекту. Наведений підхід складається з трьох етапів, що дозволяють виявляти кіберзалякування у тексті з врахуванням значущості та неупередженості щодо різних груп людей, як за віковими, так і за етнічними чи віковими ознаками, а також надає пояснення рішень моделі щодо визначених у текстовому контенті типів кіберзалякувань.

Етап оцінювання та коригування репрезентативності датасету для виявлення кіберзалякувань дозволяє оцінювати репрезентативність вибірок з точки зору етичних критеріїв та здійснювати відповідне коригування датасету, забезпечуючи його відповідність етичним вимогам. Отримані відхилення розподілів зразків за класами етичних аспектів датасету, трансформованого за кроками етапу, від ідеального репрезентативного розподілу становили: мінімальне 0.00%, максимальне 0.04%, середнє 0.02%. Результати дослідження підтверджують, що виконання кроків етапу дозволяє проводити аналіз репрезентативності текстових датасетів та адаптувати їх до репрезентативного.

Етап нейромережевого виявлення і класифікації кіберзалякувань у текстовому контенті дозволяє аналізувати текстове повідомлення та визначати загальний рівень прояву кіберзалякувань у ньому, а також виконувати мультилейблову класифікацію, надаючи окремі показники для різних типів кіберзалякувань, таких як вікові, релігійні, етнічні, гендерні залякування та інші. У завданні бінарної класифікації модель BiLSTM, досягнувши таких показників: Accuracy – 0,96, Precision – 0,96, Recall – 0,959, F1 – 0,957. А у завданні мультилейблової класифікації модель BERT показала наступні результати ефективності: Accuracy – 0,94, Precision – 0,93, Recall – 0,93, F1 Score – 0,93.

Етап візуальної інтерпретації результатів нейромережевого виявлення кіберзалякувань дає змогу здійснювати інтерпретацію результатів для кожного виявленого типу кіберзалякування окремо, що досягається завдяки використанню мультилейблового класифікатора на основі нейромережевої архітектури трансформер та інтерпретаційної моделі машинного навчання. Запропонована інтерпретація результатів виявлення кіберзалякувань у текстовому контенті відноситься до категорії інструментів візуальної аналітики рішень штучного інтелекту, розробка яких є необхідною для забезпечення етичності, прозорості та довіри до таких систем у суспільстві, особливо у контексті чутливих тем, як виявлення кіберзалякувань.

Отже, запропонований підхід для виявлення та класифікації кіберзалякувань у цифрових текстах є комплексним і враховує різноманітні соціальні групи, зокрема за віковими та етнічними ознаками. Він дозволяє ефективно виявляти різні типи кіберзалякувань у текстовому контенті та забезпечує пояснення рішень моделі через візуальну інтерпретацію, що підвищує прозорість та довіру до результатів.

Література

1. Молчанова М.О. Метод нейромережевого виявлення кібербулінгу з використанням хмарних

сервісів та об'єктно-орієнтованої моделі / М.О. Молчанова, О.В. Мазурець, О.В. Собко, В.І. Кліменко, В.І. Андрощук // Вісник Хмельницького національного університету. Серія: Технічні науки. – 2024. – № 2 (333). – С. 200–206.

2. Собко О.В. Метод інтелектуального виявлення та класифікації кіберзалякувань у текстовому контенті / О.В. Собко // Інформаційні управляючі системи та технології ІУСТ-ОДЕСА-2024: матеріали XII Міжнар. наук.-практ. конф.– Одеса, 2024. – С. 262–265.

3. Krak I. 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. – С. 16–28.

4. Abood M.M. Explainable Multimodal Deep Learning Model for Cyberbullying Detection (EMDL-CBD) / M.M. Abood, M.A. Al-Bayati // Journal Port Science Research. – 2024. – Vol. 7, № 1.

5. Sheanoda V. Sexuality, gender and culturally diverse interpretations of cyberbullying / V. Sheanoda, K. Bussey, T. Jones // New Media & Society. – 2024. – Vol. 26, № 1. – С. 154–171.

6. Sen M. From Tweets to Insights: BERT-Enhanced Models for Cyberbullying Detection / M. Sen, J. Masih, R. Rajasekaran // 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS): Proc. – 2024. – С. 1289–1293.

7. Nuthalapati P. Cyberbullying Detection: A Comparative Study of Classification Algorithms [Електронний ресурс] – Режим доступу: https://www.authorea.com/doi/full/10.22541/au.170664263.38254624.

8. Khairy M. The effect of rebalancing techniques on the classification performance in cyberbullying datasets / M. Khairy, T.M. Mahmoud, T. Abd-El-Hafeez // Neural Computing and Applications. – 2023. – Vol. 36, № 3. – С. 1049–1065

9. Perera A. Cyberbullying Detection System on Social Media Using Supervised Machine Learning / A. Perera, P. Fernando // Procedia Computer Science. – 2024. – Vol. 239. – С. 506–516.

10. Aggarwal P. Shielding Social Media: BERT and SVM Unite for Cyberbullying Detection and Classification / P. Aggarwal, R. Mahajan // Journal of Information Systems and Informatics. – 2024. – Vol. 6, № 2. – С. 607–623.

11. Molchanova M. Object-Oriented Approach for Ethnic Enmity Detection in Text Messages by NLP / M. Molchanova, O. Mazurets, O. Sobko, I. Boiarchuk // Scientific Achievements and Innovations as a Way to Success: Proc. XXI Int. Scientific and Practical Conf., May 1–3, 2024, Vilnius, Lithuania. – Vilnius, 2024. – С. 73–77.

12. B. Memarian B. Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review / B. Memarian, T. Doleck // Computers and Education: Artificial Intelligence. – 2023. – Vol. 5.

13. Мазурець О.В. Метод автоматизованого підбору відповідей на користувацькі запитання за семантичною подібністю / О.В. Мазурець, О.В. Козенко, О.В. Собко // Глушковські читання: матеріали XII Всеукр. наук.-практ. конф., Київ, 2023. – Київ, 2023. – С. 106–109.

14. Alissa S. Text Simplification Using Transformer and BERT / S. Alissa, M. Wald // Computers, Materials & Continua. – 2023. – Vol. 75, № 2. – С. 3479–3495.

15. Kovalchuk O. Visual Analytics-Based Method for Sentiment Analysis of COVID-19 Ukrainian Tweets / O. Kovalchuk, V. Slobodzian, O. Sobko, M. Molchanova, O. Mazurets, O. Barmak, I. Krak, N. Savina // Lecture Notes on Data Engineering and Communications Technologies. – 2023. – Vol. 149. – С. 591–607.

16. Собко О.В. Дослідження ефективності методу оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості / О.В. Собко // Перспективи сучасної науки: теорія і практика: матеріали VIII Міжнар. наук.-практ. конф.– Львів, 2024. – С. 217–221.

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

References

1. Molchanova M.O. Metod neiromerezhevoho vyiavlennia kiberbulinhu z vykorystanniam khmarnykh servisiv ta obiektno-oriientovanoi modeli / M.O. Molchanova, O.V. Mazurets, O.V. Sobko, V.I. Klimenko, V.I. Androshchuk // Visnyk Khmelnytskoho natsionalnoho universytetu. Seriia: Tekhnichni nauky. – 2024. – № 2 (333). – S. 200–206.

2. Sobko O.V. Metod intelektualnoho vyiavlennia ta klasyfikatsii kiberzaliakuvan u tekstovomu kontenti / O.V. Sobko // Informatsiini upravliaiuchi systemy ta tekhnolohii IUST-ODESA-2024: materialy XII Mizhnar. nauk.-prakt. konf.– Odesa, 2024. – S. 262–265.

3. Krak I. 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. – S. 16–28.

4. Abood M.M. Explainable Multimodal Deep Learning Model for Cyberbullying Detection (EMDL-CBD) / M.M. Abood, M.A. Al-Bayati // Journal Port Science Research. – 2024. – Vol. 7, № 1.

5. Sheanoda V. Sexuality, gender and culturally diverse interpretations of cyberbullying / V. Sheanoda, K. Bussey, T. Jones // New Media & Society. – 2024. – Vol. 26, № 1. – S. 154–171.

6. Sen M. From Tweets to Insights: BERT-Enhanced Models for Cyberbullying Detection / M. Sen, J. Masih, R. Rajasekaran // 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS): Proc. – 2024. – S. 1289–1293.

7. Nuthalapati P. Cyberbullying Detection: A Comparative Study of Classification Algorithms [Elektronnyi resurs] – Rezhym dostupu: https://www.authorea.com/doi/full/10.22541/au.170664263.38254624.

8.Khairy M. The effect of rebalancing techniques on the classification performance in cyberbullying datasets / M. Khairy, T.M. Mahmoud, T. Abd-El-Hafeez // Neural Computing and Applications. – 2023. – Vol. 36, № 3. – S. 1049–1065

9.Perera A. Cyberbullying Detection System on Social Media Using Supervised Machine Learning / A. Perera, P. Fernando // Procedia Computer Science. – 2024. – Vol. 239. – S. 506–516.

10. Aggarwal P. Shielding Social Media: BERT and SVM Unite for Cyberbullying Detection and Classification / P. Aggarwal, R. Mahajan // Journal of Information Systems and Informatics. – 2024. – Vol. 6, № 2. – S. 607–623.

11. Molchanova M. Object-Oriented Approach for Ethnic Enmity Detection in Text Messages by NLP / M. Molchanova, O. Mazurets, O. Sobko, I. Boiarchuk // Scientific Achievements and Innovations as a Way to Success: Proc. XXI Int. Scientific and Practical Conf., May 1–3, 2024, Vilnius, Lithuania. – Vilnius, 2024. – S. 73–77.

12. B. Memarian B. Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review / B. Memarian, T. Doleck // Computers and Education: Artificial Intelligence. – 2023. – Vol. 5.

13. Mazurets O.V. Metod avtomatyzovanoho pidboru vidpovidei na korystuvatski zapytannia za semantychnoiu podibnistiu / O.V. Mazurets, O.V. Kozenko, O.V. Sobko // Hlushkovski chytannia: materialy XII Vseukr. nauk.-prakt. konf., Kyiv, 2023. – Kyiv, 2023. – S. 106–109.

14. Alissa S. Text Simplification Using Transformer and BERT / S. Alissa, M. Wald // Computers, Materials & Continua. – 2023. – Vol. 75, № 2. – S. 3479–3495.

15. Kovalchuk O. Visual Analytics-Based Method for Sentiment Analysis of COVID-19 Ukrainian Tweets / O. Kovalchuk, V. Slobodzian, O. Sobko, M. Molchanova, O. Mazurets, O. Barmak, I. Krak, N. Savina // Lecture Notes on Data Engineering and Communications Technologies. – 2023. – Vol. 149. – S. 591–607.

16. Sobko O.V. Doslidzhennia efektyvnosti metodu otsiniuvannia ta koryhuvannia reprezentatyvnosti datasetu za FATE-pryntsypom spravedlyvosti / O.V. Sobko // Perspektyvy suchasnoi nauky: teoriia i praktyka: materialy VIII Mizhnar. nauk.-prakt. konf.– Lviv, 2024. – S. 217–221.

17. Sobko O.V. Metod intelektualnoho vyiavlennia kiberzaliakuvan u tekstovomu kontenti / O.V. Sobko // Rozvytky informatsiino-keruiuchykh system ta tekhnolohii: monohrafiia. – Lviv-Torun: Lina-Pres, 2024. – S. 267–287.