УДК 004.8

**Собко О.В.**, *(Хмельницький національний університет, м. Хмельницький, Україна)*

# ВІЗУАЛЬНА ІНТЕРПРЕТАЦІЯ НЕЙРОМЕРЕЖЕВОГО ВИЯВЛЕННЯ КІБЕРБУЛІНГУ У ЦИФРОВИХ ТЕКСТАХ

***Анотація:*** *Запропоновано метод візуальної інтерпретації нейромережевого виявлення кібербулінгу в цифрових текстах, що дозволяє інтерпретувати рішення моделі щодо типів кібербулінгу. Метод базується на використанні моделі BERT для мультилейблової класифікації та інтерпретаційної моделі LIME, яка візуалізує вплив слів на рішення моделі. Метод забезпечує три формати інтерпретації: кольорову палітру, діаграми локальної та загальної важливості слів. Експерименти підтвердили, що розроблений підхід забезпечує зрозуміле пояснення рішень штучного інтелекту щодо виявлених типів кібербулінгу..*

***Ключові слова:*** *кібербулінг, інтерпретація результатів, нейронні мережі, BERT, LIME.*

***Abstract:*** *A method for explaining the results of neural network detection of cyberbullying in digital texts is proposed, which allows interpreting the model's decisions regarding the types of cyberbullying. The method is based on the use of the BERT model for multi-label classification and the LIME interpretation model, which visualizes the influence of words on the model's decisions. The method provides three interpretation formats: a color palette, diagrams of local and global word importance. Experiments have confirmed that the developed approach provides a clear explanation of artificial intelligence decisions regarding the detected types of cyberbullying..*

***Keywords:*** *propaganda objects, propaganda techniques, propaganda detection, natural language processing*

## Постановка проблеми

Проблема кібербулінгу стає дедалі актуальнішою через зростання кількості користувачів соціальних мереж, особливо серед молоді, що збільшує попит на системи нейромережевого виявлення кібербулінгу в цифрових текстах [1,2]. Завдяки прогресу у використанні моделей трансформерів, зокрема BERT, стало можливим ефективно виявляти та класифікувати типи кібербулінгу [3]. Однак складність інтерпретації таких моделей викликає сумніви щодо їх використання у чутливих контекстах. Тому інтерпретація рішень є ключовою для забезпечення довіри та прозорості. У роботі запропоновано метод пояснення рішень моделі щодо виявлених типів кібербулінгу, таких як дискримінація за віком, етнічністю чи гендером.

## Аналіз останніх публікацій

Проблема нейромережевого виявлення кібербулінгу є надзвичайно актуальною через його руйнівний вплив на психічне здоров’я, особливо підлітків та молоді. Сучасні методи ґрунтуються на технологіях обробки природної мови, що дозволяють аналізувати цифрові тексти для виявлення та класифікації різних форм кібербулінгу [4].

У дослідженні [5] розглядається задача нейромережевого виявлення кібербулінгу. Серед протестованих моделей, таких як Random Forest, XgBoost, Naive Bayes, SVM, CNN, RNN та BERT, остання продемонструвала найвищу ефективність, досягнувши 88,8% точності у бінарній класифікації та 86,6% у мультилейбловій.

Автори роботи [6] розробили новий підхід до виявлення кібербулінгу, протестувавши SVM, Naive Bayes і Logistic Regression у поєднанні з різними методами обробки тексту. Було доведено, що аналіз настроїв, N-грам, TF-IDF та визначення ненормативної лексики суттєво покращують точність, дозволяючи досягти 75,17% у задачі класифікації.

Інші автори зосередили увагу на інтерпретації результатів. Наприклад, у [7] представлено модель BiLSTM-LIME для багатокласової класифікації кібербулінгу в цифрових текстах Twitter. Використання LIME забезпечило високу якість пояснень, акцентуючи увагу на токенах, які вплинули на рішення.

Дослідження [8] запропонувало ансамбль BERT та SVM з налаштуванням параметрів для багатокласової класифікації кібербулінгу у соціальних медіа. Модель показала точність 90% на тестових даних, перевершивши альтернативні підходи. Для пояснення прогнозів було використано техніку SHAP, яка надала детальний аналіз значущості ознак.

Аналіз публікацій свідчить, що виявлення кібербулінгу в цифрових текстах є важливим і багатогранним завданням, яке активно досліджується завдяки значному впливу цієї проблеми на суспільство. Використання сучасних моделей обробки природної мови, особливо архітектур трансформерів, таких як BERT, демонструє високу ефективність у задачах як бінарної, так і мультилейблової класифікації кібербулінгу. Зокрема, BERT стабільно перевершує інші підходи, досягаючи точності понад 85% у більшості експериментів. Що стосується інтерпретації, роботи підкреслюють важливість забезпечення прозорості моделей. Методи, такі як LIME та SHAP, дозволяють не лише пояснити рішення моделі, але й зробити її застосування більш зрозумілим для користувачів. Це особливо важливо у соціально значущих контекстах, де прозорість рішень безпосередньо впливає на довіру до систем штучного інтелекту.

## Мета роботи та постановка завдань

Мета роботи полягає в розробці методу для візуальної інтерпретації результатів нейромережевого виявлення кібербулінгу в цифрових текстах, спрямованого на пояснення рішень моделі штучного інтелекту стосовно визначених типів кібербулінгу. Запропонований метод повинен забезпечувати зрозумілу інтерпретацію, яка дозволяє людині аналізувати текстові ознаки, що вплинули на рішення нейромережевої моделі щодо ідентифікації типів кібербулінгу.

## Виклад основного матеріалу

Метод інтерпретації результатів нейромережевого виявлення кібербулінгу в цифрових текстах передбачає створення візуального пояснення рішень моделі штучного інтелекту щодо визначених типів кібербулінгу [9]. Схематичне представлення цього методу наведено на рисунку 1.

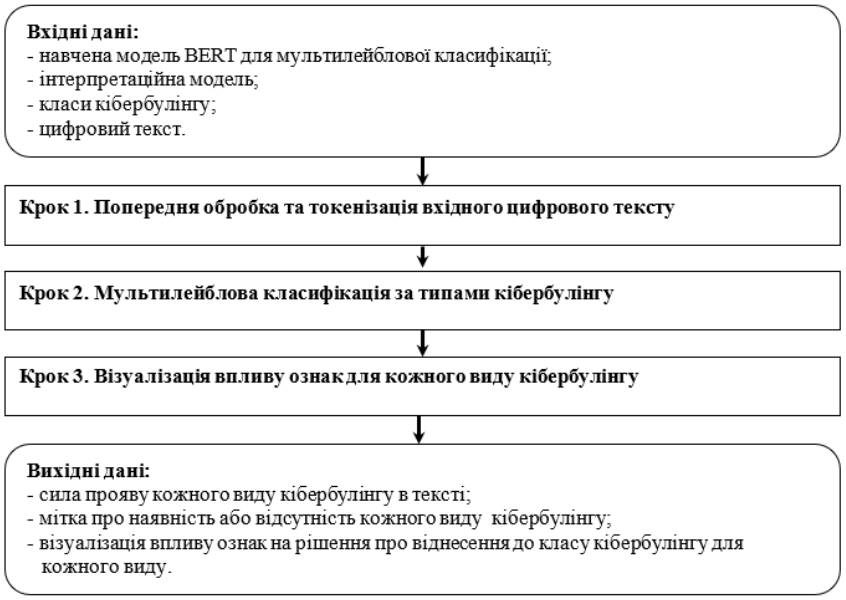


Рисунок 1. Схема методу візуальної інтерпретації нейромережевого виявлення кібербулінгу в цифрових текстах

Вхідними даними цього етапу є навчена модель трансформерної архітектури для мультилейблової класифікації, здатна визначати різні типи кібербулінгу, такі як віковий, етнічний, гендерний, релігійний та узагальнений тип, що охоплює інші види кібербулінгу. Також використовуються інтерпретаційні моделі, які пояснюють вплив окремих слів чи фраз

на результати класифікації. Вхідний текст аналізується на наявність ознак кібербулінгу, після чого результати піддаються інтерпретації.

Першим кроком є токенізація тексту, після чого текстові елементи перетворюються в числові послідовності для подальшої обробки нейромережевою моделлю.

Другий крок полягає у прогнозуванні ймовірностей належності тексту до кожного з типів кібербулінгу, оцінюючи наявність ознак, таких як вікові, етнічні чи гендерні характеристики.

Третім етапом є пояснення та візуалізація результатів класифікації за допомогою інтерпретаційної моделі, яка виявляє вплив окремих слів або фраз на ідентифікацію ознак кібербулінгу. Для мультилейблової класифікації часто застосовуються такі інтерпретаційні методи, як [10]: Local Interpretable Model-agnostic Explanations (LIME), який генерує локальні пояснення для кожного передбачення, демонструючи, які слова найбільше вплинули на результат; SHapley Additive exPlanations (SHAP), що базується на теорії ігор і обчислює внесок кожного слова у передбачення, враховуючи взаємодію між ознаками; Transformers Interpret, інтерпретаційна бібліотека, розроблена спеціально для моделей на основі трансформерів, таких як BERT, GPT, RoBERTa та інші моделі з бібліотеки Hugging Face; методи, що використовують Attention, які аналізують ваги уваги трансформерів (наприклад, у моделі BERT) для розуміння важливості окремих слів чи фраз у процесі прийняття рішень моделлю.

Вихідними даними є інтенсивність прояву кожного типу кібербулінгу в тексті, виражена через ймовірності, які демонструють ступінь наявності ознак для кожного класу кібербулінгу. Для кожного класу визначається мітка, що вказує на наявність або відсутність ознак, представлених числовими значеннями, які відображають ймовірність прояву кібербулінгу за кожним типом. Крім того, метод забезпечує візуалізацію впливу конкретних ознак на прийняте рішення про належність тексту до певного класу кібербулінгу, де важливі слова підсвічуються відповідно до їх значущості для кожного з класів.

Таким чином наведений метод візуальної інтерпретації результатів нейромережевого виявлення кібербулінгу сприятиме кращому розумінню та поясненню рішень, ухвалених моделлю щодо мультилейблової класифікації цифрових текстів та визначених типів кібербулінгу.

Для навчання моделі BERT [11], яка застосовується на кроці 2 методу візуальної інтерпретації нейромережевого виявлення кібербулінгу (рисунок 1), використовувався датасет «Cyberbullying Classification» [12]. Цей датасет містить текстові повідомлення з мітками, що визначають належність кожного повідомлення до одного з класів: Age, Ethnicity, Gender, Religion, Other type of cyberbullying, Not cyberbullying.

Для навчання моделі BERT мультилейбловій класифікації був видалений клас «Not cyberbullying» з датасету «Cyberbullying Classification», оскільки він не використовувався в навчанні. Крім того, клас «Other type of cyberbullying» був збільшений за допомогою методики SMOTE-балансування, що дозволило створити синтетичні зразки. Завдяки цьому попередньому етапу обробки даних був отриманий збалансований набір для навчання моделі BERT для завдання мультилейблової класифікації типів кібербулінгу в текстовому контенті.

Для оцінки ефективності методу візуальної інтерпретації нейромережевого виявлення кібербулінгу в цифрових текстах використовувалося середовище Google Colab. Модель BERT була навчена для класифікації таких типів кібербулінгу, як віковий, гендерний, релігійний, етнічний, а також окремо для типу «інші кібербулінги».

Показники макрометрик навченої моделі BERT для мультилейблової класифікації типів кібербулінгу становлять: Accuracy 0.956478, Precision 0.963677, Recall 0.956478 та F1 Score 0.960019. Ці значення свідчать про високу ефективність моделі у виявленні різних видів кібербулінгу в текстовому контенті.

Для дослідження був використаний англомовний цифровий текст, який було проаналізовано для виявлення різних типів кібербулінгу за допомогою навченої моделі BERT. Модель BERT виявила ймовірності наявності різних видів кібербулінгу в цифровому тексті,

зокрема віковий кібербулінг – 0.06%, етнічний – 0.08%, гендерний – 0.10%, інший тип – 0.09%, та релігійний кібербулінг – 99.86%

Застосування моделі LIME для візуальної інтерпретація нейромережевого виявлення кібербулінгу за допомогою моделі BERT для мультилейблової класифікації типів кібербулінгу в цифровому тексті дозволило отримати візуальні результати інтерпретації виявлених типів кібербулінгу, використовуючи абсолютні значення ваг, що зображені на рисунку 2. Для пояснення прийнятих рішень моделлю BERT слова в цифровому тексті виділяються різними кольорами: найбільш яскравий колір вказує на найбільшу вагу слова, що означає його найбільший вплив на результат, а найсвітліший – на найменший.


Рисунок 2. Абсолютне значення ваги для визначення яскравості кольору з метою інтерпретації результатів виявлення різних типів кібербулінгу в цифровому тексті

Як видно з рисунку 2, слова з додатними та від'ємними значеннями виділяються однаковою яскравістю. В цьому випадку для визначення яскравості використовується абсолютне значення ваги, що призводить до однакової яскравості для від'ємних та додатних значень. Від'ємні значення ваги зменшують ймовірність певного класу, тоді як додатні значення збільшують її, але обидва типи мають однаковий вплив на прийняте моделлю рішення. Для LIME важливо не лише показати силу впливу слова, а й його напрямок (позитивний або негативний). Тому реалізовано підхід, де від'ємні значення мають менш яскравий колір і окремий відтінок для додатних та від'ємних значень. Результати такої візуалізації подано на рисунку 3. Використання різних кольорів для додатних і від'ємних значень є важливим, оскільки від'ємні ваги зменшують ймовірність певного класу, а додатні – збільшують. Без цієї відмінності, однакові інтенсивності різних знаків можуть бути сприйняті як рівнозначні, що може призвести до неправильного розуміння результатів.


Рисунок 3. Підхід для інтерпретації результатів виявлення типів кібербулінгу з урахуванням негативного чи позитивного типу впливу кінцевий результат

Додатково були створені діаграми для графічної інтерпретації впливу окремих слів цифрового тексту на ймовірність віднесення цього тексту до конкретного типу кібербулінгу (рисунок 4).

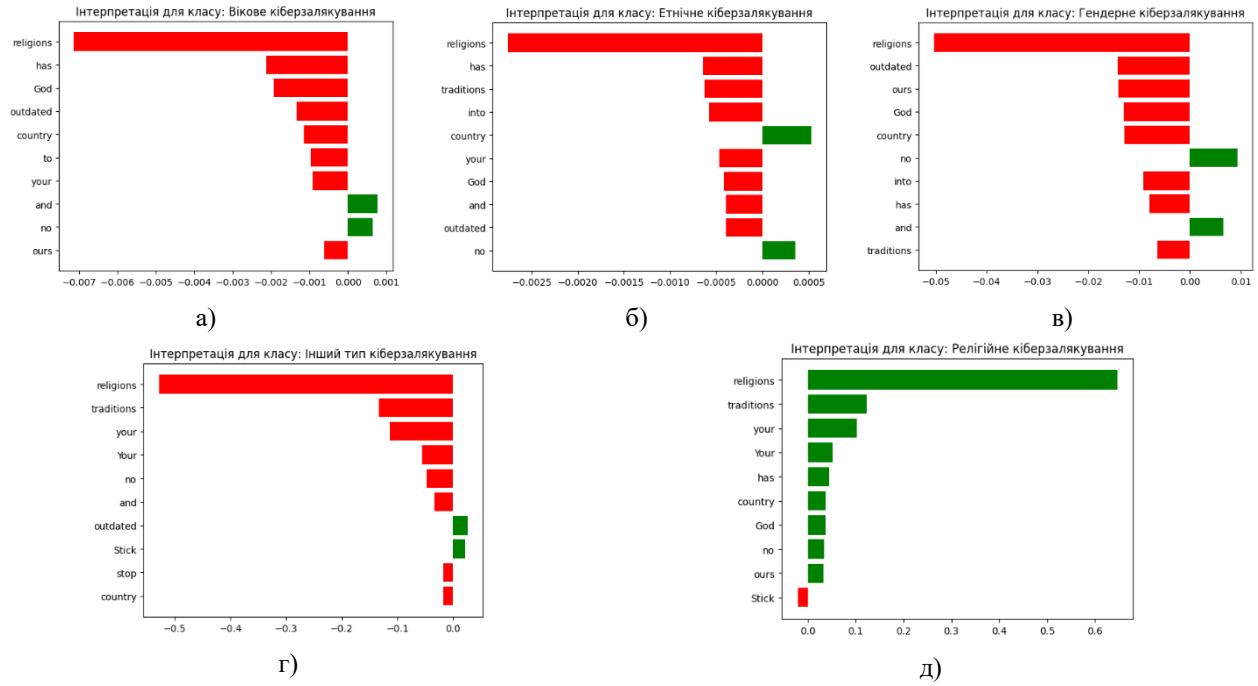


Рисунок 4. Графічна інтерпретації впливу окремих слів цифрового тексту на результат

Діаграми показують, як модель оцінює вагу кожного слова в цифровому тексті, залежно від його внеску в прийняте рішення. Вплив слів відображено у вигляді горизонтальних стовпців, довжина яких показує величину впливу (ваги), а колір – напрямок цього впливу. Червоні стовпці вказують на негативний вплив слів, що зменшують ймовірність віднесення тексту до певного класу, тоді як зелені стовпці означають позитивний вплив, який збільшує ймовірність цього віднесення. Величина впливу вимірюється числовим значенням, що відображається на горизонтальній осі графіка.

Також було обчислено середнє значення важливості кожного слова для всіх класів, що дозволяє оцінити загальний вплив кожного слова без прив'язки до конкретного типу кібербулінгу. Результати обчислень представлені у вигляді діаграми (рисунок 5).

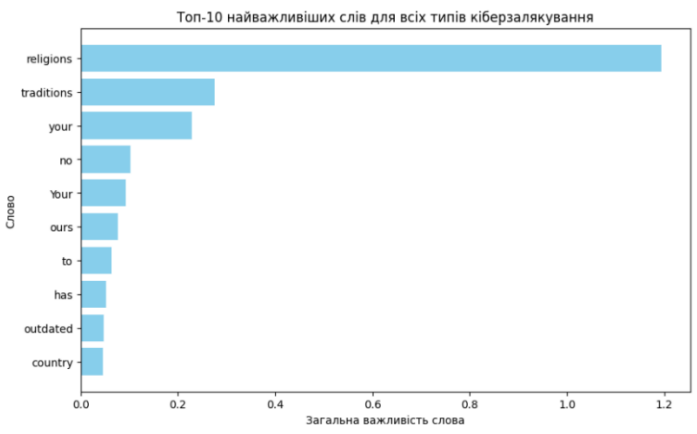


Рисунок 5. Середні значення важливості топ-10 слів для всіх класів

Обчислення загального впливу слів на результати моделі для всіх типів кібербулінгу є важливим для розуміння роботи моделі та її рішень. Аналіз здійснюється через агрегацію ваг слів, які модель оцінює для кожного класу. Використовується модуль ваги, що означає абсолютну величину впливу слова без урахування його позитивного чи негативного значення. Цей підхід дає можливість виявити слова, які модель вважає важливими незалежно від конкретного типу кібербулінгу. Наприклад, слова, що стосуються різних типів кібербулінгу, можуть мати високі ваги для кількох класів. Якщо слово має високий загальний вплив, це може свідчити про його універсальну роль у контексті кібербулінгу. Наприклад, слова, що вказують на етнічну приналежність або релігію, можуть мати великий вплив на кілька класів, таких як «етнічний кібербулінг» і «релігійний кібербулінг», що вказує на потенційну крос-модальність ознак, які модель використовує для прийняття рішень. Якщо ж слово має вплив лише на один клас, це підкреслює його специфічність і може вказувати на унікальні мовні патерни для цього виду кібербулінгу.

Отже, візуальні інтерпретації результатів нейромережевого виявлення кібербулінгу в цифрових текстах дозволяють оцінити, чи модель використовує релевантні ознаки для ухвалення рішень, чи її поведінка зумовлена випадковими чи нерелевантними факторами. Наприклад, якщо в тексті з'являються слова, що не мають змістового зв’язку з віковим кібербулінгом, але мають значний вплив, це може свідчити про наявність помилки або упередження в моделі.

## ВИСНОВКИ

У роботі запропоновано метод візуальної інтерпретації нейромережевого виявлення кібербулінгу в цифрових текстах, призначений для пояснення рішень нейромережевої моделі щодо типів кібербулінгу, виявлених у текстах. Метод є оригінальним, оскільки здійснює інтерпретацію результатів для кожного типу кібербулінгу окремо, що досягається використанням мультилейблового класифікатора нейромережевої архітектури трансформер і інтерпретаційної моделі машинного навчання.

Завдяки використанню навченої нейромережевої моделі BERT для мультилейблової класифікації типів кібербулінгу в цифровому тексті, модель виявляє різні типи кібербулінгу з указанням відсотка наявності кожного з них. Згідно з розробленим методом, для візуальної інтерпретації результатів виявлення кібербулінгу використано підхід, що базується на моделі машинного навчання LIME для локальної інтерпретованості, що дозволяє візуалізувати вплив кожного окремого слова на рішення моделі щодо належності тексту до певних типів кібербулінгу.

Метод забезпечує три способи візуальної інтерпретації нейромережевого виявлення кібербулінгу в цифрових текстах: за кольоровою палітрою, за діаграмами локальної важливості слів і за діаграмами загальної важливості слів. Інтерпретація результатів за кольоровою палітрою ґрунтується на використанні абсолютного значення ваги для визначення яскравості кольору, де найбільш яскравий колір вказує на найбільший вплив слова на прийняте рішення моделі, а найменш яскравий – на найменший вплив, незалежно від того, чи був він позитивним чи негативним. Проте, для повної інтерпретації необхідно також розуміти напрямок впливу, оскільки від'ємні ваги зменшують ймовірність певного класу, а додатні – збільшують її. Тому реалізовано інтерпретацію рішень моделі BERT з урахуванням напрямку впливу.

Візуальна інтерпретація результатів за діаграмами локальної важливості слів демонструє, як кожне слово впливає на ймовірність віднесення тексту до конкретного типу кібербулінгу, дозволяючи побачити, як модель оцінює вагу кожного слова, залежно від його впливу на прийняте рішення. Інтерпретація результатів за діаграмами загальної важливості слів показує 10 слів, які модель вважає важливими для визначення типу кібербулінгу, незалежно від конкретного класу.

Результати експериментів свідчать, що запропонований метод забезпечує візуальну інтерпретацію рішень щодо нейромережевого виявлення кібербулінгу на рівні, який дозволяє людині зрозуміти, які ознаки тексту вплинули на прийняття рішень штучним інтелектом. Розроблений метод інтерпретації виявлення кібербулінгу у цифрових текстах належить до категорії засобів візуальної аналітики рішень штучного інтелекту, що є необхідним для забезпечення етичності, прозорості та довіри до таких систем штучного інтелекту в

суспільстві, особливо коли йдеться про чутливі питання, як кібербулінг. Дослідження підкреслює важливість не тільки точності моделей, але й їхньої пояснюваності, що є ключовим для побудови довіри до систем штучного інтелекту.

## Список посилань.


1. Собко О.*В. Виявлення та класифікація кіберзалякувань у цифрових текстах засобами штучного інтелекту / О.В. Собко // Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах». – 2024. – № 4. – С. 143–152.*
2. *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.*
3. *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.*
4. *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.*
5. *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.*
6. *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, № 3.*
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 (дата звернення: 17.17.2024).*
8. *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.*
9. *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.*
10. *Kiefer S. CaSE: Explaining Text Classifications by Fusion of Local Surrogate Explanation Models with Contextual and Semantic Knowledge / S. Kiefer // Information Fusion. – 2022. – Vol. 77. – С. 184–195.*
11. *Alissa S. Text Simplification Using Transformer and BERT / S. Alissa, M. Wald // Computers, Materials & Continua. – 2023. – Vol. 75, № 2. – С. 3479–3495.*
12. *Cyberbullying Classification Dataset [Електронний ресурс]. – Kaggle. – Режим доступу: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification (дата звернення: 17.17.2024).*