DOI 10.31891/2307-5732-2024-343-6-45
УДК 004.8

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

МЕТОД ІНТЕРПРЕТАЦІЇ РЕЗУЛЬТАТІВ ВИЯВЛЕННЯ КІБЕРЗАЛЯКУВАНЬ У ТЕКСТОВОМУ КОНТЕНТІ ЗАСОБАМИ ШТУЧНОГО ІНТЕЛЕКТУ

Запропоновано метод інтерпретації результатів виявлення кіберзалякувань у текстовому контенті засобами штучного інтелекту, що призначений для пояснення рішень нейромережевої моделі щодо визначених у текстовому контенті типів кіберзалякувань. Метод оригінальний тим, що здійснює інтерпретацію результатів для кожного виявленого типу кіберзалякування окремо, що досягається шляхом використання мультилейблового класифікатора нейромережевої архітектури трансформер та інтерпретаційної моделі машинного навчання. Шляхом використання навченої нейромережевої моделі BERT для мультилейблової класифікації типів кіберзалякувань у вхідному текстовому зразку виявляються різні типи кіберзалякувань із вказанням відсотку наявності кожного з них. Згідно розробленого методу, для візуальної інтерпретації результатів виявлення кіберзалякувань використано підхід, який ґрунтується на використанні моделі машинного навчання для локальної інтерпретованості моделей LIME, що дозволяє візуалізувати вплив використання окремих слів на рішення моделі щодо належності тексту до різних типів кіберзалякувань.

Розроблений метод забезпечує три подання інтерпретації результатів виявлення кіберзалякувань: інтерпретація результатів за кольоровою палітрою, інтерпретація результатів за діаграмами локальної важливості слів, інтерпретація результатів за діаграмами загальної важливості слів. Результати експериментів показали, що створений метод забезпечує інтерпретацію рішень щодо результатів нейромережевого виявлення кіберзалякувань на рівні, достатньому для розуміння людиною ознак тексту, які виплинули на прийняття рішень штучним інтелектом щодо виявлення типів кіберзалякувань.

Ключові слова: кіберзалякування, нейронні мережі, інтерпретація результатів, BERT, LIME.

SOBKO OLENA
Khmelnytskyi National University

METHOD FOR INTERPRETING THE RESULTS OF CYBERBULLYING DETECTION IN TEXTUAL CONTENT BY MEANS OF ARTIFICIAL INTELLIGENCE

The article proposes the method for interpreting the results of cyberbullying detection in textual content by means of artificial intelligence, which is intended to explain the decisions of the neural network model regarding the types of cyberbullying identified in the textual content. The method is original in that it interprets the results for each detected type of cyberbullying separately, which is achieved by using a multi-label classifier of a transformer neural network architecture and an interpretation model of a machine learning model. By using the trained BERT neural network model for multi-label classification of cyberbullying types in the input text sample, different types of cyberbullying are detected with the percentage of each of them. According to the developed method, an approach based on the use of a machine learning model for the local interpretability of LIME models is used for the visual interpretation of the results of cyberbullying detection, which allows you to visualize the impact of the use of individual words on the model's decision regarding whether the text belongs to different types of cyberbullying.

The developed method provides three views of the interpretation of the results of cyberbullying detection: interpretation of the results according to the color palette, interpretation of the results according to the diagrams of the local importance of words, interpretation of the results according to the diagrams of the general importance of the words. The interpretation of the results according to the color palette consists in using the absolute value of the weights to determine the brightness of the color, where the brightest color indicates the greatest influence of the word on the decision made by the model, and the least bright color indicates the smallest influence, regardless of whether this influence was positive or negative. The interpretation of the results based on the diagrams of the local importance of words is provided by constructing diagrams of the influence of individual words of the text on the probability of assigning this text to a specific type of cyberbullying, which allows you to see how the model evaluates the weight of each word in the text depending on its contribution to the model's decision. The interpretation of the results from the charts of the overall importance of words is provided by forming a set of 10 words that the model considers important regardless of the specific type of cyberbullying.

The results of the experiments showed that the created method provides interpretation of decisions regarding the results of neural network detection of cyberbullying at a level sufficient for a person to understand the features of the text, which resulted in decision-making by artificial intelligence regarding the detection of types of cyberbullying.

Keywords: cyberbullying, neural networks, interpretation of results, BERT, LIME.

Аналіз предметної області та постановка задачі

Проблема кіберзалякування з часом стає дедалі актуальнішою у зв'язку зі зростанням кількості користувачів соціальних мереж, а також зі зменшенням нижньої вікової границі таких користувачів. Таким чином відбувається зростання попиту на системи виявлення кіберзалякувань у текстовому контенті [1]. Завдяки розвитку технологій обробки природної мови, зокрема моделей на основі трансформерів, таких як BERT, стало можливим розробляти системи, які ефективно виявляють випадки кіберзалякувань а також класифікують за різними типами [2]. Однак високий рівень ефективності часто супроводжується складністю в інтерпретації результатів, що ставить під сумнів використання таких моделей у чутливих і соціально важливих контекстах, як кіберзалякування. Зважаючи на це, інтерпретація результатів моделі для виявлення кіберзалякувань в текстовому контенті є важливою для забезпечення прозорості та довіри користувачів до рішень, наданих штучним інтелектом [3].

У статті пропонується метод інтерпретації результатів виявлення кіберзалякувань, за допомогою якого надаватиметься пояснення рішень нейромережевої моделі щодо визначених у текстовому контенті типів

кіберзалякувань, наприклад кіберзалякування за віком, етнічним походженням, гендером, релігією, тощо.

Останні публікації

Проблема виявлення кіберзалякувань є важливою через його значний негативний вплив на психічне здоров'я, особливо молоді та підлітків. Сучасні підходи до виявлення кіберзалякування базуються на методах обробки природної мови, що дозволяють аналізувати текстовий контент для виявлення та класифікації різних типів кібербулінгу [4].

У дослідженні [5] досліджується проблема виявлення кіберзалякувань. Серед протестованих моделей, таких як Random Forest, XgBoost, Naive Bayes, SVM, CNN, RNN та BERT, найвищі результати продемонструвала модель BERT, досягнувши 88,8% точності в задачі бінарної класифікації та 86,6% точності в мультилейбловій класифікації.

Автори дослідження [6] запропонували нову теорію для виявлення кіберзалякування, в рамках якої були протестовані моделі Support Vector Machine, Naive Bayes і Logistic Regression у поєднанні з різними методами обробки природної мови. Автори зазначають, що точність виявлення кібербулінгу підвищується завдяки використанню аналізу настроїв, аналізу N-грам, а також нетрадиційних методів виділення ознак, таких як TF-IDF і виявлення ненормативної лексики. Комбінований підхід дозволяє досягти точності виявлення на рівні 75,17%.

Деякі з авторів також пропонують підходи до інтерпретації результатів виявлення кіберзалякувань у текстовому контенті. Наприклад, у [7] пропонується уніфікована модель BiLSTM-LIME для багатокласової класифікації контенту кіберзалякування на платформі Twitter. Автори стверджують, що техніка LIME надає високого рівня пояснення, висвітлюючи найбільш доречні токени, які сприяли прийняттю рішення моделлю.

Автори дослідження [8] запропонували новий підхід до виявлення та класифікації кіберзалякування у текстах соціальних медіа за допомогою ансамблю BERT та SVM з пошуком у сітці для багатокласової класифікації. Порівняння з іншими моделями машинного та глибокого навчання показало, що запропонована модель досягає точності 90% на тестових даних, перевершуючи інші. Для інтерпретації прогнозів використано техніку SHAP.

Метою роботи є розробка методу інтерпретації результатів виявлення кіберзалякувань у текстовому контенті засобами штучного інтелекту, що призначений для пояснення рішень нейромережевої моделі щодо визначених у текстовому контенті типів кіберзалякувань. Створений метод має забезпечувати інтерпретацію рішень щодо результатів нейромережевого виявлення кіберзалякувань для розуміння людиною ознак тексту, які виплинули на прийняття рішень штучним інтелектом щодо виявлення типів кіберзалякувань.

Основна частина

Метод інтерпретації результатів виявлення кіберзалякувань у текстовому контенті засобами штучного інтелекту передбачає створення візуального пояснення результатів нейромережевої моделі щодо виявлених типів кіберзалякувань у текстовому контенті. Схему методу інтерпретації результатів виявлення кіберзалякувань подано на рисунку 1.

Рис. 1. Схема методу інтерпретації результатів виявлення кіберзалякувань у текстовому контенті

Вхідними даними наведеного на рисунку 1 методу є навчена модель BERT для мультилейблової класифікації, яка здатна розпізнавати різні типи кіберзалякування, такі як віковий, етнічної приналежності, гендеру, релігійні та окремий узагальнений тип, що містить інші типи кіберзалякувань [1]. Використовується інтерпретаційна модель, яка дозволяє пояснити вплив окремих слів чи фраз на результат класифікації [9]. Список класів кіберзалякування включає типи кіберзалякувань, за якими модель здійснює класифікацію та виконує інтерпретацію результатів інтерпретаційна модель. Також до вхідних даних належить текст, який аналізується на наявність ознак різних типів кіберзалякувань та виконується інтерпретація результатів.

На першому кроці вхідний текст перетворюється у послідовність токенів за допомогою токенізатора, який розбиває текст на окремі елементи (слова або частини слів) [10]. Так як вхідною моделлю для мультилейблової класифікації є BERT, то для нього та подібних трансформерів токенізатор перетворює текстовий вхід у числові послідовності, з якими модель може працювати [11].

На другому кроці модель BERT, натренована на мультилейбловій класифікації, прогнозує ймовірність належності тексту до кожного з можливих класів кіберзалякування [12]. Таким чином, модель визначає, чи містить текст ознаки певних типів кіберзалякування (віковий, за етнічною приналежність, гендерний, релігійний та інші види кіберзалякувань), надаючи відсоток ймовірності присутності кожного з типів кіберзалякувань.

На третьому кроці результати класифікації пояснюються і візуалізуються [13]. За допомогою інтерпретаційної моделі, показується, які саме слова чи фрази найбільше вплинули на класифікацію тексту за певним типом кіберзалякування, що допомагає зрозуміти, які частини тексту сприяли ідентифікації ознак конкретного виду кіберзалякування. У якості інтерпретаційної моделі для мультилейблової класифікації часто використовуються такі методи, як [14]:

– Local Interpretable Model-agnostic Explanations (LIME), який генерує локальні пояснення для кожного передбачення, показуючи, 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 ознакам,

– Transformers Interpret, що є інтерпретаційною бібліотекою, яка спеціально розроблена для роботи з моделями на основі нейромереж трансформерів, такими як BERT, GPT, RoBERTa та іншими моделями з бібліотеки Hugging Face;

– методи на основі Attention, які дозволяють аналізувати ваги уваги трансформерів (наприклад, у моделі BERT) для розуміння важливості окремих слів чи фраз у прийнятті рішень моделі.

Вихідними даними є сила прояву (вага) кожного виду кіберзалякування в тексті, яка визначається у вигляді ймовірностей, що показують ступінь наявності ознак кожного класу кіберзалякування, надається мітка про наявність або відсутність ознак кожного виду кіберзалякування у вигляді числових значень, що показують ймовірності присутності ознак кіберзалякування за кожним класом. Також метод забезпечує візуалізацію впливу ознак на рішення про віднесення тексту до конкретного класу кіберзалякування шляхом графічного представлення тексту, де важливі слова підсвічуються відповідно до їх значущості для кожного класу.

Отже, наведений метод інтерпретації результатів виявлення кіберзалякувань дозволить отримати пояснення щодо прийнятих рішень моделі мультилейблової класифікації текстового контенту.

Експеременти, дослідження та результати

Для навчання моделі BERT, яка використовується на кроці 2 методу інтерпретації результатів виявлення кіберзалякувань (рисунок 1), використано датасет «Cyberbullying Classification» [15], що містить текстові повідомлення з мітками про належність кожного повідомлення до одного з класів: Age, Ethnicity, Gender, Religion, Other type of cyberbullying, Not cyberbullying, детальна статистика кількості записів наведена на рисунку 2.

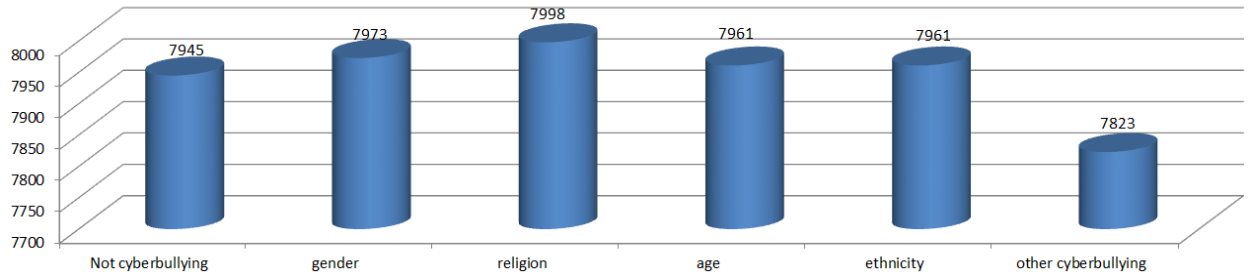
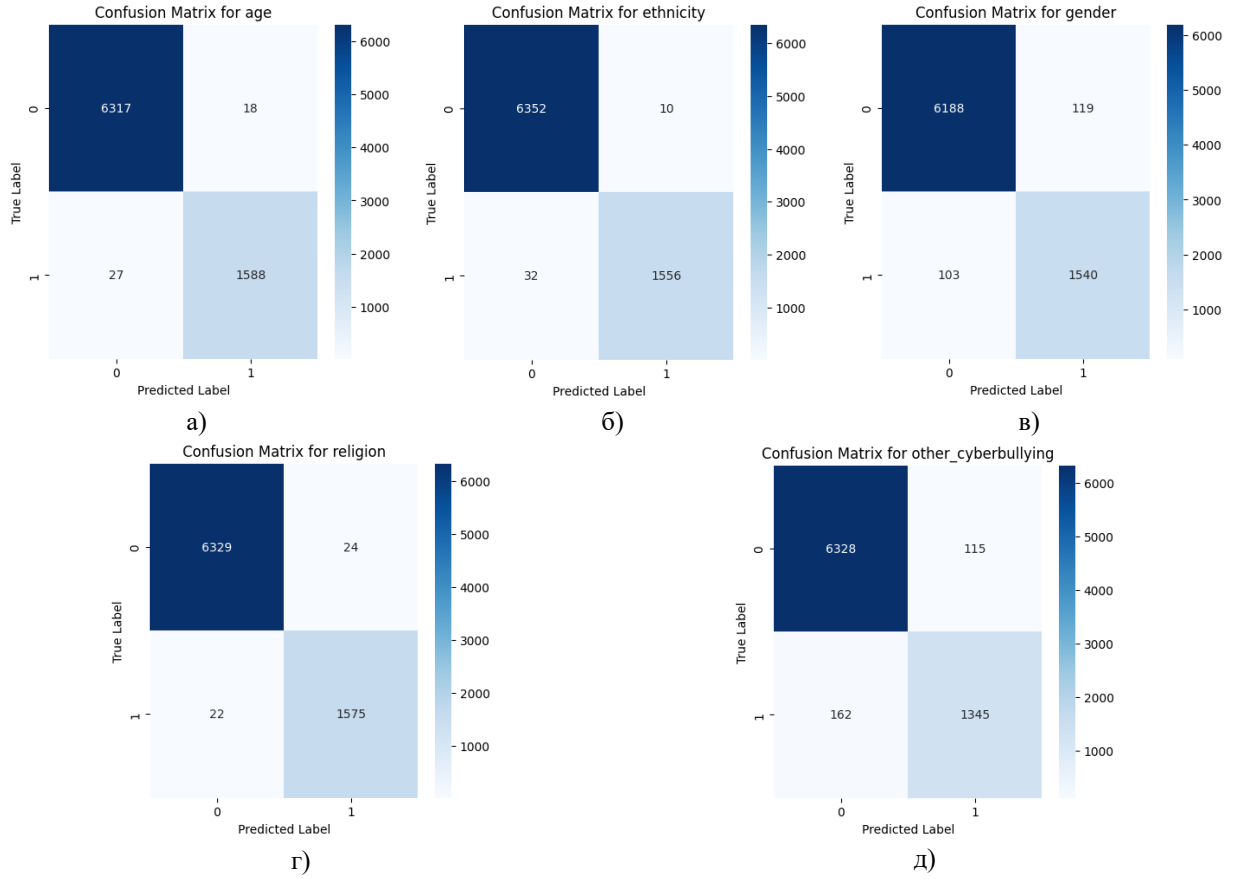


Рис. 2. Статистика кількості записів у класах датасету для виявлення кіберзалякувань

Для навчання моделі BERT мультилейбловій класифікації не використовувався клас «Not cyberbullying», тому перед навчанням він був видалений з датасету [16]. А клас «Other type of cyberbullying» був аугментований синтетичними зразками за допомогою методики SMOTE-балансування. Таким чином, шляхом попередньої обробки датасету «Cyberbullying Classification» отримано збалансовану навчальну вибірку, яка була використана для навчання моделі BERT для завдання мультилейблової класифікації типів

кіберзалякувань у текстовому контенті.

Для дослідження ефективності методу інтерпретації результатів виявлення кіберзалякувань у текстовому контенті засобами штучного інтелекту було використано середовище Google Colab. Було навчено нейромережеву модель BERT на такі типи кіберзалякувань, як вікове, гендерне, релігійне, етнічне та окремий тип – інші кіберзалякування [17]. На рисунку 3 подано матриці сплутувань для кожного типу кіберзалякувань.



а) б) в)

г) д)

Рис. 3. Матриці сплутувань для типів кіберзалякувань: а) вікове; б) етнічне; в) гендерне; г) релігійне; д) інші типи кіберзалякувань

Показники макрометрик навченої моделі BERT для мультилейбловій класифікації типів кіберзалякувань отримали значення Accuracy 0.956478, Precision 0.963677, Recall 0.956478, F1 Score 0.960019, що вказує на високі здатність моделі виявляти види кіберзалякувань у текстовому контенті.

Наприклад, для дослідження було використано англомовний текстовий зразок «Your God has no place here. Stick to your country and stop dragging your outdated traditions and religions into ours» (укр: *«Твоєму Богу тут не місце. Дотримуйся своєї країни і перестань перетягувати свої застарілі традиції та релігії в нашу»*). Модель BERT виявила наступні ймовірності наявності видів кіберзалякувань у даному текстовому зразку, як:

– вікове кіберзалякування: 0.06%

– етнічне кіберзалякування: 0.08%;

– гендерне кіберзалякування: 0.10%;

– інший тип кіберзалякування: 0.09%;

– релігійне кіберзалякування: 99.86%.

Шляхом застосування моделі LIME для інтерпретації результатів моделі BERT для мультилейблової класифікації типів кіберзалякувань у текстовому зразку, було отримано результати візуальної інтерпретації виявлених типів кіберзалякувань з використанням абсолютного значення ваг, що подані на рисунку 4. Для інтерпретації прийнятих рішень моделлю BERT слова виділяються кольорами – найбільш яскравий колір означає найбільше значення ваги слова, тобто це слово мало найбільший вплив, найсвітліше – найменший.

Як видно з рисунку 4, слова, що мають додатні та від'ємні значення виділяються однаково яскравим кольором. У такому виді візуальної інтерпретації використовується абсолютне значення ваги для визначення яскравості кольору, через що від'ємні та додатні значення мають однакову яскравість. Від'ємні значення ваг вказують на те, що слово зменшує ймовірність конкретного класу, тоді як додатні значення збільшують ймовірність цього класу і мають однаковий вплив на прийняте моделлю рішення. А значення ваги, без врахування знаку біля неї вказує на скільки він був сильний.

Рис. 4. Використання абсолютного значення ваги для визначення яскравості кольору для інтерпретації результаів виявлення різних типів кіберзалякувань

У випадку з LIME, важливо показати не лише те, наскільки сильний вплив має слово, але й чи цей вплив позитивний (збільшує ймовірність) або негативний (зменшує ймовірність). Тому реалізовано ще один підхід до зміни яскравості так, щоб від'ємні значення були менш яскравими і використовували інший колірний відтінок для від'ємних та додатних значень. Результат такої візуальної інтерпретації подано на рисунку 5.

Вибір окремих кольорових палітр для додатних і від'ємних значень у візуалізації інтерпретацій LIME є доцільним з кількох причин, які випливають із принципів сприйняття інформації та аналізу результатів моделей машинного навчання. Від'ємні ваги, за своєю природою, вказують на зменшення ймовірності певного класу, тоді як додатні – на її збільшення, тому використання однакових візуальних характеристик для цих двох типів впливу може призводити до хибного трактування результатів, якщо не було наведено додаткових типів візуальної інтерпретації, адже значення однакової інтенсивності, але протилежного знаку, можуть виглядати однаково важливими, хоча їхня роль принципово відрізняється.

Рис. 5. Використання підходу для визначення яскравості кольору для інтерпретації результатів виявлення типів кіберзалякувань з урахуванням негативного чи позитивного типу впливу на результат

Додатково створено діаграми для графічної інтерпретації впливу окремих слів тексту на ймовірність віднесення цього тексту до конкретного типу кіберзалякування (рисунок 6).

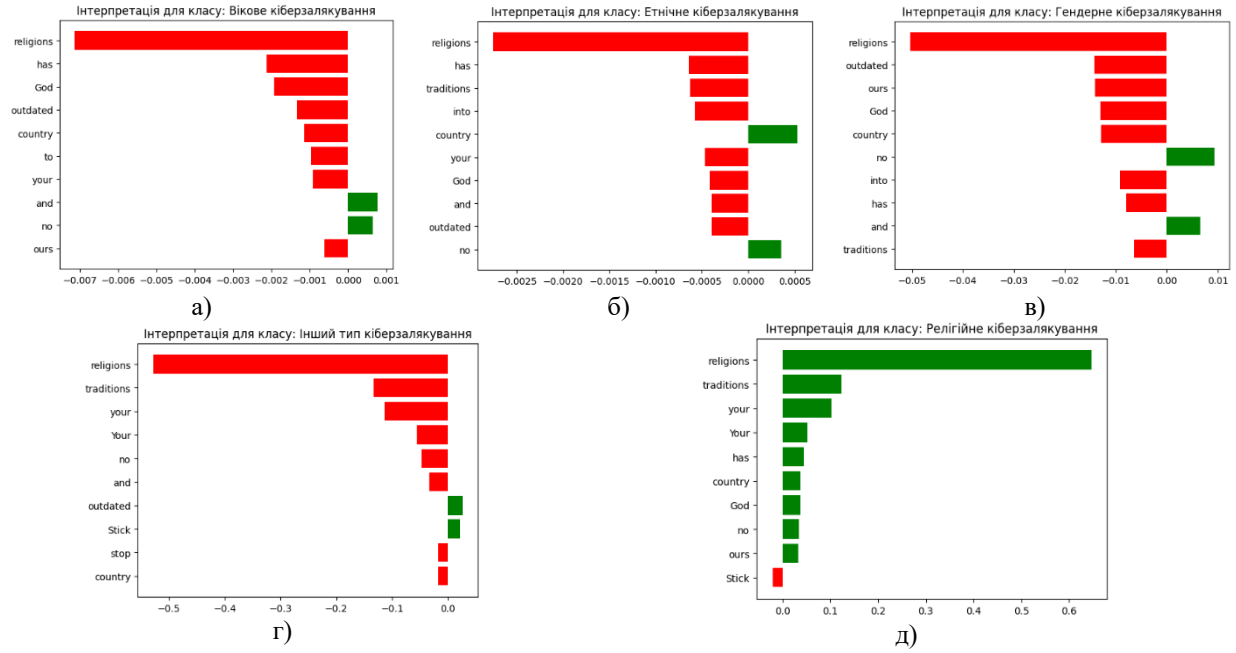


Рис. 6. Діаграми для графічної інтерпретації впливу окремих слів тексту на ймовірність віднесення цього тексту до конкретного типу кіберзалякування:
а) вікове; б) етнічне; в) гендерне; г) релігійне; д) інші типи кіберзалякувань

Діаграми ілюструють, як модель оцінює вагу кожного слова в тексті, залежно від його внеску в прийняте рішення. Вплив слів представлено у вигляді горизонтальних стовпців, довжина яких відповідає величині впливу (ваги), а колір – напрямку цього впливу. Червоні стовпці, відображають негативний вплив слів, тобто зменшення ймовірності віднесення тексту до обраного класу, тоді як зелені стовпці відображають позитивний вплив, збільшуючи цю ймовірність. Величина впливу вимірюється числом, і ці значення представлені на горизонтальній осі графіка.

Також обчислено середнє значення важливості кожного слова для всіх класів, що дає зрозуміти загальний вплив кожного слова незалежно від конкретного типу кіберзалякування. Обчислені значення візуалізовано через відповідну діаграму (рисунок 7).

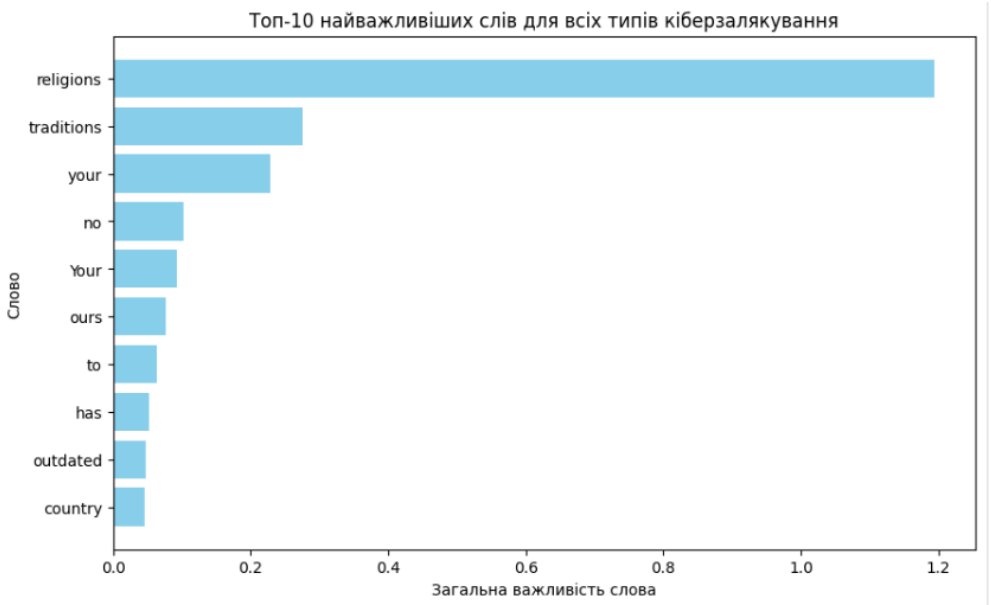


Рис. 7. Діаграма з відображенням середнього значення важливості топ-10 слів для всіх класів

Обчислення загального впливу слів на результати моделі для всіх типів кіберзалякувань також є важливим для інтерпретації роботи моделі та розуміння її прийнятих рішень. Аналіз здійснюється шляхом агрегації ваг слів, які модель оцінює для кожного класу. Використовується модуль ваги, тобто абсолютна величина, яка вказує на інтенсивність впливу слова незалежно від його позитивного чи негативного внеску. Такий підхід дозволяє виявити слова, які модель вважає важливими незалежно від конкретного типу кіберзалякування. Наприклад, слова, що відображають різні типи кіберзалякувань, можуть мати високу вагу для кількох класів. Якщо слово має високий загальний вплив, це може вказувати на його універсальну роль у контексті кіберзалякування. Наприклад, слова, що вказують на етнічну приналежність або релігію, можуть

мати високий вплив для кількох класів, таких як «етнічне кіберзалякування» і «релігійне кіберзалякування», що може вказувати про потенційну крос-модальність ознак, які модель використовує для прийняття рішень. Якщо ж слова впливають тільки на один клас, це підкреслює їхню специфічність, що може свідчити про унікальні патерни мовлення для цього типу кіберзалякування.

Отже, запропоновані візуальні інтерпретації результатів виявлення кіберзалякувань у текстовому контенті дозволяють оцінити, чи модель використовує релевантні ознаки для ухвалення рішень, чи її поведінка може бути обумовлена випадковими або нерелевантними факторами. Наприклад, якщо у тексті виявляються слова, які не мають змістового зв'язку з віковим кіберзалякуванням, але мають високий вплив, це може свідчити про наявність помилки або упередження в моделі.

Висновки

У статті запропоновано метод інтерпретації результатів виявлення кіберзалякувань у текстовому контенті засобами штучного інтелекту, що призначений для пояснення рішень нейромережевої моделі щодо визначених у текстовому контенті типів кіберзалякувань. Метод оригінальний тим, що здійснює інтерпретацію результатів для кожного виявленого типу кіберзалякування окремо, що досягається шляхом використання мультилейблового класифікатора нейромережевої архітектури трансформер та інтерпретаційної моделі машинного навчання.

Шляхом використання навченої нейромережевої моделі BERT для мультилейблової класифікації типів кіберзалякувань у вхідному текстовому зразку виявляються різні типи кіберзалякувань із вказанням відсотку наявності кожного з них. Згідно розробленого методу, для візуальної інтерпретації результатів виявлення кіберзалякувань використано підхід, який ґрунтується на використанні моделі машинного навчання для локальної інтерпретованості моделей LIME, що дозволяє візуалізувати вплив використання окремих слів на рішення моделі щодо належності тексту до різних типів кіберзалякувань.

Розроблений метод забезпечує три подання інтерпретації результатів виявлення кіберзалякувань: інтерпретація результатів за кольоровою палітрою, інтерпретація результатів за діаграмами локальної важливості слів, інтерпретація результатів за діаграмами загальної важливості слів. Інтерпретація результатів за кольоровою палітрою полягає у використанні абсолютного значення ваг для визначення яскравості кольору, де найбільш яскравий колір вказує на найбільший вплив слова на прийняте моделлю рішення, а найменш яскравий колір вказує на найменший вплив, не беручи до уваги позитивним чи негативним був цей вплив. Проте використання лише такого типу інтерпретації недостатньо, адже необхідно також і розуміти яким чином слово впливало на рішення моделі, адже від'ємні ваги вказують на зменшення ймовірності певного класу, а додатні – на її збільшення. Тому було реалізовано інтерпретацію рішень моделі BERT з врахуванням впливу на результат. Інтерпретація результатів за діаграмами локальної важливості слів забезпечується шляхом побудови діаграм впливу окремих слів тексту на ймовірність віднесення цього тексту до конкретного типу кіберзалякування, що дозволяє побачити, як модель оцінює вагу кожного слова в тексті залежно від його внеску в прийняте рішення моделі. Інтерпретація результатів за діаграмами загальної важливості слів забезпечується шляхом формування множини з 10 слів, які модель вважає важливими незалежно від конкретного типу кіберзалякування.

Результати експериментів показали, що створений метод забезпечує інтерпретацію рішень щодо результатів нейромережевого виявлення кіберзалякувань на рівні, достатньому для розуміння людиною ознак тексту, які виплинули на прийняття рішень штучним інтелектом щодо виявлення типів кіберзалякувань. Запропонований метод інтерпретації результатів виявлення кіберзалякувань у текстовому контенті належить до категорії засобів візуальної аналітики рішень штучного інтелекту, створення якої є практичною необхідністю для забезпечення етичності, прозорості та довіри до таких систем штучного інтелекту у суспільстві, особливо стосовно таких чутливих тем як виявлення кіберзалякувань. Відповідно, дослідження підкреслює важливість не лише точності моделей, але й їхньої поясненності, яка є ключовим фактором у побудові довіри до систем штучного інтелекту.

Література

1. 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.

2. 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.

3. 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.

4. 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.

5. 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.

6. 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.

7. Gongane V.U. Explainable AI for Reliable Detection of Cyberbullying / V.U. Gongane, M.V. Munot, A. Anuse // 2023 IEEE Pune Section International Conference (PuneCon): Proc., Pune, India. – 2023. – С. 1–6.

8. 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.

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. 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.

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

12. 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.

13. 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.

14. 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.

15. Cyberbullying Classification Dataset [Електронний ресурс]. – Kaggle. – Режим доступу: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification (дата звернення: 17.17.2024).

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. 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.

2. 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.

3. 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.

4. 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.

5. 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 (data zvernennia: 17.17.2024).

6. 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.

7. Gongane V.U. Explainable AI for Reliable Detection of Cyberbullying / V.U. Gongane, M.V. Munot, A. Anuse // 2023 IEEE Pune Section International Conference (PuneCon): Proc., Pune, India. – 2023. – S. 1–6.

8. 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.

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. – S. 73–77.

10. 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.

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

12. 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.

13. 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.

14. 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. – S. 184–195.

15. Cyberbullying Classification Dataset [Elektronnyi resurs]. – Kaggle. – Rezhym dostupu: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification (data zvernennia: 17.17.2024).

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.