УДК 004.8

Собко О.В.

Хмельницький національний університет

ІНТЕРПРЕТАЦІЯ РЕЗУЛЬТАТІВ ВИЯВЛЕННЯ КІБЕРЗАЛЯКУВАНЬ У ТЕКСТАХ З ВИКОРИСТАННЯМ НЕЙРОННИХ МЕРЕЖ

Розроблено та досліджено метод інтерпретації результатів виявлення кіберзалякувань, за допомогою якого надається можливість подавати результати в зрозумілому для користувача вигляді й пояснювати рішення нейромережевої моделі щодо визначених у текстовому контенті типів кіберзалякувань. Запропоновані візуальні інтерпретації результатів виявлення кіберзалякувань дозволяють оцінити, використовує нейромережева модель релевантні ознаки для ухвалення рішень чи результати обумовлені нерелевантними факторами. Результати прикладних 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.

Developed and researched the method of interpreting the results of cyberbullying detection has been, which makes it possible to present the results in a user-friendly form and to explain the decisions of the neural network model regarding the types of cyberbullying identified in the textual content. The proposed visual interpretations of cyberbullying detection results allow you to assess whether the neural network model uses relevant features for decision-making or whether the results are due to irrelevant factors. The results of applied experiments showed that the created method provides interpretation of decisions regarding the results of neural network detection of cyberbullying at a level sufficient for human understanding of text features that resulted in decision-making by artificial intelligence regarding the detection of types of cyberbullying.

Проблема кіберзалякування стає все більш актуальною через зростання кількості користувачів соціальних мереж та зниження вікового порогу їхнього використання [1]. Це зумовлює підвищений інтерес до розробки систем для автоматичного виявлення кібербулінгу у текстовому контенті [2, 3]. Завдяки прогресу в галузі обробки природної мови, зокрема появі трансформерних моделей, таких як BERT, стало можливим створення систем, які ефективно ідентифікують випадки кібербулінгу та класифікують їх за різними типами [4]. Однак висока продуктивність таких систем нерідко супроводжується труднощами в інтерпретації їхніх результатів [4, 5], що ускладнює їхнє застосування в чутливих і соціально значущих сферах, таких як боротьба з кіберзалякуванням [6]. У цьому контексті інтерпретація результатів аналізу є ключовим фактором для забезпечення прозорості рішень та довіри до використання штучного інтелекту [7].

Значущість виявлення кіберзалякування обумовлена його серйозним негативним впливом на психічне здоров'я, особливо серед підлітків і молоді. Сучасні методи для ідентифікації кібербулінгу базуються на технологіях обробки природної мови, які дозволяють аналізувати текстовий контент і класифікувати різноманітні форми кіберзалякування [8].

Метою роботи є розробка та дослідження методу інтерпретації результатів виявлення кіберзалякувань, за допомогою якого надається можливість подавати результати в зрозумілому для користувача вигляді й пояснювати рішення нейромережевої моделі щодо визначених у текстовому контенті типів кіберзалякувань.

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

Рисунок 1 – Схема методу інтерпретації результатів виявлення кіберзалякувань

Вхідними даними 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 [9]. У представленому підході використовується інтерпретаційна модель, яка дозволяє пояснювати вплив окремих слів або фраз на результати класифікації [10]. Система класифікації кіберзалякувань працює з набором класів, які відображають різні типи кібербулінгу, забезпечуючи не лише класифікацію, але й інтерпретацію результатів. Аналіз здійснюється на основі вхідного тексту, який перевіряється на наявність ознак кіберзалякувань різних типів із подальшим поясненням результатів.

На першому етапі текст проходить токенізацію, під час якої він розбивається на токени (слова чи їх частини) за допомогою спеціалізованого токенізатора. Оскільки основною моделлю для мультилейблової класифікації є BERT, текст перетворюється на числові послідовності, з якими трансформер може працювати.

На другому етапі модель BERT, натренована для мультилейблової класифікації, прогнозує ймовірність належності тексту до кожного класу кіберзалякування. Це дозволяє визначити, чи присутні у тексті ознаки конкретних типів кібербулінгу, таких як віковий, етнічний, гендерний чи релігійний, з поданням ймовірностей для кожного класу.

Третій етап передбачає пояснення та візуалізацію отриманих результатів. Інтерпретаційна модель дозволяє виділити слова чи фрази, які найбільше вплинули на класифікацію тексту за певним типом кіберзалякування. Це допомагає зрозуміти, які елементи тексту сприяли ідентифікації ознак кожного класу. Результати представлені у вигляді графічної візуалізації: важливі слова підсвічуються кольорами відповідно до їх впливу — яскраві кольори вказують на більший вплив, світліші — на менший.

Вихідними даними є ймовірності наявності ознак кожного виду кіберзалякування в тексті, представлені у числовій формі, а також візуалізація впливу слів на рішення моделі. Графічне представлення тексту дозволяє інтуїтивно оцінити вагу слів у процесі класифікації.

Для навчання моделі BERT використовувався датасет «Cyberbullying Classification» [11], який містить текстові повідомлення з мітками про належність до певного виду кіберзалякування. Клас «Not cyberbullying» було виключено перед навчанням для оптимізації задачі, а клас «Other type of cyberbullying» був доповнений синтетичними зразками з використанням SMOTE-балансування, що забезпечило збалансованість навчальної вибірки.

Для оцінки ефективності запропонованого підходу було створено тестове програмне забезпечення, яке реалізує інтерпретацію результатів BERT за допомогою методу LIME. Отримані результати візуалізуються у вигляді абсолютних значень ваг слів, де колірна шкала демонструє вплив кожного слова на класифікацію тексту. На рисунку 2 представлено приклад такої візуалізації, де найбільш насичений колір відповідає найвищій значущості слова у рішенні моделі.

Рисунок 2 – Використання значення ваги для визначення яскравості кольору для інтерпретації результатів виявлення різних типів кіберзалякувань

Як видно на рисунку 2, слова з додатними та від'ємними значеннями ваг виділяються однаковою яскравістю кольору. Це пояснюється тим, що для візуальної інтерпретації використовується абсолютне значення ваги, яке визначає інтенсивність кольору, незалежно від знаку. Від'ємні ваги вказують на те, що слово зменшує ймовірність належності тексту до конкретного класу, тоді як додатні ваги, навпаки, підвищують цю ймовірність. Обидва типи значень мають однакову силу впливу на рішення моделі, яка оцінюється за абсолютною величиною ваги.

Також створено діаграми для графічного відображення внеску окремих слів у класифікацію тексту за конкретним типом кіберзалякування. Приклад такої діаграми наведено на рисунку 3.

Рисунок 3 – Приклад діаграми для графічної інтерпретації впливу окремих слів тексту на ймовірність віднесення цього тексту до певного типу кіберзалякування

На діаграмах демонструється, як модель визначає значущість кожного слова в тексті залежно від його впливу на прийняте рішення. Червоні стовпці представляють слова з негативним впливом, які зменшують ймовірність віднесення тексту до відповідного класу, тоді як зелені стовпці відображають слова з позитивним впливом, що збільшують цю ймовірність. Рівень впливу кожного слова представлений числовими значеннями, які розташовані вздовж горизонтальної 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 4).

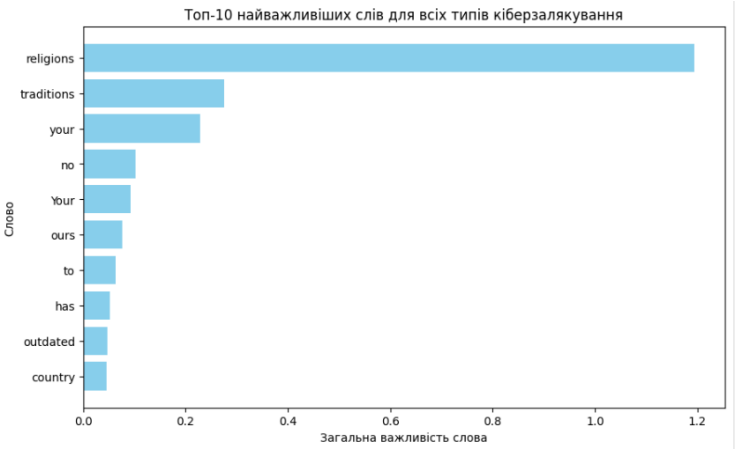

Рисунок 4 – Діаграма з візуалізацією середнього значення важливості найбільш значущих слів

Аналіз здійснюється шляхом підсумовування ваг слів, які модель оцінює для кожного класу. При цьому використовується абсолютне значення ваги, що дозволяє оцінити силу впливу слова, незалежно від його позитивного чи негативного внеску. Такий метод дає змогу визначити ключові слова, які модель вважає значущими, незалежно від конкретного типу кіберзалякування. Наприклад, одне й те саме слово, пов'язане з різними типами кіберзалякувань, може мати високі ваги для кількох класів одночасно.

Таким чином, запропоновані способи візуалізації результатів виявлення кіберзалякувань у текстовому контенті допомагають зрозуміти, чи використовує модель релевантні ознаки для ухвалення рішень або ж її результати базуються на

випадкових чи нерелевантних характеристиках. Наприклад, якщо модель надає високий вплив словам, які не мають змістового зв'язку з віковим кіберзалякуванням, це може свідчити про потенційні помилки або упередження в її роботі.

Отже, було розроблено та досліджено метод інтерпретації результатів виявлення кіберзалякувань, за допомогою якого надається можливість подавати результати в зрозумілому для користувача вигляді й пояснювати рішення нейромережевої моделі щодо визначених у текстовому контенті типів кіберзалякувань. Запропоновані візуальні інтерпретації результатів виявлення кіберзалякувань дозволяють оцінити, використовує нейромережева модель релевантні ознаки для ухвалення рішень чи резальтати обумовлені нерелевантними факторами. Результати прикладних 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.

Перелік посилань

1. 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, 2024, vol. 3688, pp. 16-28.
2. Молчанова М.О., Мазурець О.В., Собко О.В., Віт. Р.В., Назаров В.В. Алгоритм виявлення аб'юзивного вмісту в україномовному аудіоконтенті для імплементації в об'єктно-орієтовану інформаційну систему. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №1 (331). С. 101-106.
3. Мазурець О.В., Козенко О.В., Собко О.В. Метод автоматизованого підбору відповідей на користувацькі запитання за семантичною подібністю. Матеріали XII Всеукраїнської науково-практичної конференції «Глушковські читання». Київ – 2023. с. 106-109.
4. Molchanova M., Mazurets O., Sobko O., Boiarchuk I. Object-Oriented Approach for Ethnic Enmity Detection in Text Messages by NLP. Proceedings of XXI International Scientific and Practical Conference «Scientific Achievements and Innovations as a Way to Success». May 1-3, 2024. Vilnius, Lithuania. 2024. Pp. 73-77.
5. Mazurets O.V., Sobko O.V., Molchanova M.O., Zalutska O.O., Yurchak A.V. Practical Implementation of Neural Network Method for Stress Features Detection by Social Internet Networks Posts. Global Science: Prospects and Innovations. Proceedings of the II International Scientific and Theoretical Conference «Scientific Review of the Actual Events, Achievements and Problems». May 31, 2024. Berlin, Federal Republic of Germany: International Center of Scientific Research. 2024. Pp. 160-167.

6. Zalutska O., Molchanova M., Sobko O., Mazurets O., Pasichnyk O., Barmak O., Krak I. Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network. CEUR Workshop Proceedings, 2023, vol. 3387, pp. 344-356.
7. Kovalchuk O., Slobodzian V., Sobko O., Molchanova M., Mazurets O., Barmak O., Krak I., Savina N. Visual Analytics-Based Method for Sentiment Analysis of COVID-19 Ukrainian Tweets. Book Chapter. Lecture Notes on Data Engineering and Communications Technologies. 2023. Vol. 149. pp. 591–607.
8. Молчанова М.О., Мазурець О.В., Собко О.В., Кліменко В.І., Андрощук В.І. Метод нейромережевого виявлення кібербулінгу з використанням хмарних сервісів та об'єктно-орієнтованої моделі. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2024. №2 (333). С. 200-206.
9. Собко О. Метод інтелектуального виявлення та класифікації кіберзалякувань у текстовому контенті. Матеріали XII Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології ІУСТ-ОДЕСА-2024». 23-25.09.2024. Одеса. 2024. С.262-265.
10. Собко О.В. Метод інтелектуального виявлення кіберзалякувань у текстовому контенті. Розвитки інформаційно-керуючих систем та технологій.: монографія. Львів-Торунь : Lina-Pres, 2024. – С.267-287.
11. Cyberbullying Classification Dataset URL: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification