МЕТОД ВИЯВЛЕННЯ ТА КЛАСИФІКАЦІЇ КІБЕРЗАЛЯКУВАНЬ У ТЕКСТОВИХ ДАНИХ НЕЙРОМЕРЕЖЕВИМИ ЗАСОБАМИ

Собко О.В., olenasobko.ua@gmail.com

Хмельницький національний університет

Виявлення кіберзалякувань у текстових даних цифрового середовища зумовлена стрімким зростанням кількості користувачів соціальних мереж, месенджерів та інших онлайн-платформ. Кіберзалякування, або булінг у цифровому форматі, є серйозною соціальною проблемою, яка може спричинити значний негативний вплив на психоемоційний стан жертв, зокрема призводити до депресій, тривожних розладів, зниження самооцінки й навіть суїцидальних думок [1].

В умовах, коли масштаби поширення таких явищ ускладнюють їхнє своєчасне виявлення традиційними методами, на передній план виходять автоматизовані підходи на основі штучного інтелекту. Нейромережеві технології здатні ефективно аналізувати великі обсяги текстових даних, виявляючи характерні ознаки кіберзалякувань. Сучасні моделі машинного навчання, такі як глибокі нейронні мережі, використовуються для створення систем, які автоматично класифікують повідомлення за їхнім змістом та рівнем загрози [2].

Інструменти на основі штучного інтелекту дозволяють не лише виявляти кіберзалякування, а й сприяти їхньому попередженню шляхом надання користувачам рекомендацій або блокування шкідливого контенту. Завдяки високій точності та швидкості обробки даних, ці системи стають ключовим елементом боротьби з негативними проявами в онлайн-просторі, забезпечуючи більш безпечне цифрове середовище [3].

Існують два основні підходи до виявлення кіберзалякувань: бінарна класифікація («залякування» або «не залякування») та мультикласова, яка деталізує види залякувань (за релігійною, віковою, гендерною чи етнічною ознаками). Дослідження показують високу ефективність моделі BERT у цих завданнях, хоча проблеми дисбалансу класів залишаються актуальними. Для їх вирішення застосовуються методи балансування вибірок, такі як SMOTE, що

покращують точність алгоритмів, проте моделі потребують подальшого вдосконалення для зменшення хибних результатів [1].

Саме тому метою дослідження є створення методу виявлення та класифікації кіберзалякувань у текстових даних нейромережевими засобами.

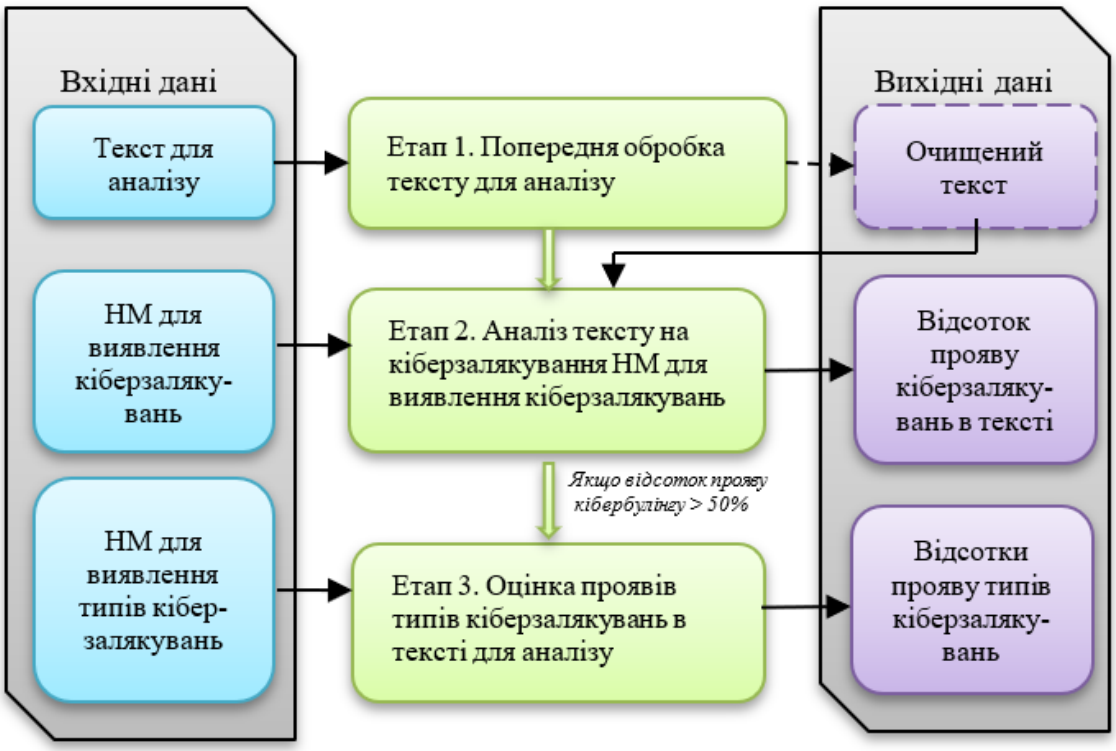

Рис.1. Схема методу виявлення та класифікації кіберзалякувань у текстових даних нейромережевими засобами

Метод виявлення та класифікації кіберзалякувань у текстових даних нейромережевими засобами складається з трьох основних етапів.

На першому етапі здійснюється попередня обробка текстуових даних Вхідні дані очищуються від зайвих символів, таких як знаки пунктуації, емодзі та пробіли. Після цього текст перетворюється у формат, придатний для аналізу нейромережевими методами.

Другий етап передбачає аналіз текстових даних на наявність ознак кіберзалякувань. Алгоритм перевіряє зміст повідомлення, оцінюючи ймовірність його належності до категорії з проявами кіберзалякувань. Якщо ця ймовірність перевищує певний поріг, текст передається на наступний етап для детального аналізу.

На третьому етапі відбувається класифікація типів кіберзалякувань. Алгоритм визначає конкретні категорії, до яких належить текст, та оцінює ступінь їх вираженості. Вихідними даними є як загальна оцінка рівня кіберзалякувань у тексті, так і розподіл за окремими типами.

Такий підхід дозволяє виявляти не лише факт кіберзалякування, а й його різновиди, що важливо для створення безпечного середовища в онлайн-комунікації.

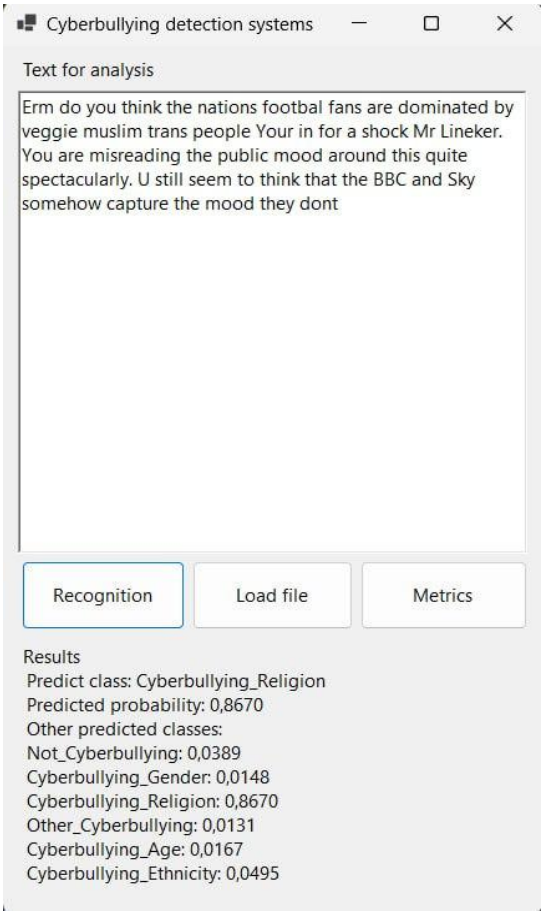

Рис. 2. Програмне забезпечення для оцінки ефективності запропонованого методу

Для оцінки ефективності запропонованого методу розроблено програмний застосунок (рис. 2), який аналізує текстові повідомлення на наявність кіберзалякувань та ідентифікує їхні типи у досліджуваних текстових зразках.

При виконанні етапу 2 метода у задачі бінарної класифікації було досягнуто найкращих результатів за допомогою нейромережевої моделі BiLSTM із такими показниками: Accuracy – 96%, Precision – 96%, Recall – 95,9% і F1 – 95,7%.

Для класифікації типів кіберзалякувань згідно етапу 3 методу нейромережева модель BERT продемонструвала такі показники: Accuracy – 94%, Precision – 93%, Recall – 93% і F1 – 93%.

Отже, для досягнення мети дослідження розроблено виявлення та класифікації кіберзалякувань у текстових даних нейромережевими засобами. Наведений метод дозволяє оцінювати загальний рівень кіберзалякувань у текстових даних та виконувати мультикласову класифікацію, надаючи окремі показники для різних типів залякувань, таких як вікові, релігійні, етнічні чи гендерні. Ефективність методу підтверджено розробленим програмним забезпеченням, яке аналізує текстові дані і визначає загальний рівень кіберзалякувань та рівні прояву кожного типу кіберзалякувань.

Список використаних джерел

1. Собко О.В. Виявлення та класифікація кіберзалякувань у цифрових текстах засобами штучного інтелекту. *Вимірювальна та обчислювальна техніка в технологічних процесах*. 2024. № 4. С. 143-152.
2. Собко О. Метод інтелектуального виявлення та класифікації кіберзалякувань у текстовому контенті. *Інформаційні управляючі системи та технології ІУСТ-ОДЕСА-2024.* (23-25 вересня 2024. Одеса). 2024. С. 262-265.
3. Собко О.В., Бармак О.В. Виявлення кіберзалякувань в інформаційному середовищі засобами машинного навчання. *Інформаційна, функційна і кібербезпека СКІФіК2024* : матеріали IV Всеукраїнської науково-технічної конференції (29-30 листопада 2024. Харків). 2024. С. 96-97.