Голові разової спеціалізованої
вченої ради PhD 8733
Хмельницького національного університету
доктору технічних наук, професору
Тетяні ГОВОРУЩЕНКО

РЕЦЕНЗІЯ
на дисертаційне дослідження Собко Олени Віталіївни
на тему «Методи виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту»,
подане на здобуття ступеня доктора філософії
з галузі знань 12 Інформаційні технології
за спеціальністю 122 Комп'ютерні науки

Актуальність теми та зв'язок з науковими планами наукових робіт університету. Сучасна інформаційна інфраструктура характеризується високим рівнем проникнення цифрових технологій у всі сфери суспільного життя, що супроводжується масштабним використанням електронних каналів комунікації. Одночасно з цим спостерігається зростання кількості проявів кіберзалякувань, які мають системний характер і справляють суттєвий вплив на психоемоційний стан користувачів цифрових платформ. Особливу загрозу становить контент, що містить вербальні форми насильства, які часто залишаються поза увагою традиційних засобів модерації через складність їх автоматизованого виявлення. У зв'язку з цим набуває актуальності застосування інструментарію штучного інтелекту для автоматизованої обробки текстових даних з метою виявлення та класифікації ознак кіберзалякувань.

Використання сучасних методів штучного інтелекту, зокрема архітектур на основі трансформерів, забезпечує високий рівень точності при виявленні семантичних конструкцій агресивного змісту. Водночас зберігається проблема репрезентативності навчальних вибірок, що обумовлює ризики дискримінаційного характеру рішень моделей. Наявність упередженості, обумовленої недостатнім урахуванням вікових, гендерних, релігійних та інших соціокультурних чинників, знижує довіру до результатів автоматизованого аналізу. Крім того, непрозорість алгоритмічного висновку, характерна для більшості сучасних нейромережевих моделей, унеможливлює верифікацію

результатів з боку кінцевих користувачів та обмежує можливості впровадження таких систем у відповідальні сфери застосування.

Таким чином, наукова задача, пов'язана з розробленням методів виявлення та класифікації кіберзалякувань у текстовому контенті, які забезпечують етичну нейтральність, репрезентативність навчального середовища та інтерпретованість результатів, є актуальною. Її вирішення сприятиме підвищенню ефективності систем моніторингу інформаційного простору, забезпеченню прозорості прийняття рішень у завданнях модерації контенту та реалізації принципів справедливості у системах штучного інтелекту.

Дослідження, результати яких викладено в дисертації, виконано в рамках науково-дослідної теми Хмельницького національного університету «Розроблення інформаційної технології прийняття контрольованих людиною критично-безпекових рішень за ментально-формальними моделями машинного навчання» (ДР № 0121U112025). Роль авторки, яка була безпосереднім виконавцем, полягала у розробленні архітектур нейромережевих моделей та інструментарій їхньої інтерпретації.

Формулювання наукової задачі, мети й задач дослідження. Здобувачкою сформульовано науково-прикладну задачу, як забезпечення автоматизованих виявлення та класифікації кіберзалякувань з подальшою інтерпретацією прийнятих рішень, шляхом до розв'язання якої пропонується розробка методів і засобів виявлення та класифікації кіберзалякувань у текстовому контенті, що сприятиме підвищенню точності та якості виявлення кіберзалякувань у текстовому контенті з подальшою інтерпретацією прийнятих рішень.

На початку дослідження, відповідно до науково-прикладної задачі, коректно визначено об'єкт та предмет дослідження. Відтак, об'єктом дослідження виступає процес інтелектуального аналізу текстового контенту для виявлення кіберзалякувань, а предметом дослідження є методи та засоби обробки природної мови для виявлення кіберзалякувань у текстовому контенті. Метою дисертаційного дослідження є підвищення точності та якості виявлення кіберзалякувань у текстовому контенті засобами штучного інтелекту з подальшою інтерпретацією прийнятих рішень

Для досягнення поставленої мети були поставлені та вирішені наступні задачі:

– проведено аналіз методів, засобів та технологій для автоматизованого виявлення кіберзалякувань у текстовому контенті;

– розроблено новий метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості, що забезпечуватиме недискримінацію за віковою, гендерною і релігійною приналежністю;

– розроблено новий метод виявлення кіберзалякувань у текстовому контенті;

– удосконалено метод інтерпретації результатів виявлення кіберзалякувань;

– створено інтелектуальну інформаційну систему для валідації розроблених методів і провести експерименти та порівняння.

Наукова новизна одержаних авторкою результатів полягає у розробленні методів:

1) вперше запропоновано метод оцінювання та коригування репрезентативності датасету за FATE-принципом справедливості, що забезпечує недискримінацію за віковою, гендерною, релігійною приналежністю, що дозволило підвищити якість навчання класифікаторів для виявлення кіберзалякувань;

2) розроблено новий метод виявлення кіберзалякувань у текстовому контенті, який відрізняється від існуючих двоетапним виявленням кіберзалякувань, що полягає у нейромережевій ідентифікації наявності кіберзалякувань та подальшій нейромережевій мультилейбловій класифікації окремих типів кіберзалякувань, що дало можливість підвищити точність та якість виявлення кіберзалякувань;

3) удосконалено метод інтерпретації результатів виявлення кіберзалякувань, який відрізняється від існуючих можливістю надавати візуальні пояснення для мультилейблової класифікації виявлених типів кіберзалякувань в альтернативних поданнях.

Обґрунтованість і достовірність наукових положень, висновків і рекомендацій, які захищаються. Обґрунтованість і достовірність наукових положень, висновків і рекомендацій, які захищаються в дисертації, забезпечуються цілісною методологією дослідження, що базується на системному підході до аналізу предметної області, формалізації задачі, побудові відповідних математичних моделей і застосуванні апробованих інструментів штучного інтелекту та обробки природної мови. Отримані

результати ґрунтуються на чітко визначених критеріях ефективності та верифікуються за допомогою експериментальних досліджень на контрольованих вибірках, що відповідають умовам практичного застосування. Достовірність підтверджується відтворюваністю результатів, проведенням багатократних тестувань із фіксацією статистичних показників, що демонструють стабільну перевагу запропонованих рішень над базовими підходами. Обґрунтування кожного з положень спирається на порівняльний аналіз, результатами якого встановлено істотне поліпшення якості виявлення, класифікації та інтерпретації кіберзалякувань. Методичні рішення узгоджені з сучасними вимогами до побудови етично відповідальних інформаційних систем, а також підтверджені впровадженням в прикладні програмні комплекси, використані у практичних умовах освітніх і правоохоронних організацій, що підтверджує прикладну релевантність і практичну цінність результатів.

Практичне значення одержаних результатів. Практичне значення одержаних результатів полягає у створенні функціонального інструментарію для виявлення та класифікації кіберзалякувань у текстовому контенті, що враховує соціально-етичні вимоги до справедливості й пояснюваності рішень. Запропоновані методи реалізовано у складі інтелектуальної інформаційної системи, яка забезпечує повний цикл обробки текстових повідомлень – від попереднього аналізу датасету з урахуванням показників репрезентативності до інтерпретації результатів класифікації у формі візуальних пояснень. Такий підхід дає змогу суттєво зменшити ризики упередженого аналізу, підвищити точність автоматизованих рішень і забезпечити їхню прозорість для користувачів, зокрема у сферах модерації контенту, освіти, інформаційної безпеки й правозахисної діяльності. Система апробована в умовах реального застосування у взаємодії з організаціями, що працюють із захистом користувачів цифрового середовища, та інтегрована в освітній процес, що підтверджує її практичну доцільність, технологічну реалізованість і здатність до адаптації під конкретні умови використання.

Результати дисертаційної роботи набули впровадження у відділі протидії кіберзлочинам у Хмельницькій області Департаменту кіберполіції Національної поліції України; у ПП «Авіві» (довідка про впровадження); у ГО «ІТ-кластер міста Хмельницького» (довідка про впровадження); у ТОВ «Системи для бізнесу 2» (довідка про впровадження); у навчальному процесі Хмельницького національного університету (акт впровадження).

6. Особистий внесок здобувачки полягає в розробленні нових моделей, методів та засобів, що повністю забезпечують вирішення поставлених у дисертації задач. Здобувачкою реалізовано програмні компоненти інтелектуальної інформаційної системи, проведено серію експериментальних досліджень, здійснено обґрунтування обраних архітектур нейромереж і налаштування алгоритмів для обробки природної мови. Самостійно виконано аналіз отриманих результатів, сформульовано висновки, підготовлено наукові публікації, а також забезпечено впровадження розробленого інструментарію в практичну діяльність установ-учасників апробації. Усі положення, що виносяться на захист, є результатом особистих досліджень здобувачки.

Основні наукові результати дисертації опубліковані у 11-ти наукових працях (4 статті у фахових наукових журналах України, що входили на момент публікації до переліку наукових фахових видань України категорії Б, 5 публікацій у виданнях, що індексуються у міжнародній наукометричній базі даних Scopus та засвідчують апробацію матеріалів дисертації, а також отримано 2 авторські свідоцтва на результати інтелектуальної діяльності). При вивченні рукопису не було встановлено фактів текстових запозичень без відповідних посилань на джерела.

Апробація матеріалів дисертації. Апробацію основних положень, ідей, висновків дисертаційної роботи проведено на: 6th International Conference on Computational Linguistics and Intelligent Systems «CoLInS 2022» (12–13 May, 2022, Gliwice, Poland); 16th International Scientific Conference «Intellectual Systems of Decision-Making and Problems of Computational Intelligence ISDMCI-2022)» (June 14–16, 2022, Rivne, Ukraine); Intelligent Systems Workshop at 8th International Conference on Computational Linguistics and Intelligent Systems «ISW-CoLInS 2024» (April 12–13, 2024, Lviv, Ukraine); 1st International Workshop at Advanced Applied Information Technologies «AdvAIT 2024» (December 5, 2024, Khmelnytskyi, Ukraine, Zilina, Slovakia); 7th Workshop for Young Scientists in Computer Science & Software Engineering «CS&SE@SW 2024» (December 27, 2024, Kryvyi Rih, Ukraine).

Структура та обсяг дисертації. Дисертаційна робота складається з анотації, змісту, переліку умовних скорочень, вступу, чотирьох розділів, висновків, списку використаних джерел із 162 найменувань на 22 сторінках і 4 додатків. Загальний обсяг дисертаційної роботи становить 174 сторінки друкованого тексту, із них 137 сторінок основного тексту. Містить 45 рисунків та 11 таблиць.

Зауваження. В результаті вивчення рукопису мною сформовано наступні зауваження:

1. У розділі 1 с. 21 та с. 24 містять висловлення про «низьку достовірність прийнятих рішень», яке є оціночним і потребує конкретизації (йдеться про точність, узагальнюваність чи стабільність результатів ?).

2. У розділі 4 с. 103-104 містять фрагмент «визначений як такий, що містить в *більшій мірі* вікове та гендерне кіберзалякування. Проте, виведено також і ймовірності наявності інших типів кіберзалякувань, які мають *значно менші значення*.». Твердження про «значно менші значення», або ж «більшій мірі» є неточним – варто наводити конкретні числові пороги або діапазони для інтерпретації результатів.

3. У розділі 4 с. 116-117 містять фрагмент «дані виявились *добре роздільні*, за гендерним аспектом з використанням класифікатора LSTM дані виявились *середньороздільні* та за віковим аспектом з використанням класифікатора SVM – *погано роздільні*.». Дані терміни потребують формального визначення.

4. У пунктах 1.4, 2.1, 2.7.1, 3.1, 3.6, 4.3 фігурує термінологія «вплив окремих слів», одна така термінологія залишається інтуїтивною. Не показано шкали абсолютних чи відносних значень, що унеможливлює порівняння між текстами.

5. Пункт 3.4 «Формування датасетів для навчання та валідування моделей машинного навчання» не є послідовним із назвою наступного пункту 3.5: один називає процеси, інший – структуру моделей. Слід дотримуватися одного рівня абстракції.

Загалом, зазначені зауваження суттєво не впливають на належний рівень і якість рукопису.

Загальний висновок. Вважаю, що дисертаційна робота Собко Олени Віталіївни «Методи виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту» містить нові науково-обґрунтовані теоретичні та експериментальні результати. Усі результати, які виносяться на захист, є достовірними та отримані авторкою особисто.

З огляду вище, вважаю, що дисертаційна робота «Методи виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту», подана на здобуття ступеня доктора філософії, за своїм науковим рівнем та практичною цінністю, змістом та оформленням повністю відповідає вимогам пп. 6, 7, 8, 9 «Порядку присудження ступеня доктора філософії та

скасування рішення разової спеціалізованої вченої ради закладу вищої освіти, наукової установи про присудження ступеня доктора філософії», затвердженого Постановою Кабінету Міністрів України від 12 січня 2022 р. №44 (зі змінами, внесеними згідно з Постановами Кабінету Міністрів України № 341 від 21.03.2022, № 502 від 19.05.2023, № 507 від 03.05.2024), а її авторка, Собко Олена Віталіївна, заслуговує на присудження ступеня доктора філософії за спеціальністю 122 Комп'ютерні науки.

Рецензент
кандидат технічних наук, доцент
кафедри комп'ютерної інженерії
та інформаційних систем
Хмельницького національного університету _____Андрій НІЧЕПОРУК

«Підпис Андрія НІЧЕПОРУКА засвідчую»:
Проректор з наукової роботи
Хмельницького національного університету _____ Олег СИНЮК