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

**ВІДГУК**

офіційного опонента на дисертаційну роботу
Собко Олени Віталіївни
за темою «Методи виявлення та класифікації кіберзалякувань
у текстовому контенті засобами штучного інтелекту»,
подану на здобуття наукового ступеня доктора філософії
за спеціальністю 122 – «Комп'ютерні науки»

**Актуальність теми та зв'язок з науковими планами та програмами.** У сучасному цифровому світі, де обмін текстовою інформацією відбувається миттєво і в глобальних масштабах, кіберзалякування перетворилося на суттєву загрозу як для психологічного здоров'я окремих користувачів, так і для стабільності онлайн-спільнот загалом. Масове розповсюдження меседжів страху, загроз та образ у соціальних мережах, форумах і месенджерах не лише створює неприпустимий емоційний тиск, але й стимулює поширення токсичної культури спілкування, що ускладнює реалізацію освітніх, професійних та соціальних ініціатив в інтернет-просторі. У цьому контексті завдання своєчасного виявлення текстових одиниць із ознаками залякування набуває критичного значення для запобігання ескалації конфліктів і збереження безпеки віртуальних спільнот.

Традиційні алгоритмічні підходи, що базуються на жорстко заданих ключових фразах або лексичних шаблонах, виявилися недостатньо гнучкими перед обличчям адаптивних форм агресивної мови, яка використовує евфемізми, сарказм та контекстуальні натяки. Натомість методи глибокого навчання – зокрема трансформерні архітектури та гібридні нейромережі – демонструють здатність виявляти тонкі семантичні та синтаксичні патерни, що характеризують кіберзалякування. Використання багатовимірних векторних подань слів і речень відкриває можливість формувати семантичні простори, де агресивні та залякувальні конструкції займають окремі кластери, що істотно підвищує точність автоматичної класифікації.

Для українськомовного сегмента Інтернету існує додаткова вимога адаптації моделей до лінгвістичних особливостей мови та регіональних варіацій стилю спілкування. В умовах, коли переважна більшість досліджень зосереджена на англомовних корпусах, розробка і апробація спеціалізованих корпусів текстів із маркуванням кіберзалякувальних патернів та тональних особливостей стають ключовим завданням. Інтеграція механізмів передавального навчання із наперед навченими моделями української мови дозволяє поєднати накопичений знаннєвий фонд зі специфікою локального контексту, підвищуючи адаптивність та стійкість алгоритмів до нових форм загрозливої комунікації. Крім того, використання інструментів пояснювального штучного інтелекту дозволить не лише автоматизувати виявлення повідомлень із залякуваннями, а й забезпечить прозорість рішення для кінцевого користувача чи модератора, що сприятиме підвищенню довіри до системи та формуванню більш відповідального ставлення до спілкування в онлайні. Це відкриває перспективу до створення модулів для модерації контенту, які можуть бути інтегровані в системи управління освітніми платформами, корпоративні рішення з внутрішніх комунікацій та сервіси моніторингу соціальних мереж.

Дисертаційна робота присвячена розробці методів виявлення та класифікації кіберзалякувань й виконана в межах держбюджетної теми Хмельницького національного університету «Розроблення інформаційної технології прийняття контрольованих людиною критично-безпекових рішень за ментально-формальними моделями машинного навчання» (ДР № 0121U112025). Загалом, розробка методів виявлення і класифікації кіберзалякувань із застосуванням сучасних підходів штучного інтелекту є не лише науково обґрунтованим і актуальним, але й має безпосереднє практичне значення для забезпечення інформаційної безпеки, підтримки здорового комунікаційного середовища та розвитку професійних і освітніх онлайн-спільнот.

**Аналіз змісту дисертації та основні результати роботи.** Дисертаційна робота складається з анотації, змісту, переліку умовних скорочень, вступу, основної частини, 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. Дисертація містить 45 рисунків та 11 таблиць. Дисертація Собко О.В. структурована логічно, послідовно і складається з чотирьох основних розділів, у яких розкрито теоретичні засади, розроблено методичні рішення та проведено

експериментальне дослідження поставленої науково-прикладної задачі – виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту з урахуванням етичних аспектів та вимог пояснювальності.

У ***першому розділі*** здійснено комплексний огляд наукових джерел, присвячених тематиці виявлення агресивного контенту в мережі, зокрема з використанням методів обробки природної мови та глибокого навчання. Обґрунтовано вибір предметної області, виявлено існуючі проблеми, пов'язані з нерепрезентативністю наборів даних для навчання нейромереж, низькою точністю класифікації та відсутністю пояснювальності моделей. Результати огляду стали підґрунтям для формулювання мети та завдань дослідження.

У ***другому розділі*** наведено розроблену авторкою методику виявлення та класифікації кіберзалякувань. Зокрема, розроблено метод коригування репрезентативності текстових вибірок за принципами справедливості (FATE), що дозволяє уникнути дискримінації за віковими, гендерними, релігійними ознаками під час навчання моделей. Запропоновано рішення у вигляді двоетапної нейромережевої системи, яка спочатку ідентифікує наявність кіберзалякування, а потім класифікує його за типами за допомогою мультилейблового підходу. Висвітлено також розроблений авторкою метод інтерпретації рішень, що базується на візуалізації семантичних вагових ознак у моделі.

***Третій розділ*** присвячений побудові інтелектуальної інформаційної системи, яка реалізує запропоновані підходи. Наведено архітектуру системи, описано її модулі, засоби збору та обробки даних, реалізацію моделей машинного навчання, а також способи інтеграції з інтерфейсом користувача.

У ***четвертому розділі*** проведено ґрунтовне експериментальне дослідження ефективності запропонованих методів. Здійснено оцінювання моделей на реальних та штучно згенерованих вибірках, продемонстровано підвищення точності класифікації в порівнянні з базовими моделями, зокрема на понад 12 % за метрикою $F_1$-score. Засвідчено ефективність інтерпретаційних механізмів за участі експертної оцінки. Наведено довідки про впровадження результатів у практичну діяльність, зокрема в освітньому та комерційному середовищі.

Таким чином, аналіз змісту дисертації дозволяє зробити висновок, що здобувачка комплексно і на високому науковому рівні реалізувала всі поставлені задачі дослідження. Структура дисертації повністю відповідає логіці й послідовності рішення поставлених задач. Основні результати дисертації є логічно пов'язаними, належно верифікованими, мають значний науково-

прикладний потенціал та підтверджені апробацією на конференціях і публікаціями у фахових виданнях.

**Основні результати роботи:**

– 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 проведено експерименти та дослідження.

**Наукова новизна, оцінка обґрунтованості наукових положень дисертації та їх достовірності.**

Основні наукові положення, результати та висновки дисертації отримані здобувачкою самостійно, є новими, достатньо обґрунтованими та підтверджуються даними комп'ютерних експериментів та апробацією основних положень на міжнародних конференціях, а також впровадженням у діяльність підприємств та освітній процес. Достовірність наукових положень, висновків і результатів, отриманих здобувачем, обумовлена коректними та доцільним використанням математичного апарату, методології проектування інформаційних систем, успішною програмною реалізацією розроблених методів виявлення та класифікації кіберзалякувань у текстовому контенті. Отримані в дисертаційній роботі наступні результати, які мають наукову новизну:

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.

**Теоретичне та практичне значення одержаних результатів.** Теоретичне значення одержаних результатів дисертаційного дослідження полягає в розробці підходу до виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту з подальшою інтерпретацією прийнятих рішень. Особливу вагу має інклюзивний підхід авторки, яка інтегрує в розробку принципи соціальної відповідальності та етичної нейтральності штучних моделей, застосуючи концепцію FATE до мовних моделей. В результаті досягається мета дисертаційного дослідження – підвищення точності та якості виявлення кіберзалякувань у текстовому контенті засобами штучного інтелекту з подальшою інтерпретацією прийнятих рішень.

Практичне значення одержаних результатів дисертаційного дослідження полягає у доведенні теоретичних результатів роботи та розробці інтелектуальної інформаційної системи виявлення та класифікації кіберзалякувань у текстовому контенті, що надає можливість оцінювати та коригувати репрезентативність наборів даних для навчання моделей машинного навчання за етичними аспектами FATE-принципу справедливості; виявляти та класифікувати типи кіберзалякувань у текстовому контенті. Також інтелектуальна інформаційна система дозволяє отримувати візуальні пояснення для мультилейблової класифікації виявлених типів кіберзалякувань, що сприяє підвищенню довіри до одержаних результатів їх класифікації.

Для аналізу репрезентативності за етичними аспектами використано моделі машинного та глибокого навчання, що показали кращі показники точності під час експериментального дослідження. Підвищення якості навчання класифікаторів для виявлення кіберзалякувань полягає у формуванні репрезентативного та недискримінаційного за FATE-принципом справедливості набору даних для навчання нейромереж.

Підвищення якості виявлення кіберзалякувань полягає у застосуванні двоетапної перевірки – спочатку виявлення кіберзалякування, а потім визначення наявних типів кіберзалякувань у текстовому контенті. За результатами досліджень, для випадку бінарної класифікації кращі результати показала нейромережева модель BiLSTM з показниками Accuracy – 94 %, Precision – 94 %, Recall – 94 %, $F_1$-міри – 94 %; для випадку мультилейблової класифікації кращі результати показала нейромережева модель BERT з

показниками макрометрик Accuracy – 94 %, Precision – 93 %, Recall – 93 %, $F_1$-міри – 93 %. Виявлення кіберзалякувань запропонованим у дослідженні методом має показник Accuracy щонайменше на 0,8 % вищий, ніж у відомих підходах. Відтак, підвищення якості виявлення кіберзалякувань здійснюється шляхом використання мультилейблового класифікатора для класифікації типів кіберзалякувань на основі нейромережевої архітектури трансформер та інтерпретаційної моделі. Забезпечується візуальна інтерпретація результатів у вигляді колірного представлення ваг слів, що вплинули найбільше на рішення моделі, а також у вигляді діаграм впливу окремих слів тексту на ймовірність віднесення цього тексту до конкретного типу кіберзалякування та середнього значення важливості ключових слів для всіх класів.

Авторкою одержано довідки про впровадження результатів дисертаційної роботи впроваджено у відділі протидії кіберзлочинам у Хмельницькій області Департаменту кіберполіції Національної поліції України, у ПП «Авіві», у ГО «ІТ-кластер міста Хмельницького», у ТОВ «Системи для бізнесу 2»; також одержано акт впровадження результатів роботи у навчальному процесі Хмельницького національного університету.

Представлена в роботі сукупність методів, направлених на виявлення та класифікації кіберзалякувань у текстовому контенті, є важливим теоретичним і практичним внеском.

**Повнота викладу результатів дисертації в опублікованих працях.** Основні результати дисертації опубліковані у 11 наукових працях, серед яких: 4 статті у фахових наукових журналах України, включених на дату опублікування до переліку наукових фахових видань України категорії Б; 5 публікацій, які засвідчують апробацію матеріалів дисертації (публікації, що індексуються в наукометричній базі Scopus); 2 авторських свідоцтва.

Основні результати цілком висвітлені в 4 наукових статтях у фахових виданнях України. Одна стаття виконана у співавторстві (два співавтори), три статті одноосібні, тому згідно Підпункту 1 пункту 8 в редакції Постанови КМ № 507 від 03.05.2024, ці статті зараховуються повністю. Таким чином, авторка має 4 публікації, у яких викладено основні результати дисертації, чого достатньо згідно чинних вимог.

Положення роботи апробовано на п’яти міжнародних конференціях, за їх результатами виконано п’ять публікацій, що індексуються в наукометричній базі «Scopus». В роботі чітко зазначено особистий внесок дисертанта у кожній спільній публікації. Одержані здобувачкою два авторських свідоцтва на комп’ютерну програму є одноосібними.

**Зауваження та побажання.**

1. В п.2.5 розглядається використання FATE-принципу справедливості для оцінювання та коригування репрезентативності наборів даних для навчання нейромереж, проте не приділено увагу іншим FATE-принципам. Для підвищення академічного рівня дослідження, було б корисним розглянути запропонований підхід у розрізі всіх FATE-принципів.

2. Запропонований в п.2.6 метод виявлення кіберзалякувань в межах роботи сфокусований на виявленні вікових, гендерних, релігійних та етнічних кіберзалякувань, не розглядаючи як окремі класи інші типи кіберзалякувань. Було б корисним дослідити роботу метода для виявлення інших актуальних типів кіберзалякувань (шеймінг, тролінг, мобінг тощо).

3. У п. 3.5 (стор. 91) лише перелічено використані архітектури (LSTM, BERT тощо), без достатнього обґрунтування їх вибору саме для задачі кіберзалякувань.

4. Формули для оцінки моделей (точність, влучність тощо), наведені в п.3.7 (стор. 110-111) є відомими. Тому їх опис доцільніше винести у Розділ 1 або навести посилання на джерела цих формул.

5. В роботі виявлено ряд технічних помилок при наборі тексту. Наприклад, на стор. 84 наприкінці 3-го абзацу присутні 2 варіанта закриваючих лапок поряд.

6. На рис. 3.2 (стор. 83) наведено англомовні назви підсистем, хоча ці назви не є назвами програмних компонентів. Більш коректно ці назви подавати українською мовою.

7. В дисертації п.1.5 «Висновки. Постановка задачі» є агрегованим. Відповідно до його вмісту, для покращення сприйняття та структурування контенту, було доречним розділити його окремо на пункти «Висновки до розділу 1» та «Постановка задачі».

8. У дисертації надається перевага нейромережевим методам класифікації кіберзалякувань, тоді як альтернативні підходи (наприклад, методи на основі лінгвістичного аналізу) згадані побіжно. В п.1.3 варто було навести критерії порівняння можливих варіантів методів класифікації та об'єктивне обґрунтування вибору.

Однак наведені вище зауваження не мають принципового значення та не зменшують наукової цінності дисертаційної роботи в цілому.

**Загальний висновок.** Викладені вище міркування дають можливість стверджувати, що дисертаційна робота Собко О. В. є завершеним, цілісним, самостійно виконаним науковим дослідженням, містить елементи наукової

новизни і важливі практичні результати, які є суттєвими при вирішенні важливої науково-прикладної задачі – виявленні та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту. Розроблені методи і засоби сприяють підвищенню точності та якості виявлення кіберзалякувань у текстовому контенті з подальшою інтерпретацією прийнятих рішень.

За актуальністю розглянутих завдань, обсягом досліджень, науковим рівнем, системністю досліджень, розробленим програмним забезпеченням і практичною цінністю отриманих результатів робота відповідає рівню дисертацій на здобуття ступеня доктора філософії. Обрану тему дисертації належним чином розкрито, мету досягнуто, завдання виконані. Тема і зміст дисертації відповідають спеціальності 122 – «Комп'ютерні науки» в галузі знань 12 – «Інформаційні технології».

Актуальність розглянутих завдань, а також науковий рівень, новизна та практична цінність проведених досліджень дають право вважати, що дисертаційна робота «Методи виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту» відповідає як обраній спеціальності, так і встановленим вимогам пунктів 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 присудження ступеня доктора філософії за спеціальністю 122 – «Комп'ютерні науки».

Офіційний опонент,
доктор технічних наук, доцент,
доцент кафедри
інформаційно-обчислювальних
систем і управління
Західноукраїнського
національного університету _____ Христина ЛІП'ЯНІНА-ГОНЧАРЕНКО