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. Актуальність дослідження, присвяченого методам виявлення та класифікації кіберзалякувань у текстовому контенті із застосуванням засобів штучного інтелекту, обумовлена низкою ключових трансформацій у цифровому суспільстві, що відбуваються впродовж останніх десятиліть. Інтенсивна діджиталізація усіх сфер життєдіяльності, зокрема освіти, соціальних комунікацій, бізнесу та державного управління, супроводжується появою нових соціальних ризиків, серед яких особливе місце займає феномен кібербулінгу. Ця форма насильства має здатність до швидкого поширення, маскування та ескалації, що ускладнює її виявлення традиційними методами моніторингу та контролю.

Штучний інтелект як галузь інформатики пропонує новий рівень інструментарію для автоматизованого аналізу великих обсягів текстової інформації. Особливої актуальності набуває застосування моделей глибокого навчання та трансформерних архітектур, здатних виявляти латентні патерни агресивної поведінки в мовленнєвих конструкціях, враховуючи контекст, семантичну гнучкість та іронічну двозначність, що часто присутні в повідомленнях кіберзалякування. Таким чином, на перетині штучного інтелекту, когнітивної лінгвістики та соціальних наук формується нове міждисциплінарне поле дослідження, орієнтоване на вирішення актуальних проблем цифрової безпеки.

Сучасні інформаційні системи потребують інтеграції інтелектуальних механізмів, які здатні не лише виявляти факт агресії, а й класифікувати її за характером спрямованості. Це відкриває перспективи для створення динамічних моделей, що адаптуються до мовних і культурних особливостей користувацького середовища. Еволюція таких систем має критичне значення для

формування цифрових сервісів нового покоління, орієнтованих на захист психологічного добробуту користувачів.

Дисертаційна робота Собко О.В. 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). Робота присвячена підвищенню точності та якості виявлення кіберзалякувань у текстовому контенті засобами штучного інтелекту з подальшою інтерпретацією прийнятих рішень. Таким чином, запропонована тематика є не лише актуальною, а й стратегічно важливою для розвитку інтелектуальних технологій у глобальному вимірі.

Аналіз змісту дисертації та основні результати роботи.

Дисертаційна робота присвячена методам виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту є ґрунтовним і актуальним дослідженням, що відповідає сучасним викликам цифрової безпеки. Авторка продемонструвала високий рівень наукової зрілості, поєднавши теоретичні засади з практичними аспектами застосування інтелектуальних технологій для виявлення агресивної поведінки в онлайн-середовищі.

У роботі запропоновано комплексний підхід, який включає оцінювання та коригування репрезентативності навчального набору даних за FATE-принципом справедливості, нейромережеве виявлення та класифікацію кіберзалякувань за різними типами (віковими, релігійними, етнічними, гендерними тощо), а також візуальну інтерпретацію результатів, отриманих нейромережевими моделями. Такий підхід дозволяє забезпечити неупередженість та відповідність етичним вимогам, а також надає пояснення рішень моделі щодо кожного виявленого типу кіберзалякування, що підвищує прозорість і довіру до систем штучного інтелекту.

Особливу увагу заслуговує розроблена інтелектуальна інформаційна система, яка реалізує запропоновані методи та демонструє високу ефективність у виявленні та класифікації кіберзалякувань. Зокрема, точність не нижче 94% для моделей BiLSTM і BERT підтверджує практичну значущість дослідження та його потенціал для впровадження в реальні системи моніторингу та модерації контенту.

Загалом, дисертаційна робота О.В. Собко є вагомим внеском у розвиток методів обробки природної мови та штучного інтелекту в контексті забезпечення інформаційної безпеки. Результати дослідження мають як теоретичну, так і практичну цінність, що підтверджує актуальність та наукову новизну виконаної роботи.

Основні результати роботи:

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

2) розроблено новий метод оцінювання та коригування репрезентативності навчального набору даних за FATE-принципом справедливості, що забезпечуватиме недискримінацію за віковою, гендерною і релігійною приналежністю;

3) розроблено новий метод виявлення кіберзалякувань у текстовому контенті;

4) удосконалено метод інтерпретації результатів виявлення кіберзалякувань;

5) створено інтелектуальну інформаційну систему для валідації розроблених методів і проведення експериментів та порівнянь.

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

Авторка претендує на наукову новизну наступних отриманих результатів:

– 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;

– 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;

– 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 результатів, висновків та рекомендацій, наведених у дисертації, досягається проведеним аналізом існуючих результатів у сфері наукових досліджень. Коректне використання методів дослідження, математичного апарату й програмних результатів підтверджується отриманими результатами числових експериментів.

Теоретичне та практичне значення одержаних результатів.

Теоретичне значення результатів дисертаційного дослідження полягає в розробці концептуальної моделі системи виявлення та класифікації кіберзалякувань, яка базується на застосуванні сучасних алгоритмів штучного інтелекту. Важливим є те, що авторка враховує аспекти соціальної

відповідальності та етичної нейтральності моделей, що виводить дослідження за межі суто інженерної проблематики та інтегрує його в ширший науковий контекст. Наукова новизна також виявляється в застосуванні принципів FATE до лінгвістичних моделей, що дозволяє зменшити упередженість систем автоматизованого аналізу текстів.

З практичного погляду, результати дисертації мають значний потенціал до впровадження в системи автоматизованої модерації контенту соціальних платформ, освітніх онлайн-середовищ, корпоративних каналів внутрішньої комунікації. Запропонована інтелектуальна інформаційна система не лише демонструє високі показники точності, але й забезпечує інтерпретованість рішень, що є важливим чинником для її реального використання в умовах правової та етичної відповідальності. Практичні напрацювання, викладені у дисертації, можуть бути адаптовані для розробки освітніх курсів, зокрема в підготовці фахівців з інформаційної безпеки, цифрової етики та комп'ютерної лінгвістики.

У цілому, як теоретичні напрацювання, так і практична реалізація засвідчують високий рівень професійної зрілості дисертантки, а також значний внесок в область комп'ютерних наук і суміжні галузі знань.

Повнота викладу результатів дисертації в опублікованих працях. Основні результати дисертації достатньо повністю наведено у 4-х наукових статтях у виданнях, включених до Переліку наукових фахових видань України, які, на дату опублікування, віднесені до категорії «Б».

Три статті одноосібні, тому згідно Підпункту 1 пункту 8 в редакції Постанови КМ № 507 від 03.05.2024, зараховуються повністю. Одна стаття має 2 (два) співавтори, тому згідно наведеної вище постанови, також зараховується повністю.

Виходячи з наведеного вище, здобувачка має 4 публікації, у яких викладені основні результати дисертації, чого достатньо, згідно чинних вимог. Зазначу, що для кожної публікації, де авторка мала співавторів, чітко вказаний особистий внесок.

Додатково положення роботи апробовано на 5 Міжнародних конференціях, за їх результатами виконано 5 (п'ять) публікацій, що індексуються в міжнародній наукометричній базі «Scopus». У кожній спільній публікації чітко зазначено особистий внесок дисертантки, а загальна кількість та рівень джерел повністю відповідають вимогам до дисертаційних робіт.

Зауваження та побажання.

1. Запропоновані в розділі 2 методи призначені для роботи з україномовним контентом, що обмежує їх застосування в багатомовних або міжнародних проєктах. Варто навести підхід до адаптації методології для роботи з багатомовними текстами.

2. У роботі для інтерпретації результатів виявлення кіберзалякувань (п. 3.3) та під час їх експериментального дослідження (п. 4.3) використовується виключно метод LIME. Проте в 1 розділі (п. 1.4) не обґрунтовано вибір жодної конкретної моделі для подальшого застосування, а опис процедури інтерпретації (п. 2.7) також не обґрунтовує пріоритет LIME над іншими підходами. Доцільно або розширити огляд і розглянути альтернативні техніки, наприклад SHAP, або чітко аргументувати, чому саме LIME є оптимальним вибором для даного завдання.

3. Категорія класифікації розглядається без визначення семантичних характеристик. На наш погляд використання онтологічної метрики при визначенні термінологічних класів більш сприяло б візуалізації процесів класифікації.

4. Хоча подано довідки про впровадження, у тексті дисертації варто навести аналіз того, як система інтегрувалась у реальні процеси та якими були результати цього впровадження (кількісні та якісні).

5. При викладені підходів до інтерпретації результатів (стор. 33-36, п.1.4) бракує структурованості – огляд подано списком без табличного узагальнення характеристик методів (наприклад, LIME, SHAP, XAI), що ускладнює порівняння.

6. Деякі рисунки мають низьку роздільну здатність. Наприклад, рис. 2.8 (стор. 74) варто було б навести з більшим масштабом та контрастністю.

7. В огляді відомих засобів формування репрезентативних датасетів (п.1.2), виявлення та класифікації кіберзалякувань (п.1.3) й інтерпретації результатів (п.1.4) було розглянуто алгоритми та моделі, але недостатньо приділено увагу аналізу метрик їх ефективності.

8. У тексті дисертації зустрічаються не зовсім коректні трактування деяких категорій та понять, що досліджуються. Так фраза «…поняття репрезентативності, справедливості та недискримінаційності є важливими у створенні етичних і справедливих моделей машинного навчання», але більш коректне буде «…поняття репрезентативності, справедливості та недискримінаційності є важливими у створенні етичних і справедливих мовних моделей на засадах машинного навчання…» (стор.24).

Однак вказані зауваження не є суттєвими щодо проведеного дисертанткою дослідження, й не зменшують високій рівень та якість отриманих наукових результатів, істотно не впливають на загальну позитивну оцінку дисертаційної роботи, й не знижують її наукову й практичну цінність.

Загальний висновок.

Подана дисертаційна робота є завершеною науково-дослідною роботою, у якій розроблено системний підхід до виявлення та класифікації кіберзалякувань у текстовому контенті засобами штучного інтелекту. Вирішується суперечність між можливістю точного виявлення кіберзалякувань у текстовому контенті та

відсутністю довіри до навчальних даних, які не можуть гарантувати репрезентативність результатів через відсутність перевірки та можливостей приведення навчальних даних до репрезентативного вигляду. Розроблені методи і засоби сприятимуть підвищенню точності автоматизованого виявлення кіберзалякувань у текстовому контенті. Тема і зміст дисертації відповідає Стандарту спеціальності 122 – «Комп'ютерні науки» для третього рівня вищої освіти.

Оцінюючи дисертаційну роботу в цілому, є всі підстави стверджувати, що за актуальністю теми, обсягом виконаних досліджень, науковою новизною, цінністю одержаних результатів і науково-теоретичним рівнем їх обґрунтованості робота цілком відповідає 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 на присудження їй ступеня доктора філософії за спеціальністю 122 – «Комп'ютерні науки» в галузі знань 12 – «Інформаційні технології».

Офіційний опонент
доктор технічних наук, професор
головний науковий співробітник
Центрального науково-дослідного
інституту озброєння та військової техніки
Збройних Сил України Олександр СТРИЖАК

Начальник відділу
персоналу та стройового
ЦНДІ ОВТ ЗС України
Є. НОВОЖЕНІН