

## **ВІДГУК**

офіційного опонента професора кафедри інформаційних технологій проектування та прикладної математики Київського національного університету будівництва і архітектури,  
доктора технічних наук, професора **Терейковської Людмили Олексіївни** на дисертаційну роботу **Никоненка Андрія Олександровича** на тему «Гібридний метод ідентифікації автоматично згенерованих природномовних текстів», що представлена на здобуття наукового ступеня доктора філософії з галузі знань 12 «Інформаційні технології» за спеціальністю 122 «Комп'ютерні науки»

### **Ступінь актуальності обраної теми**

Проблема автоматизованого виявлення згенерованого контенту та протидії соціотехнічним атакам залишається у фокусі уваги світової наукової спільноти та інституцій у сфері кібербезпеки. Сучасні технології розвиваються надзвичайно швидко, і текстові маніпуляції та фішинг продовжують залишатися одними з домінуючих векторів кібератак в умовах активного застосування методів соціальної інженерії та психоемоційного впливу. Докорінна трансформація вітчизняного та глобального інформаційного простору, зумовлена розповсюдженням великих мовних моделей, або діалогових інтелектуальних помічників (ДІП), значно ускладнює своєчасне виявлення загроз традиційними сигнатурними та статистичними методами.

Класичні методи детекції, які орієнтуються на ізольовані атрибути, є малоефективними проти текстів, що імітують когнітивний стиль людини та генеруються зі специфічними цільовими предикатами. Розроблена в дисертації Никоненка А. О. інформаційна технологія розв'язує цю актуальну науково-прикладну проблему шляхом формалізованого подання багатовимірного сліду штучного інтелекту (ШІ), що дозволяє виявляти маніпулятивний вплив і автоматично згенерований контент навіть за умови цілеспрямованих трансформаційних атак та маскування.

### **Ступінь обґрунтованості наукових положень, висновків і рекомендацій, сформульованих у дисертації**

Наукові положення, висновки та рекомендації, представлені у дисертаційній роботі, є глибоко обґрунтованими та спираються на сучасні досягнення нейромережових технологій і методів машинного навчання. Автором проведено ґрунтовний аналіз вразливостей існуючих засобів захисту інформації до методів маніпулятивного впливу та семантичного маскування.

Запропонована концепція багатовимірного аналізу текстів дозволяє уникнути ручного формування шаблонів і забезпечує автоматичне виділення складних

залежностей у текстових даних. Завдяки об'єднанню лексичної, синтаксичної та семантичної модальностей у єдиний простір ознак, каскадний ансамбль здатний враховувати глибинну семантику повідомлення та ефективно протидіяти загрозам вітчизняного кіберпростору. Достовірність результатів переконливо підтверджується проведеними експериментальними дослідженнями на базі репрезентативних вибірок, які включають тексти від сучасних діалогових інтелектуальних помічників та спеціалізованих моделей-парафразерів.

## **Новизна наукових положень, висновків і рекомендацій, сформульованих у дисертації**

Наукова новизна дисертаційної роботи полягає у розробці та теоретико-методичному обґрунтуванні моделей і гібридного методу ідентифікації автоматично згенерованих природномовних текстів, що підвищує стійкість детекції до трансформаційних атак та забезпечує інтерпретованість прийнятих рішень в умовах невизначеності. Отримано такі нові наукові результати:

1. Вперше розроблено гібридний метод ідентифікації автоматично згенерованих текстів, який, завдяки застосуванню каскадного ансамблю з механізмами адаптивного зважування та аналізу міжрівневої узгодженості ознак, дає змогу виявляти дисонанс між лексичним, синтаксичним та семантичним рівнями тексту, забезпечуючи стійкість до процесів маскування текстових даних, що реалізуються комерційними системами.
2. Вперше розроблено інтерпретаційну модель обґрунтування доказів детекції, яка використовує внутрішні ваги метакласифікатора для прямої генерації глобального та локальних профілів атрибуції та карт доказів, що дає змогу знизити часову складність побудови пояснень до  $O(1)$ .
3. Одержав подальший розвиток метод забезпечення робастності моделей обробки природної мови в умовах агресивного маскування та епістемічної невизначеності шляхом інтеграції процедури індуктивного конформного передбачення.
4. Удосконалено концептуальну модель оцінки робастності систем детекції шляхом введення формалізованого поняття «Точки зламу», яке встановлює математичний взаємозв'язок між інтенсивністю структурних змін тексту, збереженням його семантичної цілісності та ймовірністю виявлення.

## **Оцінка наукових публікацій здобувача**

Проведений аналіз публікаційної активності здобувача показав, що основні результати дисертації повною мірою висвітлені у наукових працях, які відповідають вимогам до опублікування результатів. Зокрема опубліковано 8 наукових праць, з

них: 1 наукова стаття у виданнях, що індексуються у міжнародних базах Scopus; 3 наукові статті у фахових виданнях України; 4 тези доповідей на міжнародних наукових конференціях. Означені публікації відповідають тематиці дослідження та відображають його основні результати.

## **Відсутність порушення академічної доброчесності**

Результати перевірки дисертаційної роботи за допомогою сервісів Turnitin, аналізу наукових публікацій здобувача, а також дослідження тексту дисертації й використаних джерел засвідчують відсутність порушень принципів академічної доброчесності з боку автора.

## **Структура та зміст дисертаційного дослідження**

Дисертаційна робота складається з анотації, вступу, чотирьох розділів, загальних висновків, списку використаних джерел та п'яти додатків. Загальний обсяг роботи становить 248 сторінок, з яких основний текст займає 188 сторінок. Робота містить 45 рисунків та 40 таблиць в основному тексті. Матеріал викладено з дотриманням принципів системного підходу, логічної послідовності та коректного використання математичного апарату.

У **вступі** обґрунтовано актуальність ідентифікації згенерованого контенту як засобу протидії маніпулюванню суспільною думкою та академічній недоброчесності. Визначено об'єкт, предмет, мету та завдання дослідження.

**Перший розділ** присвячено комплексному системному аналізу поточного стану предметної області. Автором детально досліджено еволюцію архітектур великих мовних моделей, зокрема Mixture-of-Experts та міркуючих агентів. Проведено критичний аналіз існуючих статистичних та нейромережевих підходів до детекції згенерованих текстів, виявлено їхні ключові обмеження та вразливості до трансформаційних атак, що дозволило чітко сформулювати та обґрунтувати задачі поточного дослідження.

У **другому розділі** сформовано фундаментальний теоретико-математичний базис розробленої інформаційної технології. Автором формалізовано багатовимірну математичну модель сліду штучного інтелекту, що враховує лексичний, синтаксичний та семантичний рівні. Розроблено концептуальну модель «Точки зламу» для прогнозування стійкості детекції в умовах навмисного маскуванню тексту. Також наведено алгоритмічний опис архітектури гібридного каскадного ансамблю з інтегрованим механізмом адаптивного зважування та запропоновано інтерпретаційну модель пояснюваного ШІ.

**Третій розділ** зосереджено на експериментальній валідації розроблених неймережових та математичних моделей методами чисельного моделювання і комп'ютерного експерименту. Варто відзначити ґрунтовний підхід автора до формування репрезентативних тестових вибірок, що включають тексти сучасних архітектур MoE та Reasoning, а також синтез трансформаційних атак за допомогою парафразерів T5 та DIPPER. Експериментально доведено факт зміщення «Точки зламу» для гібридного методу та спростовано низку гіпотез щодо виявлення мікроструктурних архітектурних артефактів у згенерованих текстах.

У **четвертому розділі** висвітлено питання практичної реалізації програмного комплексу та об'єктивної оцінки його ефективності у складних умовах конкуренції. Детально описано інтеграцію фреймворку індуктивного конформного передбачення для розв'язання проблеми епістемічної невизначеності під час аналізу даних поза навчальним розподілом. Проведено аналіз обчислювальної складності та часових витрат системи. Продемонстровано результати роботи інтерпретаційної моделі ХАІ на конкретних прикладних кейсах із генерацією інтерактивних «Карт доказів».

## **Практична цінність результатів роботи**

Практичне значення одержаних результатів полягає у створенні готового до промислового використання інструментарію ідентифікації автоматично згенерованого контенту, який може бути ефективно інтегрований у системи контент-модерації, інструменти забезпечення інформаційної безпеки та платформи перевірки академічної доброчесності.

Розгортання системи реалізовано з використанням хмарної інфраструктури Amazon Web Services, що забезпечує високу відмовостійкість та можливість паралельної обробки значних обсягів запитів. Завдяки раціональній оптимізації алгоритмів, загальна асимптотична складність функціонування гібридного методу зведена до  $O(N^2)$ . Середній час обробки одного документа у хмарному середовищі AWS становить 0.33 секунди, що гарантує придатність розробленої технології до повноцінного використання в режимі реального часу. Експериментально доведено беззаперечну перевагу розробленого гібридного методу у протидії сучасним комерційним засобам маскування тексту, таким як StealthGPT, AIHumanize та Phrasly. В умовах активного маскування система забезпечила найнижчий серед досліджуваних світових аналогів рівень успішності атак на рівні 22.84%, стабільно зберігаючи середній показник повноти виявлення на рівні 77.16%. Використання процедури індуктивного конформного передбачення (ICP) є вкрай важливим для практичного застосування системи, оскільки дає змогу здійснювати кількісну оцінку епістемічної невизначеності та ефективно розпізнавати дані поза навчальним

розподілом. Під час тестування на складних текстах неносіїв мови впровадження процедури ICP дозволило суттєво знизити рівень хибнопозитивних спрацювань з 4.81% до 0.98%. Коефіцієнт порятунку авторів-людей становив 79.71%, що радикально мінімізує ризики автоматичної дискримінації та несправедливих звинувачень. Інтеграція підсистеми пояснювального штучного інтелекту (XAI), що автоматично генерує глобальні та локальні профілі атрибуції ознак, а також інтерактивні «Карти доказів», надає експертам вичерпну та прозору аналітичну базу для ухвалення остаточних рішень. Це дозволяє успішно перетворити архітектуру системи з «чорної скриньки» на прозорий інтерпретований інструмент аудиту ознак, що є критично необхідним для використання технології у спірних академічних та інституційних ситуаціях.

### **Дискусійні положення, зауваження та рекомендації**

Позитивно оцінюючи теоретичну та прикладну значущість дисертаційної роботи Никоненка А. О., вважаю за доцільне висловити такі зауваження та рекомендації:

1. У підрозділі 3.1 роботи (с. 125-136) ґрунтовно описано протидію комерційним системам маскуванню тексту. Проте додаткового висвітлення потребує питання ефективності гібридного методу у виявленні персоналізованих фішингових електронних листів та маніпулятивних повідомлень в каналах електронної комунікації, де зловмисники генерують контент через специфічний *prompt engineering* задля імітації емоційно-експресивного стилю та обходу спам-фільтрів. Доцільно було б конкретизувати цей аспект під час опису формування тестових наборів даних.
2. Дискусійним залишається питання стійкості підсистеми індуктивного конформного передбачення, що розглядається у підрозділі 4.2.1 (с. 188-193), до атак типу «отруєння даних». Робота значно виграла б, якби автор проаналізував ризики масової інжекції синтетичних текстів зі специфічним маскувальним патерном у калібрувальну вибірку та їхній вплив на зсув порогової зони системи.
3. У підрозділі 2.2 (с. 88-90) в описі семантичного рівня недостатньо повно розкрито особливості застосування гібридного методу в умовах реалізації атак стилістичної імітації, коли велика мовна модель практично повністю відтворює синтаксис та специфічний словниковий запас автентичного тексту. Потребує додаткового обґрунтування теза про те, що сили сигналу дисонансу  $S_{mismatch}$  буде достатньо для виявлення такої глибокої семантичної мімікрії.
4. Як зазначається у підрозділі 4.4.1 (зокрема, на рис. 4.13, с. 206), розроблено глобальний та локальний профілі атрибуції ознак. Оскільки діалогові

інтелектуальні помічники генерують контент у прямій залежності від структури ініціалізуючої та уточнюючої частин запиту, здобувачу варто було б детальніше пояснити вплив специфіки цих цільових предикатів на розподіл вагових коефіцієнтів ознак у глобальному профілі моделі-детектора.

5. У процесі формування багатовимірному простору ознак (підрозділ 2.3, с. 96-109) виникає проблематика обробки текстового шуму, наприклад, артефактів парсингу, спецсимволів або вузькоспеціалізованої термінології. Метод можна було б значно покращити, застосувавши підходи на основі алгоритмів вибору ознак або семантичної сегментації, що дозволило б відфільтрувати надлишкову інформацію та підвищити чутливість системи до складних маніпулятивних загроз.

Висловлені зауваження мають переважно дискусійний характер і не знижують загальної наукової цінності та практичної значущості виконаного дослідження.

### **Загальний висновок**

Дисертаційна робота Никоненка Андрія Олександровича є завершеною, цілісною та самостійно виконаною науково-дослідною працею, яка має вагоме теоретичне та прикладне значення для розвитку сфери кібербезпеки та комп'ютерних наук. За рівнем наукової новизни, глибиною проведеного аналізу предметної області та практичною цінністю впроваджених результатів, робота повністю відповідає всім критеріям та вимогам «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 року № 44.

Здобувач Никоненко Андрій Олександрович заслуговує на присудження наукового ступеня доктора філософії (PhD) за спеціальністю 122 «Комп'ютерні науки» галузі знань 12 «Інформаційні технології».

### **Офіційний опонент,**

доктор технічних наук, професор,

професор кафедри інформаційних технологій проєктування

та прикладної математики

Київського національного університету

будівництва і архітектури



Л.О. Терейковська