

ВІДГУК ОФІЦІЙНОГО ОПОНЕНТА

кандидата економічних наук, доцента **Бойко Наталії Іванівни** 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

Докорінна трансформація глобального інформаційного простору пов'язана зі швидким розвитком сучасних технологій продукує зростання обсягів та гетерогенності неструктурованих текстових даних. Згенеровані тексти часто різняться за форматом, деталізованістю, структурою і семантичною глибиною, що створює труднощі під час розробки уніфікованих підходів для їхнього багатовимірного аналізу. У реальних завданнях інформаційної безпеки та академічної доброчесності важливо враховувати одразу кілька вимог: швидкість, точність, стійкість до шуму та маніпуляцій і обчислювальну ефективність.

Класичні методи детекції, які орієнтуються лише на один критерій або один вимір, наприклад, суто лексичний, виявляються малоефективними у таких випадках. Відсутність комплексних моделей, методів і засобів підтримки прийняття рішень на основі багатовимірного аналізу лінгвістичних даних, які б одночасно враховували різноманітність рівнів тексту та багатокритеріальний характер класифікації, створює актуальну науково-прикладну проблему. Розроблена в дисертації Никоненка А. О. інформаційна технологія розв'язує цю проблему шляхом формалізованого подання багатовимірного сліду ШІ, що враховує його складність і багатшаровість, зберігаючи аналітичну точність навіть за умови агресивного маскування контенту.

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

Наукові положення, висновки та рекомендації, представлені у дисертаційній роботі, є достатньо обґрунтованими та методологічно цілісними. Здобувачем проведено ґрунтовний багатовимірний аналіз існуючих підходів.

Запропонована концепція аналізу текстів полягає у виділенні ознак для кожної лінгвістичної модальності (лексичної, синтаксичної, семантичної) та їх об'єднанні в єдиний багатовимірний простір ознак $\mathbf{F} = \mathbf{F}_{lex} \times \mathbf{F}_{syn} \times \mathbf{F}_{sem}$. Такий інтегрований підхід забезпечує формування окремих множин ознак для розробки методів класифікації,

що дозволяє підвищити інформативність масивів вхідних даних та збільшити кількість правильно ідентифікованих текстів.

Достовірність результатів підтверджується масштабними комп'ютерними експериментами на репрезентативних наборах даних та застосуванням строгих статистичних метрик.

Новизна наукових положень, висновків і рекомендацій, сформульованих у дисертації

Наукова новизна результатів дисертаційної роботи полягає в вирішенні важливої науково-прикладної проблеми - розробленні методології багатовимірного аналізу сліду ШІ. Отримано такі нові наукові результати:

1. Вперше розроблено гібридний метод ідентифікації згенерованих текстів, який, на відміну від існуючих, будує точки спостереження шляхом адаптивного, погодженого та поетапного об'єднання ознак із різних модулів у єдиний вектор метакласифікатора.
2. Вперше розроблено інтерпретаційну модель пояснювального ШІ, що базується на внутрішніх вагах класифікатора і забезпечує формування прозорі картографії доказів для експертних платформ.
3. Одержав подальший розвиток метод забезпечення надійності класифікації в умовах епістемічної невизначеності, зокрема на OOD-даних, шляхом інтеграції фреймворку Індуктивного конформного передбачення.
4. Удосконалено принципи багатовимірної оцінки робастності систем через концептуальну модель «Точки зламу», що встановлює чіткий математичний зв'язок між коефіцієнтом заміни RR , семантичною цілісністю I_{sem} та ймовірністю виявлення P_{det} .

Оцінка наукових публікацій здобувача

Аналіз публікаційної активності здобувача підтверджує, що основні результати дисертації повно висвітлені у наукових працях. Особливої уваги заслуговують публікації у виданнях, що індексуються у базах Scopus та фахових виданнях України. Концептуальні засади гібридного методу та методики багатовимірного аналізу лінгвістичних даних пройшли належну апробацію та сліпе рецензування у фаховому середовищі.

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 та списку використаних джерел. Матеріал викладено з дотриманням принципів системності та логічної послідовності.

У **вступі** обґрунтовано актуальність інтеграції різнорівневих лінгвістичних ознак, які виступають аналогом мультимодальних даних для ідентифікації згенерованого контенту. Сформульовано науково-прикладну задачу, визначено об'єкт, предмет, мету та завдання дослідження.

Перший розділ присвячено системному аналізу існуючих інформаційних технологій детекції текстів. Доведено вразливість одновимірних підходів до трансформаційних атак маскування. Показано необхідність парадигмального переходу до методів багатовимірної аналізу лінгвістичних об'єктів.

Другий розділ формує теоретико-методологічний базис. Запропоновано інноваційну математичну модель сліду ШП, що формалізує його як вектор у багатовимірному просторі лексичних, синтаксичних та семантичних ознак $\mathbf{F} = \mathbf{F}_{lex} \times \mathbf{F}_{syn} \times \mathbf{F}_{sem}$. Розроблено концептуальну модель «Точки зламу» для прогнозування чутливості системи та архітектуру гібридного каскадного ансамблю з адаптивним зважуванням.

Третій розділ описує методологію комп'ютерного експерименту та етап формування наборів даних, що включає тексти МоЕ, Reasoning-моделей та результати парафразерів T5 і DIPPER. Експериментально доведено зміщення «Точки зламу» гібридного методу екстремально праворуч, що математично підтверджує нівелювання зони успішного маскування.

Четвертий розділ розкриває практичну реалізацію інформаційної технології у хмарному середовищі AWS з асимптотичною складністю $O(N^2)$ та середнім часом обробки одного тексту на рівні 0.33с. Вирішено проблему епістемічної невизначеності шляхом інтеграції фреймворку індуктивного конформного передбачення. Також формалізовано інтерпретаційну модель ХАІ, що генерує глобальний і локальний профілі атрибуції та «Карту доказів» із низькою обчислювальною вартістю.

Практична цінність результатів роботи

Практичне значення одержаних результатів полягає у створенні готового до промислового використання інструментарію багатовимірної аналізу для систем контент-модерації, інформаційної безпеки та платформ перевірки академічної доброчесності. На основі запропонованих математичних моделей розроблено програмний комплекс ідентифікації автоматично згенерованих текстів, який розгорнуто у хмарному середовищі AWS. Завдяки глибокій оптимізації алгоритмів загальна асимптотична складність системи становить $O(N^2)$, а середній час обробки одного документа зведено до 0,33 секунди.

Експериментально підтверджено ефективність програмної реалізації гібридного методу у протидії комерційним засобам маскування тексту (StealthGPT, AIHumanize, Phrasly) із доведеним зниженням рівня успішності атак до 22.84%. Впровадження методики застосування фреймворку індуктивного конформного передбачення дало змогу суттєво знизити рівень хибнопозитивних спрацювань. Так, FPR на складних OOD-текстах датасету IELTS було знижено з 4.81% до 0.98%, що забезпечило підсумковий коефіцієнт порятунку на рівні 79.71%. Це підтверджує високу прикладну цінність розробки для мінімізації ризиків несправедливих звинувачень.

Дискусійні положення, зауваження та рекомендації

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

- Багатокритеріальне прийняття рішень:** У підрозділі 2.3.3 (с. 105) та на Рисунку 2.2 «Архітектура гібридного методу» (с. 101) наведено логіку роботи системи, де гібридний метод використовує метакласифікатор на базі логістичної регресії для прийняття фінального рішення. З огляду на необхідність балансування між метриками FPR та Recall у багатовимірному просторі ознак, доцільним було б розглянути можливість застосування методів багатокритеріального аналізу, наприклад, методу аналізу ієрархій для експертного калібрування чи налаштування вагових коефіцієнтів системи.
- Управління ризиками в умовах невизначеності:** У підрозділі 4.2.1 «Валідація методу ICP для ідентифікації OOD-даних» (с. 188) та в Таблиці 4.9 (с. 189) вказується, що використання фреймворку індуктивного конформного передбачення суттєво зменшує ризик хибнопозитивних спрацювань шляхом переведення нетипових OOD-даних у стан "невизначеності". Проте у тексті

роботи недостатньо розкрито механізми управління ризиками при масовому надходженні таких даних і подальшої автоматизації роботи з порожніми множинами.

3. **Статистична валідація $O(1)$ моделі пояснень:** У підрозділі 4.4.1 (зокрема, на Рис. 4.13, с. 206) та підрозділі 2.4 (с. 109-117) розроблено локальний профіль атрибуції з часовою складністю $O(1)$. Оскільки класичні методи, на кшталт SHAP, базуються на строгій теорії кооперативних ігор для усунення зміщення, у роботі потребує додаткового математичного обґрунтування питання відсутності зміщення впливу залежних ознак у запропонованій прямій моделі розрахунку пояснень.
4. **Візуалізація багатовимірного простору ознак:** У другому розділі (підрозділ 2.1, с. 73) ґрунтовно описується математична формалізація багатовимірного простору ознак $\mathbf{F} = \mathbf{F}_{lex} \times \mathbf{F}_{syn} \times \mathbf{F}_{sem}$. Для кращого розуміння природи розподілу класів (генерованих текстів та текстів авторів-людей) у цьому просторі, доцільно було б доповнити дослідження візуалізаціями з використанням алгоритмів зниження розмірності, наприклад, t-SNE або UMAP. Це дозволило б наочно продемонструвати щільність кластерів та рівень їхнього перекриття.
5. **Порівняльний аналіз альтернативних архітектур агрегації:** У підрозділі 2.3.3 (с. 105) описано застосування метакласифікатора на базі логістичної регресії для фінальної агрегації результатів окремих модулів. Проте для підтвердження оптимальності такого вибору варто було б навести результати порівняльного експерименту з іншими популярними методами багатовимірного класифікування, такими як Random Forest та Support Vector Machines у ролі агрегуючого модуля.

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

Загальний висновок

Дисертаційна робота Никоненка Андрія Олександровича на тему «Гібридний метод ідентифікації автоматично згенерованих природномовних текстів» є завершеним, самостійним інноваційним науковим дослідженням. 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) з галузі знань 12 «Інформаційні технології» за спеціальністю 122 «Комп'ютерні науки».

Офіційний опонент,

кандидат економічних наук,

доцент кафедри систем штучного інтелекту

Національного університету «Львівська політехніка»

Наталія БОЙКО

підпис
засвідч.
профес.



Вона зрештеш. ЖЕХУХА