Голові разової спеціалізованої
вченої ради PhD 8729
Хмельницького національного університету
доктору технічних наук, професору
Тетяні ГОВОРУЩЕНКО

ВІДГУК

офіційного опонента на дисертаційну роботу
Молчанової Марини Олексіївни
за темою «Методи виявлення та класифікації прийомів та об'єктів пропаганди 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.

На сучасному етапі публікації в інтернеті стають не лише засобом передачі фактів, але й полем для прихованих маніпуляцій, спрямованих на формування потрібних масам установок. Традиційні підходи експертної перевірки та семантичного аналізу, що спираються на людський фактор, виявляються занадто повільними та ресурсомісткими, аби протидіяти шалено зростаючому потоку повідомлень у соціальних мережах, блогах і новинних агрегаторах. У цьому контексті впровадження методів штучного інтелекту набуває вирішального значення: машинне навчання та глибоке навчання дозволяють автоматизувати розпізнавання як явних прийомів пропаганди, так і тонких семантичних маніпуляцій.

Сучасні нейромережеві архітектури в поєднанні з технологіями обробки природної мови здатні навчатися на великих корпусах текстів, виокремлювати характерні ознаки пропагандистських стратегій та класифікувати інформаційні об'єкти за ступенем їх маніпулятивного потенціалу. Це відкриває перспективи створення систем оперативного моніторингу інформаційного простору, які можуть стати важливим інструментом для журналістів, аналітиків безпеки та освітніх платформ з медіаграмотності.

Застосовування нейромереж архітектури трансформер відкриває можливості до виявлення як поверхневих патернів лексичних маркерів маніпуляції, так і глибинних семантичних аномалій. Використання багатовимірних ембедингів дозволяє будувати простори представлення текстів, де пропагандистські фрейми виявляють свою корелюючу структуру. Розробка та апробація алгоритмів на основі донавчання попередньо навчених мовних

моделей дає змогу адаптувати рішення під специфіку локальних інформаційних векторів.

Особливої гостроти завдання виявлення пропаганди набуває в умовах сучасного інформаційного протистояння, коли ефективність впливу на суспільну свідомість стає одним із вирішальних чинників національної безпеки. Український контекст, де гібридні загрози супроводжуються масштабними кампаніями дезінформації, висуває нагальну потребу в інструментах, які дозволять виявляти не лише шаблонні прийоми впливу, але й тонко масковані семантичні конструкції. Розробка автоматизованих підходів до виявлення і класифікації пропагандистських технік сприяє підвищенню стійкості суспільства до інформаційних атак та зміцненню інформаційного суверенітету.

Дисертаційна робота Молчанової М. О. присвячена розробці методів і засобів для автоматизованого виявлення прийомів та об'єктів пропаганди, що є актуальною науково-прикладною задачею й дозволяє комплексно аналізувати взаємозв'язки виявлених прийомів і об'єктів пропаганди, та сприяє підвищенню точності автоматизованого виявлення пропаганди. Таким чином, тема дисертаційної роботи є значущою й актуальною, що підтверджується її відповідністю пріоритетним тематичним напрямам наукових досліджень і науково-технічних розробок.

Аналіз змісту дисертації та основні результати роботи.

Метою дисертаційного дослідження є підвищення точності та якості виявлення прийомів та об'єктів пропаганди за семантичними маркерами у текстовому контенті засобами штучного інтелекту з подальшим поясненням прийнятих рішень.

Загалом робота, за деякими незначними винятками, коректно структурована відповідно до мети і завдань дослідження. Робота написана на достатньому мовно-стилістичному рівні. Зміст дисертації в повній мірі дозволяє скласти уявлення про основні положення, запропоновані авторкою, їх практичне значення, а також результати дослідження запропонованих рішень за рядом критеріїв.

У *першому розділі* роботи проведено комплексний аналіз методів, засобів та технологій для автоматизованого виявлення пропаганди у текстовому контенті. Визначено важливість подальшого розвитку та вдосконалення методів штучного інтелекту для автоматизованого виявлення пропаганди, що є критичним для забезпечення інформаційної безпеки та протидії маніпулятивному впливу на суспільство. Проаналізовано існуючі підходи до виявлення та класифікації прийомів пропаганди, що дозволило виокремити семантичні особливості, притаманні прийомам. Виявлено низку проблем,

пов'язаних з виявленням та класифікацією пропагандистських прийомів, таких як складність автоматизації процесу через різноманіття мовних конструкцій та контекстів. Обґрунтовано використання класифікації за 17-ма прийомами пропаганди у межах дисертаційного дослідження. За результатом проведеного аналізу метою роботи визначено підвищення точності та якості виявлення прийомів та об'єктів пропаганди з подальшим поясненням прийнятих рішень.

У *другому розділі* розроблено концепцію виявлення та класифікації прийомів та об'єктів пропаганди у текстовому контенті, яка використовує нейромережеві засоби та дозволяє виявляти приналежність об'єктів до використаних прийомів пропаганди Запропонований підхід містить етапи класифікації текстів за вмістом пропаганди, виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень та виявлення об'єктів пропаганди з візуальною інтерпретацією прийнятих рішень. Наведено модель семантичної структури пропаганди для формалізованого подання семантичної моделі пропаганди у тексті. Висвітлено удосконалений авторкою метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання, що призначений для автоматизованої ідентифікації текстів, які містять пропагандистські елементи. У межах удосконаленого методу розглянуто особливості використання нейромережевих моделей BiLSTM та GRU, а також нейромережевої моделі BiLSTM гібридної архітектури, для класифікації текстів за вмістом пропаганди.

У *третьому розділі* наведено методи виявлення прийомів та об'єктів пропаганди. Наведено розроблений метод виявлення прийомів пропаганди за маркерами, призначений для аналізу текстового контенту на предмет наявності пропагандистських прийомів та визначення сили їх проявів. Відмінність запропонованого методу від існуючих полягає у використанні додаткової множини маркерів при навчанні нейромережевих класифікаторів. Розглянуто особливості формування навчальних множин даних для нейромережевих моделей класифікації прийомів пропаганди. Також в розділі наведено розроблений авторкою метод виявлення об'єктів пропаганди з візуальною інтерпретацією, який дозволяє забезпечити комплексний аналіз взаємозв'язків прийомів та об'єктів пропаганди в текстах, а також забезпечує узагальнення для об'єктів пропаганди та їх альтернативних текстових згадувань. Розроблені методи дозволяють не лише ідентифікувати наявність пропаганди, а й забезпечити візуальну інтерпретацію отриманих результатів.

У *четвертому розділі* наведено опис розробленого прикладного експериментального програмного забезпечення, що реалізує розроблені методи. Досліджено ефективність комплексного підходу до нейромережевого

виявлення і класифікації прийомів та об'єктів пропаганди у текстовому контенті. Виявлено, що запропонований метод класифікації текстів за вмістом пропаганди дозволяє досягнути точності 0.978, що є на 0.035 вище за реалізовані відомі аналоги. Розроблений метод виявлення прийомів пропаганди за маркерами забезпечує виявлення різних пропагандистських прийомів з мінімальною точністю 0.82 та з середньою точністю 0.886, що краще за відомі аналоги виявлення пропаганди мінімум на 0.368. Розроблений метод виявлення об'єктів пропаганди дозволяє отримувати результати, які цілком корелюють із результатами, одержаними експертами.

У висновках подано підсумок по проведеній роботі, зокрема перелік вирішених завдань, результати експериментального тестування запропонованих методів та відомості про практичне застосування розроблених у роботі положень. Список літератури, в загальному, відповідає рівню досягнень галузі. Додатки виглядають доречними.

Основні результати роботи:

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

– удосконалено метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання;

– розроблено метод виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень;

– розроблено метод виявлення об'єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень;

– розроблено інтелектуальну інформаційну систему для валідації запропонованих методів і проведено експериментальні дослідження.

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

Обґрунтованість наукових висновків дисертації забезпечується глибоким аналізом існуючих досліджень. Правильність застосування методів, математичних моделей і програмних інструментів підтверджується результатами експериментів. Наукова новизна отриманих результатів полягає в наступному:

1. Удосконалено метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання, який відрізняється від існуючих модифікованою архітектурою нейромережі та обсягом вихідних даних, що дало змогу підвищити точність класифікації.

2. Розроблено новий метод виявлення прийомів пропаганди за маркерами з візуальною інтерпретацією прийнятих рішень, який відрізняється від існуючих використанням доповненої множини семантичних маркерів для виявлення

прийомів пропаганди, що дало змогу пояснити отримані результати і підвищити точність та якість виявлення пропаганди.

3. Розроблено новий метод виявлення об'єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень, який відрізняється від існуючих асоціативним групуванням об'єктів пропаганди, що дало змогу покращити результати виявлення об'єктів пропаганди та візуально їх інтерпретувати.

Теоретичне та практичне значення одержаних результатів.

Наукова цінність роботи полягає в тому, що науково обґрунтовано, теоретично доведено і підтверджено практичними дослідженнями підхід до виявлення та класифікації прийомів та об'єктів пропаганди у текстовому контенті, який дозволяє підвищити точність та якість виявлення прийомів та об'єктів пропаганди за семантичними маркерами у текстовому контенті засобами штучного інтелекту з подальшим поясненням прийнятих рішень.

Практичне значення отриманих результатів полягає в реалізації теоретичних результатів дисертаційної роботи та безпосередньому використанні їх у виробничій діяльності. Розроблено інтелектуальну систему, яка надає користувачам можливість автоматизованого аналізу текстів, забезпечуючи комплексний результат у вигляді загальної оцінки прояву пропаганди у тексті, виявлених пропагандистських прийомів, оцінок приналежності виявлених об'єктів пропаганди до використаних прийомів. Також система дозволяє отримувати візуальну інтерпретацію прийнятих рішень, що сприяє підвищенню прозорості та довіри до отриманих результатів.

Результати дисертаційної роботи впроваджено у діяльності підприємств (наявні 4 довідки про впровадження), в навчальному процесі Хмельницького національного університету (наявний акт про впровадження), при виконанні держбюджетної теми Хмельницького національного університету «Розроблення інформаційної технології прийняття контрольованих людиною критично-безпекових рішень за ментально-формальними моделями машинного навчання».

Розроблені в роботі методи дозволили у повній мірі досягти мети дисертаційного дослідження, яка полягала у підвищенні точності та якості виявлення прийомів та об'єктів пропаганди з подальшим поясненням прийнятих рішень. Підвищення точності досягається методом класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання (підвищує точність на 0.035 за метрикою Accuracy та щонайменше на 0.06 за метрикою F1 порівняно з існуючими підходами) та методом виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих

рішень (підвищує в середньому точність за метрикою Accuracy на 0.368 порівняно з відомими аналогами). Підвищення якості полягає у додатковій поясненості нейромережевих рішень шляхом порівняння отриманих значень з еталонними значеннями маркерів, а також у застосуванні візуальної аналітики (LIME) в контурі реалізації методу виявлення прийомів пропаганди за маркерами із візуальною інтерпретацією прийнятих рішень. Також підвищення якості полягає у візуальній інтерпретації прийнятих рішень щодо виявлених об'єктів пропаганди, а також групуванні об'єктів пропаганди і визначенні їх приналежності до виявлених прийомів пропаганди в контурі реалізації методу виявлення об'єктів пропаганди нейромережевими моделями глибокого навчання з візуальною інтерпретацією прийнятих рішень.

Повнота викладу результатів дисертації в опублікованих працях.

Основні результати дисертації опубліковані у 12 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); 3 авторських свідоцтва. Ознайомлення з дисертацією, копіями статей і тез дозволяє зробити висновок про повноту викладення здобутих наукових результатів в опублікованих працях.

Основні результати повністю висвітлено у 4 наукових статтях у фахових виданнях України. Одна стаття має двох співавторів, три статті одноосібні, тому згідно Підпункту 1 пункту 8 в редакції Постанови КМ № 507 від 03.05.2024, ці статті зараховуються повністю. Одна стаття має 2 (два) співавтори, тому згідно наведеної вище постанови, також зараховується повністю. Тобто здобувачка має 4 публікації, у яких викладені основні результати дисертації, чого достатньо згідно чинних вимог.

Положення роботи апробовано на 5 Міжнародних конференціях, за їх результатами виконано п'ять публікацій, що індексуються в міжнародній наукометричній базі «Scopus». Чітко зазначено особистий внесок дисертанта у кожній спільній публікації.

Одержані здобувачкою два авторських свідоцтва є одноосібними.

Зауваження та побажання.

При цілком позитивній оцінці роботи, вважаю за необхідне зробити такі зауваження:

1. У списку використаних джерел зазначено низку публікацій автора, зокрема й одноосібних, у виданнях, включених до Переліку наукових фахових видань України, які, на дату опублікування, віднесені до категорії «Б» та тези

науково-практичних конференцій, які не включено до списку публікацій здобувача за темою дисертації. Хоча вже включені до списку публікації цілком відображають одержані результати роботи, було б доцільним додати до списку й інші публікації автора, виконані за темою дисертації.

2. У 1.4. Проблеми виявлення та класифікації прийомів і об'єктів пропаганди:

- «LIME є методом пояснення прогнозів моделей машинного навчання, який фокусується на локальній інтерпретації рішень. Його основна ідея полягає в апроксимації складної моделі простою лінійною моделлю в околі конкретного передбачення.»

LIME, як і інші методи, що апроксимують складну модель простішою, має сенс застосовувати коли є апріорні підстави вважати, що спрощена модель близька до істинної, а ускладнені моделі вивчають шум на рівні з простішою моделлю. З тексту не очевидно, що прихована справжня залежність є лінійною. Тому отримані лінійні моделі дадуть результат, який буде схожим на правду, але без будь-яких підстав вважати його правдивим.

3. У 2.5. Метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання:

- «Обмеження за обсягами текстової інформації також частково пов'язане з наявними навчальними даними і в межах дослідження працює з текстовими дописами довжиною від 200 до 6300 символів. Також обмеження пов'язане з самою сутністю пропаганди. Тексти менше 200 символів не можуть повною мірою розкрити її зміст і можуть призводити до хибних результатів, а тексти понад 6300 символів є більш складними і на ряду з короткими навчальними текстами їх аналіз також буде нерелевантним. »

Видається дивним обмеження на довжини текстів.

а) короткі слогани дуже добре передають пропагандистські твердження. Наприклад, слогани часів Другої Світової «Loose lips sink ships», «Careless Talk Costs Lives» (американський та британський, відповідно) чітко підпадають під апеляцію до страху, і не містять нічого окрім пропаганди.

б) довгі твори можуть містити багато різних елементів пропаганди, тому важливо коректно визначити фокусні об'єкти, які представлені, зазвичай, іменованими сутностями або займенниками. Пропаганда переважно буде міститися у їх найближчому околі, окрім випадків складних ланцюжків думок.

4. В 3.3. Метод виявлення об'єктів пропаганди з візуальною інтерпретацією прийнятих рішень:

- «Метод виявлення об'єктів пропаганди використовує нейромережеві моделі глибокого навчання та містить етапи: побудова набору об'єктів пропаганди PO шляхом розпізнавання іменованих сутностей; попередня обробка тексту та розширення PO за рахунок слів-представлень об'єктів; побудова контекстних вікон CW і їх об'єднання за PO′; виявлення рівня використання пропагандистських прийомів у CW′ нейромережевими моделями NM′; побудова множини важливості віднесення RTO′ між TT′ та об'єктами PO′ для тестового тексту T (рис. 3.11).»

Повністю підпадає під зауваження до «2.5. Метод класифікації текстів за вмістом пропаганди нейромережевими моделями глибокого навчання». Дослідниця виділяє фокусні об'єкти та вікна уваги. Швидше за все, проблеми з короткими текстами зумовлені використанням FastText замість більш потужного засобу, наприклад, на основі BERT. Імовірно, що проблеми з довгими текстами зумовлені використанням інших риторичних прийомів, що не були досліджені в роботі, та нелінійною структурою дискурсу, що робить задачу перебірною, оскільки для людини допустимо вести виклад перестрибуючи з думки на думку та повертаючись назад. Для надійної обробки такої структури треба групувати віднайдені вікна в потенційні ланцюжки, що робить задачу повноперебірною відносно кількості віднайдених фокусних об'єктів та вікон.

5. Опора на прийоми замість віднаходження та верифікації фактів є справді дієвим методом. Проте наявність базових засобів контролю фактажу має посилити ефективність.

6. Одержані результати було впроваджено в діяльності ІТ-фірм, про що свідчать довідки впровадження, наведені у Додатку Б. Проте в роботі не наведено одержані кількісні результати впровадження та їх аналіз, хоча ці дані можуть свідчити про позитивний ефект від використання одержаних у роботі результатів.

Наведені зауваження та побажання не є суттєвими, істотно не впливають на загальну позитивну оцінку дисертаційної роботи і якість одержаних результатів.

Загальний висновок.

У цілому, дисертаційна робота «Методи виявлення та класифікації прийомів та об'єктів пропаганди у текстовому контенті засобами штучного інтелекту» є завершеною науково-дослідною працею і відповідає Стандарту вищої освіти України зі спеціальності 122 – «Комп'ютерні науки» для третього (освітньо-наукового) рівня вищої освіти, зокрема об'єкту вивчення та діяльності «процеси обробки інформації у комп'ютерних системах». Дисертація

присвячена розв'язанню актуальної науково-прикладної задачі виявлення та класифікації пропагандистських прийомів і об'єктів у текстовому контенті. Розроблені в роботі методи дозволяють ефективно виявляти та класифікувати прийоми пропаганди та аналізувати інформаційні загрози для використання у медіа-ресурсах та соціальних мережах, що є важливим кроком у боротьбі з дезінформацією та пропагандою. Розроблені методи дозволили у повній мірі досягти мети дисертаційного дослідження, яка полягала у підвищенні точності та якості виявлення прийомів та об'єктів пропаганди за семантичними маркерами у текстовому контенті засобами штучного інтелекту з подальшим поясненням прийнятих рішень.

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

Офіційний опонент
професор кафедри математичної інформатики
факультету комп'ютерних наук та кібернетики
Київського національного університету
імені Тараса Шевченка,
доктор фізико-математичних наук,
професор _____ Олександр МАРЧЕНКО

Підпис О. Марченка засвідчую.
Зав. канц. _____ /Дмитро ЗАТУЛА/