РЕЦЕНЗІЯ

на дисертаційну роботу
Шаптали Романа Віталійовича
на тему «Класифікація документів на основі векторних представлень словників при обробці природної мови у малоресурсному середовищі»,
представлену на здобуття ступеня доктора філософії
в галузі знань Інформаційні технології
за спеціальністю 122 – Комп'ютерні науки

Актуальність теми дисертації.

Тема дисертаційної роботи є актуальною тому, що на даний момент існує велика кількість малоресурсних середовищ, у яких потрібно класифікувати документи. Збільшення точності існуючих моделей класифікації, а також розробка нових методів для обробки малоресурсних мов є актуальною задачею поточного етапу розвитку методів інтелектуального аналізу даних та машинного навчання.

Причинами малоресурсності мовних середовищ можуть бути непопулярність мови у інформаційних джерелах чи вузька тематика прикладної області, а отже і документів, що її описують. Багато сучасних методів обробки природних мов показують низьку якість при класифікації таких текстів, тому розвиток даної області також сприяє розумінню недоліків поточних підходів. Варто зазначити, що обмеження, які накладає малоресурсність середовища на ефективність моделей класифікації документів значні, що суттєво ускладнює вирішення типових завдань, а отже покращення методів обробки природної мови у таких середовищах має високу прикладну цінність.

Оцінка обгрунтованості наукових результатів дисертації, їх достовірності та новизни.

Наукова новизна результатів дисертаційного дослідження полягає в наступному:

1. Запропоновано метод класифікації документів на основі векторних представлень словників при обробці природної мови, який через поєднання векторних представлень документів та векторних представлень слів зі словника синонімів покращує F1-міру якості класифікації документів у малоресурсному середовищі.
2. Запропоновано векторну модель слів зі словника синонімів, яка будується на основі графового представлення словника за допомогою методів кодування вузлів графу.
3. Модифіковано методи злиття векторних представлень слів, а саме конкатенації та зваженої суми, через додатковий крок пошуку

відповідника слову з документа у словнику синонімів на основі критеріїв міжрядкової відстані.

Достовірність результатів забезпечена коректним застосуванням відповідного математичного апарату та обчислювальних експериментів на основі набору даних петицій до Київської міської ради з підтвердженням статистичної значущості отриманих результатів за допомогою Хі-квадрат тесту за методом МакНемара. Контроль метрик здійснювався на незалежній тестовій вибірці, що підсилює достовірність отриманих наукових результатів.

Наукові положення, висновки та рекомендації достатньо обгрунтовані, базуються на сучасних вітчизняних та зарубіжних джерелах, а також правильно використовують методи інтелектуального аналізу даних, обробки природних мов, математичної статистики та штучного інтелекту.

У дисертації здобувач запропонував нові наукові результати, основним з яких є метод класифікації документів, що працює на основі характеристик утворених поєднанням векторних представлень слів у документі та векторних представлень слів зі словника. Моделювання словника як графу з навчанням векторних представлень слів на його основі є новим та перспективним підходом, який можна повторно використовувати у інших середовищах. Цікавою є також ідея обходу побудови моделей визначення частини мови та пошуку словоформ при пошуку відповідних векторних представлень словника через застосування міжрядкової відстані.

Наукові дослідження були виконані здобувачем на кафедрі Системного проектування КПІ ім. Ігоря Сікорського в рамках тематичного плану науково-дослідних робіт кафедри системного проектування під керівництвом к.т.н., ст.н.с. Кисельова Геннадія Дмитровича. Результати дисертації використані в проектах ННК ІПСА з підтримки та супроводження грід-центру засвідчення сертифікатів користувачів і грід-сайтів національної грід-інфраструктури: НДР № 2299/20 (номер держреєстрації 0120U103046), НДР № 2302/21 (номер держреєстрації 0121U110624), НДР № 2307/22 (номер держреєстрації 0122U002655), які виконувались згідно Програми інформатизації НАН України на 2020 – 2024 р.

Отже, в дисертаційній роботі поставлене наукове завдання розробки методів обробки природної мови на основі векторних представлень словників у малоресурсному середовищі виконано повністю, здобувач повною мірою оволодів методологією наукової діяльності.

Оцінка змісту дисертації, її завершеність та дотримання принципів академічної доброчесності.

За своїм змістом дисертаційна робота здобувача Шаптали Р.В. повністю відповідає Стандарту вищої освіти зі спеціальності 122 – Комп'ютерні науки та напрямкам досліджень відповідно до освітньої програми Комп'ютерні науки.

Дисертаційна робота є завершеною науковою працею і свідчить про наявність особистого внеску здобувача у науковий напрям Комп'ютерні науки.

Розглянувши звіт подібності за результатами перевірки дисертаційної роботи на текстові співпадіння, можна зробити висновок, що дисертаційна робота Шаптали Романа Віталійовича є результатом самостійних досліджень здобувача і не містить елементів фальсифікації, компіляції, фабрикації, плагіату та запозичень. Використані ідеї, результати і тексти інших авторів мають належні посилання на відповідне джерело.

Мова та стиль викладення результатів.

Дисертаційна робота написана українською мовою. Робота написана грамотно, послідовно та доступно. Використано науковий стиль та застосована сучасна загальноприйнята наукова термінологія. Текст дисертації чітко структуровано, що полегшує її сприйняття, дає можливість сфокусуватись на конкретному аспекті дослідження та посиланнях на відповідні джерела. Таке подання наукових результатів на основі теоретичних та експериментальних досліджень забезпечує доступність їх сприйняття та подальший розвиток наукового напряму.

Дисертація складається з вступу, трьох розділів, висновків, списку літератури та додатків. Загальний обсяг дисертації 151 сторінка.

У вступі визначаються мета та основні завдання дослідження, обгрунтовується актуальність теми. Також описуються проблематика існуючих підходів та наводиться наукова і практична новизна отриманих результатів. Здобувач також зазначає особистий внесок, апробацію матеріалів дисертації та перелічує публікації за темою роботи.

У першому розділі висвітлено аналіз методів та досліджень в галузі обробки природних мов у малоресурсних середовищах. На основі даного аналізу побудована класифікація методів обробки природних мов у малоресурсних середовищах. Виділено обмеження кожного з методів, а також припущення на яких вони базуються, що визначає можливу область застосувань.

Другий розділ присвячено методиці побудови векторних представлень словників, а також методам злиття векторних представлень. Здобувач побудував класифікацію методів побудови векторних представлень графів та виділив їх особливості. Класифікацію і методи на основі факторизації графу, і на основі глибокого навчання, і на основі випадкових блукань. Саме у цьому розділі пропонується модифікація методу злиття векторних представлень слів через додатковий етап пошуку відповідності слів на основі міжрядкових відстаней.

У третьому розділі детально описано експериментальні результати запропонованих автором методів. Для перевірки ефективності запропонованих методів було використано множину даних петицій до Київської міської ради з

розміткою за темами звернень. Проведено розвідувальний аналіз набору даних та його поділ на вибірки для тестування і тренування. У розділі продемонстровано підсумки практичних дослідів з різними архітектурами та компонентами, їх порівняльний аналіз та оцінка статистичної значущості результатів.

Дисертаційна робота оформлена відповідно до вимог наказу МОН України від 12 січня 2017 р. № 40 «Про затвердження вимог до оформлення дисертації».

Оприлюднення результатів дисертаційної роботи.

Наукові результати дисертації висвітлені у 7 наукових публікаціях здобувача, серед яких: 5 статей у наукових фахових виданнях України, 4 з яких включені на дату опублікування до переліку наукових фахових видань України за спеціальністю 122 Комп'ютерні науки, та 1 стаття у періодичному науковому виданні, проіндексованому у базі даних Scopus.

Також результати дисертації апробовано на одній науковій фаховій конференції.

Усі публікації здобувача мають високий науковий рівень. У них детально розкриваються основні наукові результати дослідження. Особистий внесок здобувача до публікацій за співавторством вагомий, особливо у описі експериментальних частин роботи. Принципів академічної доброчесності у жодній з публікацій не порушено.

Таким чином, наукові результати, описані в дисертаційній роботі, повністю висвітлені у наукових публікаціях здобувача.

Недоліки та зауваження до дисертаційної роботи.

1. У дисертаційній роботі варто було додатково розглянути існуючі векторні представлення слів для української мови, на відміну від побудови власних представлень на конкретній множині даних.
2. У тексті дисертації іноді наявні стилістичні помилки та термінологічні неузгодженості, наприклад «домен» та «прикладна область».
3. У дослідженні не розглядаються різні класифікатори, порівнюються лише ознаки, які надходять на вхід багатошаровому персептрону. Інші моделі класифікації могли б мати вплив на отримані результати.
4. Експерименти проводились лише на одному наборі даних, хоч він і відповідає поставленим завданням. Додаткові набори даних додали б упевненості у результатах.
5. Окрім F1-міри, доцільно було б додати інші метрики класифікації, такі як точність та повнота. Це могло б показати у яких випадках модель більше помиляється.

Вважаю, що висловлені зауваження не є визначальними і не зменшують загальну наукову новизну та практичну значимість результатів і не впливають на позитивну оцінку дисертаційної роботи.

Висновок про дисертаційну роботу

Вважаю, що дисертаційна робота здобувача ступеня доктора філософії Шаптали Романа Віталійовича на тему «Класифікація документів на основі векторних представлень словників при обробці природної мови у малоресурсному середовищі» виконана на високому науковому рівні, не порушує принципів академічної доброчесності та є закінченим науковим дослідженням, сукупність теоретичних та практичних результатів якого розв'язує наукове завдання, що має істотне значення для інформаційних технологій. Дисертаційна робота за актуальністю, практичною цінністю та науковою новизною повністю відповідає вимогам чинного законодавства України, що передбачені в п. 6 – 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.

Здобувач Шаптала Роман Віталійович заслуговує на присудження ступеня доктора філософії в галузі знань Інформаційні технології за спеціальністю 122 – Комп'ютерні науки.

Рецензент:
професор кафедри
математичних методів
системного аналізу
Навчально-наукового інституту
прикладного системного аналізу
Національного технічного
університету України
«Київський політехнічний інститут
імені Ігоря Сікорського»,
д.т.н., професор _____ Петро БІДЮК

Підпис гр. _____
ЗАСВІДЧУЮ
Відділ кадрів і архівної справи
підпис пр-ще

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського» Відділ кадрів і архівної справи № 6

М.П. «14» серпня 2023 року