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. Обробка великих масивів текстових даних необхідна в багатьох галузях людської діяльності. Тому дослідження з забезпечення обробки природної мови, особливо в області обробки мови у малоресурсному середовищі, є затребуваними. Вдосконалення сучасних методів вирішення задач обробки природної мови, а також створення методів, що дозволяють використовувати апробовані засоби вирішення задач у високоресурсному середовищі для досягнення таких цілей у малоресурсному середовищі, є актуальним.

У дисертації розглядаються виклики та обмеження, пов'язані з обробкою природної мови в середовищах з низьким рівнем ресурсів. Із зростаючим попитом на системи, здатні працювати з текстовою інформацією різними мовами та у різних прикладних областях, брак тестових даних стає критичною проблемою. Дослідження дисертації спрямовано на розробку та вдосконалення методів класифікації документів із використанням векторних представлень графів лінгвістичних словників, що є новим підходом до усунення обмежень існуючих методів.

Ще одним аспектом актуальності дисертації є дослідження, спрямовані на створення методів для покращення інтерпретації мови та вдосконалення класифікації документів у умовах обмежених ресурсів. Запропонований підхід може підвищити якість систем обробки природної мови для мов з обмеженими ресурсами та сприяти подоланню розриву між такими широко вивченими мовами, як англійська, та мовами, які недостатньо представлені у інформаційному середовищі.

Оцінка обґрунтованості наукових результатів дисертації, їх достовірності та новизни. Отримані наукові результати розв'язують актуальну науково-прикладну задачу підвищення точності класифікації документів на природній мові за умови недостатніх для навчання тестових даних.

Автором вперше розроблено 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.

Автором вперше запропоновано векторну модель слів зі словника синонімів, яка будується на основі графового представлення словника за допомогою методів кодування вузлів графу.

Автором вдосконалено 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 наукових результатів експериментально доведена. Використання статистичних методів перевірки статистичної значущості результатів експериментів, наприклад, Хі-квадрат тесту за методом МакНемара, підтверджує достовірність та надійність висновків даної роботи.

Також у дисертації проведено експерименти на тестовому наборі даних, результати яких доводять ефективність запропонованих методів.

Аналіз змісту розділів та використаних методів дозволяє зробити висновок про належну обґрунтованість наукових результатів. Наукові положення та висновки, представлені у дисертації, обґрунтовано теоретичним аналізом, результатами практичного використання та інформацією з науково-технічної літератури.

Дисертаційна робота є завершеною науковою працею.

Практичне значення та практична цінність отриманих результатів. Дисертаційна робота виконана на кафедрі системного проектування КПІ ім. Ігоря Сікорського в рамках тематичного плану науково-дослідних робіт. 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.

Практичне значення одержаних результатів полягає у розробці програмного забезпечення класифікації документів, яке апробоване на наборі даних петицій до Київської міської ради.

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. У дисертації прослідковується чітка послідовність та структурованість викладення матеріалу відповідно до поставленої мети та завдань.

Дисертація написана науково-правильною мовою з використанням сучасної термінології.

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

У вступі розглядається актуальність проблеми обробки природної мови в малоресурсних умовах. Окреслюється важливість розробки ефективних методів обробки текстової інформації та зазначається, що в малоресурсних умовах виникає потреба в розробці нових методів обробки природної мови, оскільки доступність даних для навчання обмежена. Визначається мета та основні завдання дослідження, обґрунтовуються наукова та практична новизна роботи.

У першому розділі дисертації представлено огляд методів обробки природної мови, починаючи з загальних методів та переходячи до розгляду особливостей малоресурсних середовищ та відповідних методів вирішення поставленої задачі, включаючи генерацію додаткової розмітки, трансферне навчання з використанням векторних представлень, а також інші підходи. У висновках розділу визначено актуальність проблеми та напрями подальших досліджень.

Другий розділ дисертації присвячений опису запропонованого методу класифікації документів на основі векторних представлень словників, а також варіантам реалізацій його компонент. Зокрема, для векторного представлення слів описуються такі підходи як унітарне кодування, Word2Vec та FastText. В розділі представлено перхід від векторних представлень слів до векторних представлень документів. Описуються методи побудови векторних представлень графів: на основі факторизації, випадкових блукань та глибокого навчання. Представляється опис запропонованого методу, включаючи методи

злиття векторних представлень, та вказується перспектива його застосування в практичних задачах обробки природної мови.

В третьому розділі представлено постановку експерименту, реалізацію запропонованого методу, та результати порівняння якості класифікації методів обробки природної мови у малоресурсному середовищі. У розділі представлено тестовий набір даних для класифікації документів - набір петицій до Київської міської ради - та словник синонімів української мови. Наведено етапи передобробки даних, моделювання, та аналіз отриманих векторних просторів, включаючи векторні простори слів та петицій на основі методів Word2Vec та FastText. В розділі також наводяться гіперпараметри запропонованого методу, проводиться аналіз статистичної значимості отриманих результатів.

Оформлення дисертаційна робота відповідає усім необхідним вимогам 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 міжнародній науковій IEEE конференції.

У всіх публікаціях були дотримані принципи наукової доброчесності. Науковий рівень публікацій здобувача - високий, суттєвий особистий внесок прослідковується у кожній роботі.

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

Недоліки та зауваження до дисертаційної роботи.

1. Часті випадки пропущених ком.
2. Дослідження спрямоване на виділення оптимальних ознак для класифікаторів документів у малоресурсних середовищах, однак не визначено їх підмножину в множині ознак класифікації в загальному випадку.
3. Здобувач пропонує використання відстані Левенштейна для порівняння слів зі словника синонімів та документів, але не обґрунтовано вибір саме цієї метрики.
4. У роботі не повністю обґрунтовано застосоване експериментальне середовище як малоресурсне. Варто було додати характеристичні ознаки тестових наборів даних.

5. У дисертації наводяться оптимальні гіперпараметри для класифікації документів з експериментального набору даних, але не наводяться межі області пошуку при їх пошуці.
6. Варто було провести експерименти на додаткових наборах даних та у інших малоресурсних середовищах. Таким чином можна додатково підвищити достовірність результатів.

Вважаю, що висловлені зауваження не є визначальними і не зменшують загальну наукову новизну та практичну значимість результатів і не впливають на безумовно 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:

доцент кафедри
цифрових технологій в енергетиці
Національного технічного
університету України
«Київський політехнічний інститут
імені Ігоря Сікорського»,
к.т.н., доцент /_____ Світлана ШАПОВАЛОВА

Підпис гр. ЗАС...
Відділ кадрів...
підпис р-ще

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

М.П. «23» серпня 2023 року