ВІДГУК

офіційного опонента 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 природної мови представляє собою одну з найбільш вагомих та практичних галузей в області комп'ютерних наук та штучного інтелекту. Розвиток методів автоматизованої обробки текстової інформації, породженої людською соціальною діяльністю, відкриває можливості аналізувати різні аспекти людської взаємодії на семантичному рівні. Але не всі середовища мають велику кількість даних для побудови мовних моделей та систем обробки природної мови, а отже важливим напрямком дослідження у обробці природних мов є саме розробка методів, які здатні якісно працювати у малоресурсних середовищах.

Актуальність теми дисертації полягає у тому, що вирішення проблем обробки природної мови у малоресурсному середовищі та покращення методів представлення елементів мови має широке застосування у практичних системах, які працюють з малопредставленими мовами, наприклад українською. Дослідження надає цінну інформацію та описує методи, які можуть принести користь як дослідницькій спільноті обробки природних мов, так і розробникам систем на основі штучного інтелекту з обмеженими ресурсами.

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 дисертації забезпечена через:

- використання методологічного, систематичного та науково обґрунтованого підходу до вивчення актуальної літератури та аналізу попередніх досліджень;
- проведення експериментальних досліджень на наборі даних петицій до Київської міської ради;
- застосування статистичної перевірки (Хі-квадрат тест за методом МакНемара) для визначення статистичної значущості результатів експериментів;
- чітке дотримання наукової методології під час аналізу, моделювання та перевірки результатів, використання відомих методів, добре визначених метрик та стандартів, а також здатність до критичної оцінки власних підходів;
- прозорість та документацію всіх етапів дослідження, що дозволяє незалежно відтворити результати.

Новизна запропонованого підходу обґрунтована тим, що у розробленому методі використовується оригінальне поєднання векторних представлень документів та лінгвістичних елементів зі словника синонімів. Це дає приріст якості класифікації у контексті обробки природної мови у малоресурсних умовах, де інші методи, зазвичай, оперують або лише документами, або лише словниковими даними. Отримані під час експериментів результати, а також порівняння із існуючими підходами до обробки природної мови, свідчать про ефективність розробленого методу в малоресурсних середовищах.

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.

Стиль мовлення, використаний для викладення концепцій та результатів дослідження, є акуратним, структурованим та послідовним, що допомагає читачеві легко розуміти методологію дослідження та запропонований метод. Розділи та параграфи передають логічну послідовність дослідження, а використання математичних методів та експериментальних результатів допомагає підкреслити обґрунтованість та наукову цінність представлених результатів. Загалом, стиль мовлення в роботі демонструє високий рівень наукової грамотності та вміння ефективно описувати складні ідеї у науковому контексті. Застосування загальноприйнятої термінології в роботі свідчить про глибоке знання та розуміння предметної області. Автор вдало використовує фахові терміни для точного виразу концепцій та понять, що допомагає забезпечити точність та наукову достовірність викладу.

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

У вступі розкривається актуальність вдосконалення методів обробки природної мови в малоресурсних середовищах. Зазначається, що нестача даних та обмеженість експертних ресурсів у таких середовищах вимагає розробки нових підходів. Також у вступі формулюються мета та задачі дослідження, описується новизна досягнутих результатів, практична цінність розробленого методу, особистий внесок здобувача, а також апробація матеріалів роботи.

Перший розділ дисертації присвячений обробці природної мови у малоресурсному середовищі. Подається класифікація методів обробки природної мови, описуються їх особливості, переваги та недоліку у контексті малоресурсності середовища. Серед них – генерація додаткової розмітки, трансферне навчання, використання векторних представлень, багатомовні моделі мов та адаптація домену. Також детально розглядаються інші підходи, такі як змагальний дискримінатор і метанавчання. Розділ завершується висновками, у яких наголошується на важливості удосконалення вищезазначених методів, та вказуються ідеї розвитку даної області досліджень.

У другому розділі дисертації висвітлено метод класифікації документів на основі доповнення векторних представлень документів за допомогою векторних представлень словників. У ньому детально розглядаються різні методи представлення слів, такі як унітарне кодування, Word2Vec та FastText. Далі описуються методи побудови векторних представлень документів на основі представлень слів, зокрема методи мішка слів та TF-IDF. Друга половина розділу присвячена графовим векторним представленням, включаючи методи на основі факторизації, випадкових блукань та глибокого навчання. Розділ також охоплює методи злиття векторних представлень, зокрема конкатенації та зваженої суми. Розділ завершується висновками, в яких підсумовуються характеристики запропонованого методу, а також описуються різні варіанти його реалізації.

Третій розділ дисертації присвячений експериментальним результатам. В ньому детально описані експериментальні дані, включаючи набір даних для класифікації документів та словник синонімів української мови. Значна увага приділяється передобробці експериментальних даних та аналізу отриманих векторних просторів, включаючи аналіз просторів слів та просторів петицій на основі методів Word2Vec та FastText. Далі висвітлено порівняльний аналіз методів класифікації петицій та аналіз гіперпараметрів запропонованого методу з перевіркою статистичної значимості отриманих результатів. У висновках до розділу підсумовано результати експериментів та виділено можливі практичні застосування запропонованого методу.

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 були апробовані на 1 науковій фаховій конференції.

Публікації здобувача мають високу якість та повноцінно висвітлюють наукове дослідження. В кожній з них визначається методологія, описуються практичні аспекти дослідження, та, де потрібно, розкриваються експериментальні результати. Чітке дотримання принципів академічної доброчесності у всіх публікаціях підтверджує високий рівень професіоналізму та дисципліни дослідника.

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

Недоліки та зауваження до дисертаційної роботи.

1. Багатошаровий перцептрон використовується в роботі у якості класифікатора, на вхід якому подаються вектори обчислені запропонованим методом. Варто було описати детальніше чому обрано саме цей класифікатор.

2. У роботі пропонується використання функції міжрядкової відстані для пошуку відповідників між словами у документі та у словнику. Але такий пошук може бути повільним, адже потрібно обчислити значення функції для усіх елементів словника. Відповідно варто було надати аналіз швидкодії запропонованого методу.

3. У дисертації наводиться аналіз повного набору даних петицій до Київської міської ради, але не описуються статистики після поділу на набір даних для тренування та набір даних для тестування. Таким чином ускладнюється незалежне відтворення результатів.

4. Робота недостатньо описує малоресурсне середовище, у якому проводяться експерименти, а саме документи у прикладній області містобудування написані українською мовою. Таким чином незрозуміло чи дійсно ресурсів недостатньо для використання інших методів, чи вони були штучно обмежені для проведення експериментів.

5. Окрім обчислення та порівняння результуючих метрик, варто було провести аналіз помилок запропонованого методу та надати приклади успішних та неуспішних результатів класифікації петицій.

Вважаю, що висловлені зауваження не є визначальними і не зменшують загальну наукову новизну та практичну значимість результатів та 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.

Офіційний опонент:

завідувач кафедри
технологій цифрового розвитку
Державного університету
інформаційно-комунікаційних
технологій,
д.т.н., доцент /_____/ Вікторія ЖЕБКА

М.П. «18» серпня 2023 року

Підпис Вікторії Жебки, д.т.н., доц. засвідчую

Учений секретар
Державного університету
інформаційно-комунікаційних
технологій Антоніна Тюменєва