ВІДГУК

офіційного опонента доктора технічних наук, професора

Єрохіна Андрія Леонідовича

на дисертаційну роботу Демидович Інни Миколаївни

«Розвиток методів та засобів для визначення авторства україномовних текстів на основі конструктивно-продукційного моделювання»,

яку подано на здобуття наукового ступеня доктора філософії

за спеціальністю 122 Комп'ютерні науки

(галузь знань 12 Інформаційні технології)

Актуальність теми дисертаційного дослідження. Впровадження інформаційних технологій у роботи різних направленостей з природньою мовою вже давно стало суттєвим явищем, а останнім часом все більше актуальним. Насамперед це стосується вирішення науково-практичних задач пошуку низки різноманітних показників, які в достатній мірі б характеризували авторський стиль для забезпечення ефективності пошуку та встановлення авторства текстів різної довжини та стилю. Ідентифікація авторства застосовується до все більшої кількості текстів, включаючи літературні твори, розвідку, кримінальне право, цивільне право, комп'ютерну експертизу тощо.

Важливим завданням для багатьох задач в області визначення авторства текстів є робота з частинами мови та побудова моделі речення. Це є доволі складним завданням, оскільки українська мова не має чітких правил словотвору та порядку слів, на відміну від англійської, а тому тематика дисертації є актуальною.

Демидович Інна Миколаївна виявила наукову проблему, визначивши **мету дисертаційного дослідження** та часткові завдання, що дозволило системно й логічно коректно досягти мети. Для розв'язання окремих завдань були використані сучасні математичні засоби моделювання, а саме принципи й методи статистичного, семантичного та рекуррентного аналізу, а також

конструктивно-продукційне моделювання. Слід відзначити, що дослідження та пошук ефективного методу встановлення авторства природньомовних текстів проводилось протягом тривалого часу, але широке впровадження таких технологій дещо стримувалось через відсутність загального підходу, який би охоплював також і побудову формальної моделі мови.

Необхідність ефективного використання величезного обсягу інформації, її відповідний аналіз, обробка та проведення великої кількості розрахунків, обумовлює актуальність і значимість дослідження у галузі.

Наукові дослідження, які представлені в роботі, проводилися відповідно до плану держбюджетної наукової тематики «Інструментальна підтримка систем обробки природномовних документів» (2021 р. № держреєстрації 0122U002086) та «Моделювання в задачах розробки програмного забезпечення» (2021 р. № держреєстрації 0121U109167), у яких здобувачка була виконавцем та її власне дослідження стало фрагментом даної науково-дослідної роботи.

Ступінь новизни, обґрунтованості та достовірності наукових положень, висновків і рекомендацій, сформульованих у дисертаційній роботі. Наукова новизна отриманих результатів визначається теоретичним узагальненням і новим рішенням науково-практичного завдання з визначення авторства природньомонвних текстів – створенні методів та інформаційної технології інфометрії в академічному інформаційному середовищі.

Першим науковим результатом можна важати розробку метода визначення авторства тексту на основі комбінованих показників складності тексту статистичного, рекурентного аналізу та конструктивно-продукційного аналізу. Це забезпечує всебічний аналіз з врахуванням значної кількість параметрів тексту.

Другим науковим результатом стала модель природньомовного тексту у вигляді множини правил стохастичних граматик та метод порівняння текстів на основі порівняння цих правил, на відміну від існуючих моделей вона дозволяє враховувати синтаксичні та стилістичні особливості тексту автора.

Третій науковий результат полягає у тому, що в роботі розроблено метод встановлення профілю автора, який визначає найбільш суттєві показники аналізу текстів притаманні певному автору, що спрощує та полегшує подальші розрахунку та скорочує потрібний час та ресурси розрахунку;

Четвертий науковий результат полягає у встановленні статистично значимого зв'язку результатів рішення задач виявлення запозичень та встановлення авторства текстів.

Також в дисертаційній роботі вдосконалено метод багатокритеріальної оптимізації на основі генетичного алгоритму. Отримали подальшого розвитку методи конструктивно-продукційного моделювання (у частині використання параметризованих конструкторів, зв'язків між конструкторами та конструювання конструкторів), метод розпізнавання образів за критерієм мінімальної відстані (в частині формування образу текстів) та метод рекурентного аналізу для роботи з природньомовними текстами.

Обґрунтованість наукових положень, висновків і рекомендацій, сформульованих у дисертаційній роботі підтверджується: актуальністю та достатньою кількістю опрацьованих здобувачем літературних джерел, коректним формулюванням проблеми та мети дослідження, обґрунтованою постановкою експериментальних досліджень та аналізом отриманих результатів.

Достовірність отриманих в роботі результатів досліджень забезпечується коректним використанням статистичних та стілометричних методів, методу конструктивно-продукційного моделювання, методів розпізнавання образів та методів обробки природньої мови, а також полягає в експериментальному підтверджені запропонованих рішень, висновків та рекомендацій, розроблених дисертанткою.

Вищевикладене свідчить про обґрунтованість та достовірність наукових положень, висновків і рекомендацій, що викладені у дисертаційній роботі Демидович Інни Миколаївни.

Теоретичне значення дисертаційної роботи полягає в розробці моделей, методів та технологій, комплексне використання яких дозволяє автоматизувати процес аналізу природньомовних текстів, вірно інтерпретувати та застосовувати отримані показники при визначенні авторства україномовних текстів.

Практичне значення наукових положень, висновків і рекомендацій, сформульованих у дисертації. Результати дисертаційної роботи можуть бути використані для вирішення практичних завдань у галузях, де застосовуються механізми обробки природної мови, а саме у галузях лінгвістики, визначення ознак академічного плагіату, автоматичного реферування, класифікації текстових документів за стилометричними властивостями, визначення авторства тощо.

Можливість практичного використання підтверджується впровадженням результатів дослідження. Наведено використання результатів роботи у навчальному процесі при викладенні дисципліни «Ефективність інформаційних систем та комп'ютерних технологій» при підготовці аспірантів за спеціальністю 122 «Комп'ютерні науки» на кафедрі Комп'ютерних інформаційних технологій. Практичне значення роботи підтверджено відповідним Актом впровадження.

Основні результати дослідження достатньо повно викладено в публікаціях у фахових та рекомендованих Міністерством освіти і науки України для публікації результатів дисертацій (4 публікації). За результатами апробацій (за матеріалами виступів на наукових міжнародних конференціях) оприлюднено 9 робіт, у тому числі матеріалах міжнародних конференцій, що індексуються МНМБ Scopus – 3, тезах доповідей міжнародних та всеукраїнських конференцій – 6 робіт.

Оцінка змісту дисертації, її завершеності й оформлення. Оцінюючи зміст дисертаційної роботи, слід відзначити її практичну спрямованість, завершеність в цілому, внутрішню єдність матеріалу.

Дисертаційна робота написана структуровано й технічно грамотною науковою мовою та відповідає встановленим вимогам щодо оформлення. Стиль викладення матеріалів дисертаційного дослідження та наукових положень доказовий і в цілому забезпечує доступність їх сприйняття.

Дисертаційна робота складається із вступу, 4 розділів, висновків, списку використаних джерел і додатків. Загальний обсяг дисертації становить 130 сторінок, в тому числі 110 сторінок основної частини.

У вступі відображено загальну характеристика роботи, виконано обґрунтування актуальності теми, сформульовані мета і задачі дослідження, відображено наукову новизну та практичну цінність отриманих результатів і висновків, наведено дані щодо їх апробації та впровадження.

У першому розділі проведено аналіз завдань визначення авторства та атрибуції текстів та аналітичний огляд підходів до їх реалізації. Виконано огляд існуючих підходів та систем, дані яких застосовуються у процесі визначення авторства природньомовних тексті для різних мов. На основі аналізу складових процесу та факторів, що впливають на достовірність отриманих результатів, визначено задачі дослідження.

У другому розділі детально представлено розроблені методи визначення авторства. Сформовано модель представлення тексту у вигляді стилометричних властивостей тексту та множин правил стохастичних граматик. Розглянуто задачу порівняння текстів на основі цих правил. Запропоновано використання методу конструктивно-продукційного моделювання та розпізнавання образів для визначення авторства текстів та пошуку їх співпадінь.

У третьому розділі для детального розуміння задачі виявлення ознак авторського стилю наведені результати експериментальних досліджень. Перевірена та підтверджена ефективність кожного з методів та розроблених

моделей. Виконано експерименти за допомогою репрезентативних вибірків як художніх творів різних авторів, так технічних текстів різного розміру та складу. Встановлено ступінь ефективності кожного з досліджених методів окремо.

У четвертому розділі представлена практична реалізація запропонованих моделей та методів. Розроблено інструменти для автоматичного аналізу тексту, підрахунку відповідних показників та подальшого порівняння робіт за ними. Також представлені інструменти, які на основі розроблених конструкторів автоматично будують множини правил для різних текстів та порівнюють обрані на ступінь схожості.

Анотація відображає основний зміст дисертації та розкриває наукові результати та практичну цінність роботи у достатній мірі.

Зауваження до дисертаційної роботи:

1. У запропонованій інформаційній технології визначення авторства використовуються певні маркери стилю (відібрані з усіх стилометричних властивостей) та методи класифікації, але не показано, які передумови визначили вибір саме цих маркерів стилю та методів класифікації.

2. Тестування розробленого підходу проводилось на досить малій вибірці даних. Не показано, як отримані результати будуть відрізнятись на більших або інших вибірках.

3. У тексті дисертації зустрічаються окремі стилістичні неточності, орфографічні помилки та неповнота визначень.

Відповідність дисертації встановленим вимогам і загальні висновки. Вказані недоліки не впливають на загальну позитивну оцінку виконаної роботи. Дисертація є актуальною і має високу наукову цінність та практичну значущість. Під час вивчення та аналізу дисертаційної роботи випадків порушення академічної доброчесності виявлено не було.

Вважаю, що за актуальністю, новизною отриманих наукових результатів, практичною значущістю, повнотою опублікування матеріалів в статтях і доповідях на конференціях одержаних результатів дисертаційна робота «Розвиток методів та засобів для визначення авторства україномовних текстів на основі конструктивно-продукційного моделювання» відповідає вимогам «Порядку проведення експерименту з присудження ступеня доктора філософії та скасування рішення разової спеціалізованої ради закладу вищої освіти, наукової установи про присудження ступеня доктора філософії», затвердженого постановою Кабінету Міністрів України від 12.01.2022 р. № 44 з останніми змінами внесеними постановою Кабінету Міністрів України №341 від 21 березня 2022 р., а її авторка, Демидович Інна Миколаївна, заслуговує на присудження наукового ступеня доктора філософії за спеціальністю 122 - Комп'ютерні науки.

Офіційний опонент, доктор технічних наук,
професор, декан факультету комп'ютерних наук
Харківського національного
університету радіоелектроніки Андрій ЄРОХІН

Підпис професора Єрохіна А.Л. засвідчую

Проректор з наукової роботи ХНУРЕ Юрій РОМАНЕНКОВ

"31" 01 2024 р.