

ВІДГУК ОФІЦІЙНОГО ОПОНЕНТА

доктор технічних наук, професора **Теслюка Василя Миколайовича**
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.

1. Актуальність теми дослідження

Ми живемо в епоху впровадження інформаційних технологій у різні сфери людської діяльності. Також, сучасні інформаційні технології та методи і засоби штучного інтелекту широко використовуються в галузі прикладної лінгвістики для атрибуції автора за допомогою контент-аналізу, для атрибуції текстів у юриспруденції, а також для лінгвістичного аналізу тексту рекламного змісту. Особливого значення, в умовах глобалізації, набуває завдання з ідентифікації авторства текстів. Вирішити проблему непросто, оскільки кілька лінгвістичних факторів можуть збігатися. Це: фактор стилю тексту, фактор тематики та фактор авторського стилю. Проблема атрибуції авторства передбачає визначення певного стилю, певної теми та авторського стилю.

Через загальну складність поставленої задачі, наявності різної результативності використовуваних методів при роботі з різними мовами – проблема визначення авторства досі не вирішена.

Дисертаційна робота Демидович Інни Миколаївни присвячена розв'язанню актуальної наукової задачі – розробленню нових, удосконаленню існуючих методів і засобів атрибуції природньомовних текстів української мови та визначення авторства. Тематика роботи відповідає пріоритетним напрямкам розвитку науки в Україні. Робота виконувалась відповідно до плану держбюджетної наукової тематики Українського державного університету науки та технологій та відповідає науковому напрямку кафедри комп'ютерних інформаційних технологій. Актуальність дисертаційного дослідження підтверджена використанням отриманих дисертантом результатів у науково-дослідній роботі «Інструментальна підтримка систем обробки природньомовних документів» (2021 р. № держреєстрації 0122U002086) та «Моделювання в задачах розробки програмного забезпечення» (2021 р. № держреєстрації 0121U109167), у яких здобувач була виконавцем та її власне дослідження стало фрагментом даної науково-дослідної роботи.

2. Ступінь обґрунтованості та достовірності основних наукових положень, висновків і рекомендацій

Основні наукові положення дисертаційної роботи обґрунтовані та не суперечать методам проектування інформаційних систем, методам математичного аналізу та моделювання, теорії алгоритмів, принципам об'єктно-орієнтованого програмування та методам машинного навчання і штучного та обчислювального інтелекту.

Наукові результати дисертаційної роботи доповідались на шести конференціях. Результати дисертаційної роботи Демидович І. М. перевірені на практиці, що підтверджено відповідним актом впровадження.

Нові наукові результати, які отримані автором, доповнюють попередньо відомі наукові дослідження за напрямком дисертаційної роботи.

3. Наукова новизна отриманих результатів

У дисертаційній роботі Демидович І.М. отримано наступні основні наукові результати.

1. Вперше розроблено модель природньомовного тексту у вигляді множини правил стохастичних граматик та метод порівняння текстів на основі порівняння цих правил. На відміну від існуючих моделей – дана модель дає можливість враховувати синтаксичні та стилістичні особливості тексту автора.

2. Вперше розроблено метод визначення авторства тексту на основі комбінованих показників складності тексту статистичного, рекурентного аналізу та конструктивно-продукційного аналізу. Це забезпечує всебічний аналіз з врахуванням значної кількості параметрів тексту.

3. Удосконалено метод встановлення профілю автора, який визначає найбільш суттєві показники аналізу текстів притаманних певному автору, що спрощує та полегшує подальші розрахунки та скорочує необхідний час та ресурси розрахунку.

4. Удосконалено метод багатокритеріальної оптимізації на основі генетичного алгоритму.

5. Отримав подальший розвиток метод конструктивно-продукційного моделювання у частині використання параметризованих конструкторів, зв'язків між конструкторами та конструювання конструкторів.

4. Практична цінність одержаних результатів

Практичне значення полягає у розробленні програмної системи для визначення авторства україномовних текстів.

Результати експериментів визначили значення раціонального параметру – мінімальну довжину текстових фрагментів (у кількості слів), яку слід вважати запозиченням, він дорівнює п'яти словам. Це слід вважати рекомендацією для використання будь-яких програм виявлення запозичень.

Запропонований метод може використовуватись як для вирішення проблем пошуку запозичень, так і для встановлення вірогідного авторства тексту. Може використовуватись як додатковий інструмент для комплексної перевірки текстів на плагіат. Особливістю є можливість використання запропонованого метода, як самостійно для розв'язання відповідних задач, так і комплексі з іншими інструментами для покращення їх роботи та отримання більш достовірних результатів.

Запропонований підхід може використовуватися для пошуку як повних збіжностей та співпадінь у тексті, так і схожості стилів написання та побудови речення, що може вказувати на участь декількох авторів у написанні роботи.

5. Повнота викладу результатів дисертаційної роботи у наукових публікаціях

За результатами досліджень, які викладені у дисертаційній роботі, опубліковано 13 наукових праць, з них 4 статей у наукових фахових виданнях України та 3 публікацій у збірниках тез наукових конференцій (зокрема 3 – у наукометричних базах даних Scopus та Web of Science), 6 матеріалів міжнародних та всеукраїнських науково-практичних конференцій.

Опубліковані праці повною мірою висвітлюють матеріали дисертації.

6. Загальна характеристика дисертаційної роботи

В дисертаційній роботі розв'язано наукове завдання розроблення моделей, методів та програмного інструментарію для встановлення авторства природньомовних текстів, виявлення запозичень та побудови різних моделей представлення природньомовного тексту для подальшого їх аналізу та використання при визначенні авторства.

Дисертаційна робота складається із вступу, 4 розділів, висновків, списку використаних джерел і додатків. Загальний обсяг дисертації становить 130 сторінки, в тому числі 110 сторінки основної частини, 15 рисунків, 22 таблиць, 1 додаток на 1 сторінці та список використаних джерел із 148 найменувань на 14 сторінках.

У вступі подано загальну характеристику дисертаційної роботи, обґрунтовано всі процедурні положення та подано зв'язок роботи з науковими програмами, планами, темами.

У *першому розділі* дисертаційної роботи здійснено огляд літературних джерел щодо встановлення авторства природньомовних текстів на основі різних методик атрибуції тексту; проаналізовано існуючі підходи, їх особливості та недоліки. За результатом проведеного аналізу визначені основні задачі дисертаційного дослідження.

Другий розділ роботи присвячений детальному опису методів та моделей встановлення авторства природньомовних текстів. Розроблено модель представлення природньомовного тексту у вигляді правил стохастичної граматики, метод визначення авторства художніх та технічних текстів на основі цієї моделі; розроблена конструктивно-продукційна модель для автоматичного будування та порівняння цих правил для встановлення подібності текстів; розвинуто метод визначення авторства на основі декількох груп показників на основі результатів статистичного аналізу з використанням N-грамів, рекурентного аналізу, аналізу складності сприйняття тексту за допомогою розпізнавання образів; розроблено метод встановлення профілю автора для подальшого його використання у визначенні авторства текстів та пошуку їх подібностей на основі результатів статистичного аналізу з використанням N-грамів, рекурентного аналізу, аналізу складності сприйняття тексту за допомогою багатокритеріального аналізу та методу направленого випадкового пошуку глобальних екстремумів (генетичного алгоритму) для зменшення числа показників на перелік найбільш значущих для кожного з авторів;

У *третьому розділі* представлено результати експериментів та проведено порівняльний аналіз усіх запропонованих моделей та методів.

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

В подальшому, методи було об'єднано для поліпшення результатів та врахування різних особливостей авторського стилю. Було розвинуто та експериментально доведено ефективність методів роботи з великою кількістю різних показників для поліпшення результатів.

У *четвертому розділі* продемонстровано розробку інструментів для автоматичного аналізу тексту та обчислення відповідних показників. У тому числі висвітлено логіку, за якою на основі розроблених конструкторів автоматично будують множини правил для різних текстів та порівнюють обрані на ступінь схожості.

Додатки містять акт впровадження отриманих результатів в навчальний процес.

7. Недоліки та зауваження до дисертаційної роботи:

1. На мою думку, необхідно конкретизувати пункт наукової новизни “Удосконалено метод багатокритеріальної оптимізації на основі генетичного алгоритму”, а саме: додати дані щодо особливостей вдосконалення та переваг удосконалення методу. Аналогічне зауваження відноситься і до пунктів новизни, які отримали подальший розвиток.

2. В параграфі 2.4.3 “Оптимізація розпізнавання образів засобами генетичного алгоритму” варто чіткіше описати критерій, цільову функцію та обмеження оптимізаційної задачі, що значно покращило б розуміння роботи. Окрім того, відсутні дані у дисертації щодо ресурсів, які необхідні для розв’язання вищезгаданої задачі (час роботи центрального процесора, необхідний об’єм пам’яті, та інші).

3. Для кращого розуміння роботи, у другому розділі дисертації варто додати дані щодо точності, області адекватності розроблених методів і моделей та алгоритми їх застосування в процесі розв’язання практичних задач визначення авторства досліджуваного тексту (метод порівняння з еталоном теорії розпізнавання образів, ст.39).

4. У роботі варто докладніше проаналізувати використання автором знаків пунктуації, що відображають логічні паузи автора у мовленні та інтонаційні акценти на письмі, у той час як шаблон їх використання може мати велике значення для індивідуальної характеристики мовлення автора.

5. На деяких рисунках у третьому розділі дисертації (рис.3.2-3.4, рис.3.6) відсутні підписи по осях, що ускладнює їхнє розуміння. Присутні рисунки з англійськими підписами (рис.3.9, рис.4.1 – рис.4.6) – варто написати на рисунках зробити українською.

6. В четвертому розділі дисертації, на мою думку, варто додати структуру чи архітектуру розробленого програмного засобу, а також більше уваги приділити опису програмного, інформаційного, алгоритмічного та технічного забезпечень.

7. У висновках, на мою думку, доцільно додати кількісні показники переваг розроблених моделей та методів (п.1 – п.5) та чіткіше описати наукову задачу дисертаційного дослідження, яку розв’язано автором.

8. В дисертаційній роботі відсутні дані щодо впровадження отриманих наукових і практичних результатів – в наукових темах та проектах.

9. Автор дослідження досить часто використовує в роботі термін “...інформація...”, на мою думку – коректніше “...дані...”. В тексті дисертації присутні граматичні та стилістичні помилки (ст.3, ст.16 та ін.), описки (“...авторству...” – “...авторства...” ст.11, “...репрезентативних вибірок...” –

“...репрезентативних вибірок...” ст.3), не коректні вирази “...дозволяють...” – “...дають змогу...” ст.10 та ін.

Однак, ці зауваження не є принциповими, не впливають істотно на зміст дисертації та не знижують її наукову і практичну цінність.

Висновок про дисертаційну роботу

Дисертація Демидович Інни Миколаївни на тему “Розвиток методів та засобів для визначення авторства україномовних текстів на основі конструктивно-продукційного моделювання” – є завершеною науковою працею і має наукову та практичну цінність.

У дисертації розв’язано наукове завдання розроблення методів та програмного інструментарію для встановлення авторству природньомовних текстів.

З урахуванням обґрунтованості наукових положень та висновків наведених у дисертації, наукової та практичної цінності отриманих автором результатів, вважаю, що дисертаційна робота відповідає вимогам до наказу МОН України № 40 від 12.01.2017 р. “Про затвердження вимог до оформлення дисертації” та КМ від 12 січня 2022 р. №44 “Порядку проведення експерименту з присудження ступеня доктора філософії та скасування рішення разової спеціалізованої ради закладу вищої освіти, наукової установи про присудження ступеня доктора філософії” з останніми змінами внесеними постановою КМ №341 від 21 березня 2022, а її автор Демидович Інна Миколаївна заслуговує присудження ступеня доктора філософії зі спеціальності 122 – Комп’ютерні науки.

Рецензент:

Доктор технічних наук, професор,
завідувач каф. автоматизованих систем
управління Національного університету
"Львівська політехніка"

Василь ТЕСЛЮК

Підпис д.т.н., професора Теслюка В.М.

«ЗАСВІДЧУЮ»

Вчений секретар Національного Університету
«Львівська Політехніка»

Роман БРИЛИНСЬКИЙ

“5” лютого 2024 р.

