ВІДГУК

офіційного опонента
доктора технічних наук, професора **Шинкаренка Віктора Івановича**,
на дисертацію
Висоцької Вікторії Анатоліївни
«Аналіз та синтез комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту»,
подану на здобуття наукового ступеня доктора технічних наук
за спеціальністю
10.02.21 – структурна, прикладна і математична лінгвістика

1. Актуальність теми дослідження

Обсяги наукових досліджень та їх реалізація у комп'ютерних інформаційних системах з опрацювання природномовних текстів стрімко зростають. На сьогодення розроблено велика кількість комп'ютерних лінгвістичних систем (КЛС) різного призначення, зокрема для опрацювання україномовного текстового контенту.

Усі сфери життєдіяльності людини, як то наукова, технічна, соціальна, побутова, інформаційна та інші спираються на використання природної мови. Тому виникла і ще виникає величезна кількість задач автоматичного та автоматизованого опрацювання природномовного контенту.

Подальший розвиток КЛС стримується значною фрагментацією, відособленістю та закритістю існуючих систем. Ця проблема більш вагома саме для систем пов'язаних з мовами низької комп'ютерної ресурсності та своєрідністю, зокрема української.

Усвідомлення цієї проблеми дозволило започаткувати та виконати це дисертаційне дослідження.

Більш конкретно: дисертаційна робота Висоцької Вікторії Анатоліївни присвячена вирішенню важливої актуальної науково-прикладної проблеми аналізу та синтезу комп'ютерних лінгвістичних систем для розв'язання різних задач опрацювання україномовного текстового контенту, що дасть змогу підвищити рівень ресурсності природної української мови на основі розроблення нових та удосконаленні відомих моделей, методів та, як кінцевий результат, засобів NLP (Natural-Language Processing).

2. Ступінь обґрунтованості наукових положень, висновків і рекомендацій, сформульованих у дисертації, їхня достовірність і новизна

Дисертаційна робота Висоцької В.А. виконана на високому науковому рівні, є завершеною науково-дослідницькою роботою, матеріал подано в логічній послідовності, поставлені задачі сформульовано в межах визначеної науково-прикладної проблеми та глибоко опрацьовані. Аналіз змісту розділів, використаного методологічного та програмно-алгоритмічного інструментарію та способів його застосування дає підстави зробити висновок про належну обґрунтованість винесених дисертантом на захист основних наукових результатів. Наукові положення, висновки та рекомендації, сформульовані у дисертації, є обґрунтованими за рахунок проведеного теоретичного аналізу, використання відомостей і положень, отриманих з науково-технічної літератури та інформаційних ресурсів, а також вони підтверджені науковими та практичними результатами за рахунок відповідних матеріалів про впровадження дисертаційних досліджень в межах національних проектів.

3. Основні наукові результати досліджень та наукова новизна дисертаційної роботи

Маю погодитись з формулюванням наукової новизни автором, підтверджую її наявність і значимість:

вперше:

– розроблено метод ідентифікації ключових слів в україномовних текстах на основі графемного та морфологічного аналізу основ слів через регулярні вирази та N-грами;

– розроблено метод визначення стилю автора тематичного україномовного текстового контенту на основі аналізу ключових слів, стійких словосполучень, N-грам, лінгвометрії та стилеметрії;

– розроблено метод обчислення ступеня верифікації автора україномовного тексту із множини можливих на основі порівняльного аналізу стилів потенційних авторів;

– розроблено методи аналізу та синтезу КЛС на основі створення загальної типової структури системи опрацювання текстового контенту українською мовою через підтримку модульності, моделювання взаємодії основних процесів і компонентів;

одержало подальший розвиток:

– методи опрацювання інформаційних ресурсів, такі як інтеграція, управління та супровід контенту, які на відмінну від існуючих адаптовані для опрацювання україномовного тексту та враховують потреби постійної цільової аудиторії на основі аналізу історії діяльності цільової аудиторії на веб-ресурсі КЛС;

– модель лінгвістичного опрацювання текстового контенту на основі вдосконалення графемного, морфологічного, лексичного та синтаксичного аналізів, які на відмінну до існуючих адаптовані для опрацювання україномовного тексту через регулярні вирази та машинне навчання;

удосконалено:

– методи NLP, які на відмінну від існуючих реалізовані на основі розроблених регулярних виразів графемного та морфологічного аналізу україномовного тексту та модифікованого алгоритму стемінгу Портера як ефективного способу ідентифікації афіксів лем для можливості розмічування аналізованого слова;

– методи токенізації та нормалізації тексту, які на від міну від існуючих використовують каскади простих підстановок розроблених регулярних виразів узгодження з шаблонами на основі продукційних правил, скінченних автоматів та онтологічної моделі правил синтаксису української мови;

– модель інтелектуального аналізу текстового потоку, яка на відмінну від існуючої базується на процесах опрацювання інформаційних ресурсів та машинного навчання, що дало змогу адаптувати типові структури модулів інтеграції, управління та супроводу контенту, розробити конвеєр опрацювання україномовного тексту та підвищити ефективність функціонування КЛС в залежності від розв'язку конкретної задачі NLP.

4. Практичне значення результатів дисертаційної роботи

Практично цінними є такі результати:

– застосування методу ідентифікації стійких словосполучень при визначенні ключових слів в україномовних наукових текстах технічного профілю дозволяє підвищити точність пошуку ключових слів;

– розроблення формального підходу до проектування модуля контент-моніторингу для ідентифікації ключових слів в україномовних текстах на основі видобування веб-даних, ОПМ та лексичного аналізу визначених слів текстового контенту, що дозволило розробити загальну структуру типових КЛС та підвищити ефективність функціонування КЛС;

– застосування методу обчислення ступеня верифікації автора україномовного тексту на основі аналізу стилів потенційних авторів дозволило підвищити точність ідентифікації та провести декомпозицію методу через дослідження коефіцієнтів стилістики як зв'язність мовлення, ступінь синтаксичної складності, лексична різноманітність, індекси концентрації та винятковості тексту;

– розроблення модуля контент-моніторингу для ідентифікації потенційного автора тексту із множини можливих на основі порівняння

результатів аналізу шаблонного авторського тексту з досліджуваним для зменшення обсягу відповідної множини;

– експериментальна апробація методу ідентифікації стилю автора в україномовних текстах на основі видобування веб-даних та лексичного аналізу визначених стопових слів, що дозволяє виділити множину потенційно подібного за стилем контенту з множини потенційних авторських публікацій.

Практичне значення дисертаційної роботи засвідчується використанням результатів дослідження:

– у низці науково-дослідних роботах кафедри інформаційних систем та мереж Національного університету «Львівська політехніка»;

– у навчальному процесі під час викладання низки дисциплін у Національному університеті «Львівська політехніка» та Національному технічному університеті «Харківський політехнічний інститут».

5. Рекомендації щодо використання результатів дисертації

Розроблені моделі, методи та алгоритми можуть бути використані при розробці нових чи вдосконалені вже існуючих комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту для розв'язку різних NLP-задач, а також при проектуванні подібних систем згідно потреб постійної/потенційної цільової аудиторії на основі аналізу історії їх дій на Web-ресурсі комп'ютерної лінгвістичної системи. Також результати функціонування подібних комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту дозволить підвищити рівень ресурсності української мови.

6. Повнота викладення наукових положень, висновків і рекомендацій в опублікованих працях

Наведені в дисертації положення та результати в повній мірі опубліковані у 254 наукових публікаціях, із них 143 статті, де 4 статті опубліковано в журналах з квартилем Q2 відповідно до SCImago Journal, 85 включено до Scopus або Web of Science та 100 публікацій у матеріалах конференцій (зокрема, 64 із них включено до Scopus або Web of Science). Опубліковано 9 монографій та 2 розділи монографії, які включено до Scopus.

Що свідчить про надзвичайну продуктивність дисертантки.

У наукових працях повною мірою подано всі розділи рецензованої дисертації.

Зміст реферату повністю відображає основні положення дисертації.

7. Оцінка змісту дисертаційної роботи, її завершеність

Дисертаційна робота складається з анотацій, вступу, шести розділів, висновків, списку використаних джерел з 1044 назв на 52 сторінках та 6 додатків на 82 сторінках, 179 рисунків, 62 таблиці. Повний обсяг дисертації – 480 сторінок, у тому числі 306 сторінок основного тексту.

У **вступі** проведено аналіз проблеми опрацювання україномовного текстового контенту та розроблення комп'ютерних лінгвістичних систем для розв'язку різних задач опрацювання української мови. Також автором обгранковується актуальність теми, сформульована мета та основні задачі досліджень, описується короткий зміст дисертаційної роботи.

У **першому розділі** проаналізовано існуючі практичні рішення, що відображають основні підходи опрацювання високоресурних породних мов у відомих комп'ютерних лінгвістичних систем для розв'язку різних задач, що дало можливість вдосконалити загальну класифікацію відповідних систем. Проведено порівняльний аналіз класичних методів та визначені обмеження, які виникають при опрацюванні низькоресурних мов, зокрема українськомовного текстового контенту. За результатами аналізу сформульовані основні етапи та напрями досліджень, що дало змогу досягнути загальної мети дисертаційної роботи – підвищення рівня ресурсності природної української мови шляхом розроблення методології побудови комп'ютерних лінгвістичних систем на базі нових та удосконалення відомих методів опрацювання україномовного текстового контенту. Обгрунтовано актуальність розв'язання проблеми аналізу та синтезу комп'ютерних лінгвістичних систем на основі розроблення загальної структури системи опрацювання україномовного текстового контенту, яка за рахунок взаємодії основних процесів/компонентів системи та адаптованих до української мови методів лінгвістичного опрацювання текстового контенту на основі графемного, морфологічного, лексичного, синтаксичного, семантичного, структурного, онтологічного та прагматичного аналізу дозволила вдосконалити інформаційну технологію інтелектуального аналізу текстового потоку для розв'язку конкретної задачі опрацювання природної мови.

У **другому розділі** вдосконалено методи опрацювання інформаційних ресурсів як інтеграція, управління та супровід україномовного контенту, що дозволило адаптувати процес інтелектуального аналізу текстового потоку та розробити метрики ефективності функціонування комп'ютерних лінгвістичних систем для до розв'язку різних задач. Розроблені методи та засоби дають можливість будувати комп'ютерні лінгвістичні системи опрацювання україномовного текстового контенту згідно потреб постійної/потенційної цільової аудиторії на основі аналізу історії дій користувачів веб-сайту. На основі аналізу вхідних/вихідних потоків контенту комп'ютерної лінгвістичної системи визначені та сформульовані функціональні вимоги до проекту

подібних систем, їх програмних модулів, мережних, програмних та технічних інструментів програмної реалізації.

У **третьому розділі** удосконалено інформаційну технологію інтелектуального аналізу текстового потоку на основі опрацювання інформаційних ресурсів, що дало змогу адаптувати загально типову структуру модулів інтеграції, управління та супроводу контенту для розв'язку різних задач опрацювання природної мови та підвищити ефективність функціонування комп'ютерних лінгвістичних систем на 6-9%. Це стало можливим завдяки поєднанню адаптованих до української мови методів лінгвістичного аналізу, вдосконаленої інформаційної технології опрацювання інформаційних ресурсів, машинного навчання та множини метрик оцінювання ефективності функціонування комп'ютерних лінгвістичних систем. Основний принцип побудови таких комп'ютерних лінгвістичних систем полягає на модульності, що полегшує їх побудову згідно вимог щодо наявності відповідних процесів для розв'язку конкретної задачі опрацювання природної мови.

У **четвертому розділі** розроблено методи опрацювання природної мови на основі регулярних виразів узгодження з шаблонами, що дало змогу адаптувати методи токенізації та нормалізації тексту каскадами простих підстановок регулярних виразів та кінцевих автоматів. Удосконалено метод морфологічного аналізу україномовного тексту на основі сегментації та нормування слова, сегментації речення та модифікованого алгоритму стемінгу Портера як ефективного засобу ідентифікації афіксів лем для можливості розмічування аналізованого слова, що дало змогу підвищити точність пошуку ключових слів на 9%.

У **п'ятому розділі** розроблено метод визначення стійких словосполучень на основі ідентифікації ключових слів україномовного тексту та аналізу коефіцієнтів лексичного мовлення автора тексту в еталонних уривках контенту, що дало можливість на основі статистичної лінгвістики покращити точність методу визначення стилю автора тексту на 9%. Достовірність наукових та практичних результатів підтверджується відповідними матеріалами про впровадження дисертаційних досліджень, а також за рахунок порівняння одержаних практичних результатів на різних вибірках достовірних вхідних даних. Експериментальне дослідження підтвердило достовірність методу визначення ключових слів – для різних алгоритмів опрацювання первинного тексту середній збіг списків виявлених ключових слів з авторськими змінюється у проміжку 52,6-68,5%. Точність збігу ключових слів із авторськими коливається в проміжку 43,6-62,9%. Середній збіг змістовних ключових слів порівняно зі всіма знайденими системою коливається в проміжку 38,9-75,8% в залежності від етапів аналізу текстів статей. Точність

збігу ключових слів порівняно зі всіма знайденими системою коливається в проміжку 34,3-71,9% в залежності від етапів аналізу текстів статей.

У **шостому розділі** розроблено метод визначення автора в україномовних текстах на основі аналізу коефіцієнтів лексичного авторського мовлення в еталонному уривку авторського тексту, який ґрунтується на аналізі колекції ключових слів, стійких словосполучень, показників лінгвометрії, стилеметрії, а також результатів аналізу N-грам на основі порівнянь різниць вживання 2-грам та 3-грам для подібних за стилем публікацій в межах [6;7]%, а для точно не подібних – >12%), що забезпечило можливість визначити множину потенційних авторів публікацій з більш ніж одного автора (до [9;34]% із загальної кількості учасників проекту) та розробити метод ідентифікації авторського стилю.

У **висновках** подано перелік одержаних в роботі основних загальних результатів дисертаційних досліджень.

У **додатках** наведено додаткові матеріал у вигляді рисунків, табличних значень, регулярних виразів морфологічного аналізу українських іменників та дієслів, дерево закінчень українських слів та акти впровадження дисертаційної роботи.

Загалом дисертаційна робота написана на високому професійному рівні. Результати досліджень обґрунтовані, подані структуровано та послідовно. Реферат відповідає змісту дисертаційної роботи.

8. Зауваження до дисертаційної роботи

Перелік основних закувань до дисертаційної роботи такий:

1. Якість усіх інформаційних систем надзвичайно важлива. У першому розділі у розрізі КЛС вона й розглядається. Але чомусь показники якості наведені лише для вузького класу задач КЛС, а саме кластеризації текстового контенту.

2. У другому розділі наведений проєкт типової комп'ютерної лінгвістичної системи. Не зрозуміла чому з узагальненого проєкту виконана конкретизація системи саме такого типу. Він характеризується використанням користувачем Web доданків. Далі є ще глибша конкретизація до комерческих систем. Але ж існують інтелектуальні агенти, роботи в якості користувачів і носіїв КЛС.

3. Наведені у (2.10) - (2.39) моделі модулів КЛС є багато параметричними, до того ж усі параметри є функціями часу. Не зрозуміло як вирішувати задачі аналізу та оптимізації у такому випадку.

4. У роботі наявні неузгодженість позначок, навіть у досить близьких частинах тексту. Так, X у (3.1) – множина одночасно інтегрованого контенту з

Інтернет-ресурсів, у (3.2) – вхідні дані в КЛС з різних джерел інформації, у (3.5) – вхідний текстовий масив даних, Y у (3.5) – кортеж вихідного опрацьованого тексту, а у (3.6) – основний процес лінгвістичного аналізу та т.і.

У моделі (3.5) X мабуть слід розглядати як множину потоків вхідних текстових масивів даних.

5. У третьому розділі наведено не моделювання комп'ютерної лінгвістичної системи опрацювання української мови, а готова модель (моделювання – це процес).

6. Є питання щодо даних на яких ґрунтується експериментальна частина дослідження. Так на рис 4.1 наведені дані щодо еталону та двох уривків. Не зрозуміло який обраний еталон та уривки. Що буде якщо обидва уривки з одного твору. Необхідно більш детально було б описати статистичну та параметричну вибірку вхідної інформації для методу визначення стилю авторі та для методу обчислення ступеня верифікації автора науково-технічної публікації.

7. Для методу визначення стилю автора застосовані лінгвістичні коефіцієнти лексичної різноманітності стилістики тексту. Варто було б дослідити це питання більш детально та доповнити множину коефіцієнтів новими метриками, властивими лише для української мови або слов'янських.

8. Метод кількісної оцінки визначення авторства текстового контенту застосовує результати стилеметричого аналізу авторського тексту з порівнянням з еталоном на основі списку Сводеша, та не обґрунтовані мета та отримані результати використання цього списку.

9. У багатьох місцях наведені статистичні результати, наприклад у непронумерованих таблицях стор. 263-264. Вони носять ймовірнісний характер. Хотілося б бачити точність цих вимірів у вигляді довірчих інтервалів або у іншому.

Перелік цих зауважень суттєво не впливає на основні наукові результати та висновки дисертації.

9. Загальні висновки

Отже, дисертація Висоцької Вікторії Анатоліївни на тему «Аналіз та синтез комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту» є завершеним науковим дослідженням, в якому вирішено науково-прикладну проблему аналізу та синтезу комп'ютерних лінгвістичних систем для розв'язання різних задач опрацювання україномовного текстового контенту на основі розроблення нових та удосконаленні відомих моделей, методів та засобів опрацювання природної мови.

Дисертаційна робота відповідає паспорту спеціальності 10.02.21 – структурна, прикладна і математична лінгвістика у частині формули спеціальності та низки напрямів досліджень, зокрема: автоматизовані та автоматичні системи маркування лінгвістичних структур текстів (графемний, морфологічний, лексичний, семантичний, синтаксичний, концептологічний аналіз тощо), обчислювальна, статистична та квантитативна лінгвістика, лінгвістичне забезпечення інформаційних систем, взаємодія структурної, прикладної та математичної лінгвістики та суміжних наук.

Основні результати роботи повністю викладені в опублікованих працях, що пройшли належну апробацію у науково-дослідних темах, конференціях та семінарах.

Дисертаційна робота виконана на високому рівні. За актуальністю розв'язаних задач, обсягом досліджень, науковим рівнем і практичною цінністю отриманих результатів відповідає вимогам п. 7 та 9 Порядку присудження та позбавлення наукового ступеня доктора наук, затвердженого постановою Кабінету Міністрів України від 17 листопада 2021 року № 1197, а її автор, Висоцька Вікторія Анатоліївна, заслуговує на присудження наукового ступеня доктора технічних наук за спеціальністю 10.02.21 – структурна, прикладна і математична лінгвістика.

Офіційний опонент,
професор кафедри комп'ютерних
інформаційних технологій факультету
комп'ютерних технологій і систем
Українського державного університету
науки і технологій МОН України,
доктор технічних наук, професор

Особистий підпис
Шинкаренко В. І.
засвідчую