ВІДГУК

офіційного опонента доктора технічних наук, професора
Стрижака Олександра Євгенійовича на дисертацію
Висоцької Вікторії Анатоліївни
«Аналіз та синтез комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту»,
яка подана на здобуття наукового ступеня доктора технічних наук за спеціальністю 10.02.21 – структурна, прикладна і математична лінгвістика

Актуальність теми дослідження. Інтернет, мобільні додатки, інформаційні системи, соціальні мережі – навколо нас постійно присутні бездонні джерела інформації. Це з одного боку допомагає вирішувати багато буденних та професійних задач, але з іншого боку ускладнює процес життя із-за необхідності орієнтування в цьому хаосі інформаційного простору. Крім того це є джерело маніпулювання свідомістю людей через пропаганди, фейки як в повсякденності (наприклад через рекламу), так і в інформаційній війні тощо. Тому на сьогодення дуже є актуальним відомий вираз англійського банкіра, бізнесмена і фінансиста Натана Ротшильда «Хто володіє інформацією – той володіє світом», якому вже понад два століття. Сучасне століття є епохою ІТ та штучного інтелекту, які оточують пересічну людини всюди в повсякденності. А там де є людина, там і є природня мова. Тому поєднання інформаційних технологій, штучного інтелекту та опрацювання природної мови є актуальним в людському суспільстві для вирішення повсякденних задач. Розв'язок таких задач покладений на сучасний молодий науковий напрям як комп'ютерна лінгвістика. Складність полягає не лише в розв'язку не типових NLP-задач, але в адаптації або створенні нових моделей, методів та технологій опрацювання конкретної природної мови. Кожна природня мова є унікально, зі своїм колоритом правил, історії, граматики, виключень та особливостей генерування лінгвістичних одиниць для передачі сенсу. Людина в середньому вчить 10-15 років для розуміння повсякденності, ще 10-15 років для адаптації до професії, а саму мову та її глибину може вчити та досліджувати ціле життя. Для автоматизації процесів опрацювання конкретної природної мови такої розкоші як час немає. Крім того обмеженість у фінансуванні подібних проектів або взагалі їх відсутність, конкретність з відомими компаніями та присутність на ринку їх розроблених комерційних проектів значно скорочує вмотивованість роботи науковців в цьому напряму. Зазвичай кожний успішний проект розроблення КЛС призначений під конкретну задачу та одночасно є

одноразовим та закритим (наприклад, Amazon Alexa, Google Assistant, Facebook, Voice Mate, Bixby, Siri, Abby Lingvo, Microsoft Cortana, Microsoft Word, Grammarly, Google Translation, PROMT, CuneiForm, Trados, OmegaT, Wordfast, Dragon, IBM via voice, Speereo, Finereader, Tesseract, OCRopus тощо) без можливості ознайомитися з вмістом бажаючим ІТ-фахівцям/спеціалістам. Досить рідкі випадки, коли до таких проектів надають відкритий доступ та можливість ознайомитися з його структурою та іншим змістом. Створення будь-якого NLP-додатку для довільної природної мови із відомих понад 7000 мов та діалектів базується саме на досліджених даних (великих текстових одномовних/ паралельних корпусів з понад сотень мільйонів слів та лінгвістичних ресурсів) конкретної мови. І лише біля 20 природних мов (англійська, китайська, іспанська та інші західноєвропейські мови, японська тощо) мають відповідні результати досліджень та відповідають вимогам розроблення різної складності КЛС. Нажаль в сучасних реаліях українська мова вважається в міжнародному науковому суспільстві екзотичною мовою з низьким показником ресурсності, тобто не має достатньо навчальних, дослідницьких та опрацьованих даних для розроблення сучасних NLP-додатків при задоволені відповідних потреб суспільства, зокрема, в кібербезпеці (виявлення фейків та пропаганди, так званих тролів/ботів в соціальних мережах тощо), соціології (аналіз динаміки зміни громадської думки на певні тематичні питання тощо), філології (автоматичне дослідження великих масивів даних різного тематичного спрямування та різних часових періодів), психології (аналіз психологічного портрету особи за дописами в соціальних мережах, ідентифікація посттравматичного стресового розладу в учасників бойових дій або окупації тощо) та в інших важливих галузях сучасної України. Означене обумовлює актуальність теми дисертаційного дослідження.

Дисертаційна робота Висоцької В.А. присвячена науковій проблемі математичного моделювання процесів опрацювання текстових потоків україномовного контенту, що дало можливість підвищити рівень ресурсності природної української мови на основі розроблення методології аналізу та синтезу комп'ютерних лінгвістичних систем (КЛС) для розв'язку різних задач опрацювання природної мови. Вдосконалена технологія інтелектуального опрацювання україномовного текстового контенту на відміну від існуючих підтримує принцип модульності типової архітектури КЛС для розв'язку конкретної типової задачі опрацювання природної мови та аналізу множини параметрів та метрик ефективності функціонування системи відповідно до поведінки цільової аудиторії. Це дозволило розробити загальну типову структуру КЛС та концептуальну схему/модель функціонування типової КЛС на основі моделювання взаємодії основних процесів і компонентів.

Удосконалення методів опрацювання інформаційних джерел (ресурсів), таких як інтеграція, управління та супровід україномовного контенту, дозволило адаптувати процес інтелектуального аналізу текстового потоку до вирішення різноманітних задач опрацювання природної мови.

Актуальність обраної теми підтверджується зв'язком дисертаційної роботи з важливими програми та темами, зокрема, «Дослідження, розроблення і впровадження інтелектуальних розподілених інформаційних технологій та систем на основі ресурсів баз даних, сховищ даних, пристроїв даних та знань з метою прискореного формування інформаційного суспільства» кафедри інформаційних систем та мереж Національного університету «Львівська політехніка». Дисертація виконана в межах науково-дослідної роботи цієї кафедри «Розроблення інтелектуальних розподілених систем на основі онтологічного підходу з метою інтеграції інформаційних ресурсів» (№ 0115U004228, терміни: 05.2015–12.2017 рр.), держбюджетної науково-дослідної роботи «Методи та засоби функціонування систем підтримки прийняття рішень на основі онтологій» (№ 0118U000269, терміни: 01.2018–12.2019 рр.), а також держбюджетної науково-дослідної роботи «Система підтримки прийняття рішень розпізнавання мультиспектральних образів на основі технологій машинного навчання та онтологічного підходу» (№ 0120U102203, терміни: 04.2020–12.2021 рр.).

Ступінь обґрунтованості наукових положень, висновків і рекомендацій, сформульованих у дисертації, їхня достовірність. Обґрунтованість і достовірність отриманих наукових положень, висновків та рекомендацій забезпечується використанням перевірених підходів, підтверджених теорією та практикою, зокрема, для досягнення поставленої мети використано: теорію формальних граматик та автоматів, теорію множин, теорію моделей даних та знань, теорію ймовірності і математичної статистики, теорію моделей, алгоритмів та логіко-лінгвістичних числень, теорію інформації, теорію графів та методи подання знань для моделювання процесів опрацювання україномовного текстового контенту та розроблення модулів машинного навчання; моделі та методи опрацювання та аналізу текстового контенту для реалізації процесів розв'язку відповідних NLP-задач; методи об'єктно-орієнтованого та системного аналізу і проектування – для проектування та розроблення комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту; теорію реляційних баз даних, методи штучного інтелекту, об'єктно-орієнтоване програмування – для програмної реалізації розроблених моделей, методів та алгоритмів функціонування КЛС опрацювання україномовного текстового контенту.

Наукова новизна дисертаційної роботи. Ознайомлення із змістом та результатами дисертаційної роботи дає змогу зробити висновок, що дисертанткою запропоновано модель, методи та засоби опрацювання україномовного текстового контенту. Сформульовано та вирішено важливу науково-прикладну проблему аналізу та синтезу КЛС 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 засобів опрацювання природної мови. Отримано такі нові наукові результати:

вперше:

– розроблено метод ідентифікації ключових слів в україномовних текстах на основі графемного та морфологічного аналізу основ слів через регулярні вирази та N-грами, що дало змогу підвищити точність пошуку ключових слів, здійснити пошук стійких словосполучень та рубрикацію контенту;

– розроблено метод визначення стилю автора тематичного україномовного текстового контенту на основі аналізу ключових слів, стійких словосполучень, N-грам, лінгвометрії та стилеметрії, що дало змогу визначити стилістичний вклад кожного з авторів та підвищити точність атрибуції науково-технічної публікації;

– розроблено метод обчислення ступеня верифікації автора україномовного тексту із множини можливих на основі порівняльного аналізу стилів потенційних авторів, що дало змогу підвищити точність класифікації за подібністю стилю;

– 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, що дало можливість розширити колекцію розв'язків різних типових задач NLP шляхом реалізації типового програмного забезпечення таких систем;

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, що дало можливість сформувати множину метрик та показників ефективності функціонування КЛС для розв'язку різних задач NLP;

– 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, дала змогу адаптувати процеси опрацювання україномовного текстового контенту та підвищити точність отриманих результатів в залежності від конкретної задачі NLP;

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, що дало змогу оптимізувати процес та покращити точність сегментації/нормування українського слова/речення;

– 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, що дало змогу адаптувати алгоритми лексичного та синтаксичного аналізів для опрацювання україномовного контенту;

– 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.

Значення одержаних результатів для науки і практики. Практичну цінність роботи підтверджує впровадження її результатів у дослідженнях, проведених в межах проектів, а також використання її результатів при виконанні держбюджетних тематик Національного університету «Львівська політехніка». Використання розроблених в дисертації методів та засобів та інших результатів дозволило:

– підвищити точність пошуку ключових слів на 9%, здійснити пошук стійких словосполучень та рубрикацію україномовного текстового контенту на основі аналізу коефіцієнтів лексичного мовлення автора в еталонних уривках контенту;

– підвищити точність ідентифікації стилю автора тематичного україномовного текстового контенту на основі аналізу ключових слів, стійких словосполучень, N-грам, лінгвометрії та стилеметрії на 6%;

– підвищити точність класифікації за подібністю стилю на 7% при обчисленні ступеня верифікації автора україномовного текстового контенту із

множини можливих на основі порівняльного аналізу стилів потенційних авторів;

– розробити інформаційну технологію аналізу та синтезу КЛС на основі розроблення загальної типової структури КЛС опрацювання текстового контенту українською мовою, моделювання взаємодії основних процесів та компонентів, що дало можливість розширити колекцію розв'язків різних типових NLP-задач шляхом реалізації типового програмного забезпечення таких систем;

– підвищити точність отриманих результатів опрацювання україномовного текстового контенту на основі графемного, морфологічного, лексичного та синтаксичного аналізу для розв'язку конкретної NLP-задачі на 6-9% в залежності від конкретної NLP-задачі;

– розробити інформаційну технологію інтелектуального аналізу текстового потоку на основі опрацювання інформаційних ресурсів (джерел), що дало змогу адаптувати загальну типову структуру модулів інтеграції, управління та супроводу контенту та підвищити ефективність функціонування КЛС на 6-9% в залежності від розв'язку конкретної NLP-задачі.

Рекомендації щодо використання результатів дисертації.

Використання запропонованих підходів щодо опрацювання україномовного текстового контенту дасть змогу одержати результати з вищою точністю в різних задачах опрацювання природної мови (наприклад, машинний переклад, ідентифікація плагіату/рерайту, рубрикація тексту, аналіз атрибуції тексту, інформаційний пошук, реферування, голосові помічники, інтелектуальні чат-боти тощо) для різних галузей науки, зокрема, такі прикладні додатки використовуються для побудови КЛС в соціології (аналіз динаміки зміни громадської думки на тематичні питання), національній безпеці (інформаційна війна), кібербезпеці (виявлення фейків та пропаганди, так званих тролів/ботів в соціальних мережах), психології (аналіз психологічного портрету особи, ідентифікація посттравматичного стресового розладу учасників бойових дій або окупації), юриспруденції (криміналістика та судова справа), філології (автоматичне дослідження великих масивів даних різного тематичного спрямування та різних часових періодів), соціальних комунікаціях (аналіз дописів спільнот в соціальних мережах) та в інших важливих галузях сучасної України. Означене обумовлює актуальність теми дисертаційного дослідження.

Конкретні шляхи використання запропонованих в дисертації Висоцької В.А. моделей, методів, алгоритмів та програмних засобів полягають у підвищенні точності пошуку рішень на базі модифікованого опрацювання україномовного текстового контенту під час вдосконаленого лінгвістичного та інтелектуального аналізу інформаційних потоків даних та розробленні якісних

комп'ютерних лінгвістичних системи для розв'язку різних задач опрацювання природної мови.

Повнота викладу в опублікованих працях. Результати дисертаційного дослідження опубліковано у 254 наукових публікаціях, із них 143 статті (85 із них включено до Scopus або Web of Science), 100 публікацій у матеріалах конференцій (зокрема, 64 із них включено до Scopus або Web of Science). Загалом опубліковано 9 монографій та 2 розділи монографії, які включено до Scopus, а також 8 навчальних посібників.

Оцінка змісту дисертаційної роботи, її завершеність. На основі аналізу відомих підходів до опрацювання природної мови визначено об'єкт та предмет досліджень, постановку проблеми дослідження, розроблено модель комп'ютерної лінгвістичної системи опрацювання україномовного текстового контенту для розв'язку різних задач опрацювання природної мови, проведено аналіз специфіки та методів побудови комп'ютерної лінгвістичної системи шляхом їх систематизації за методами реалізації та функціонування, розроблено інформаційні технології опрацювання інформаційних ресурсів як інтеграція, управління та супровід українськомовного контенту, розроблено методи та алгоритми опрацювання україномовного текстового контенту, розроблено методи та засоби інтелектуального аналізу текстового контенту, розроблено уніфікований підхід до типової побудови комп'ютерної лінгвістичної системи опрацювання україномовного тексту, створено програмні модулі опрацювання україномовного текстового контенту для розв'язку різних задач опрацювання природної мови та проведення експериментів.

У **вступі** виконано аналіз наукової проблеми, обґрунтовано актуальність досліджень, визначено наукову новизну та практичну значимість досліджень, розглянуто загальну характеристику роботи, особистий внесок здобувача в публікації, виконані у співавторстві, апробацію роботи, кількість та якість публікацій, загальна структура та обсяг роботи.

У **першому розділі** проведено аналіз сучасного стану та перспективи розвитку інформаційних технологій опрацювання природної мови, що дало змогу визначити проблему та задачі дослідження, а також сформувати загальні напрями дослідження при відсутності некомерційних комп'ютерних лінгвістичних систем з відкритим кодом для опрацювання україномовного текстового контенту та стандартизованого підходу проектування. 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. Визначені обмеження існуючих методів опрацювання україномовного текстового контенту та розроблення комп'ютерних лінгвістичних систем для розв'язку різних задач опрацювання природної мови, що сформувало актуальну науково-прикладну проблему 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 засобів опрацювання природної мови. Обґрунтовано 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** проаналізовано особливості проектування та розроблення комп'ютерних лінгвістичних систем на основі визначення основних етапів як графемний, морфологічний, лексичний, синтаксичний семантичний аналіз/синтез україномовного тексту для розв'язку конкретної задачі опрацювання природної мови. Зроблена та конкретизована постановка проблеми опрацювання україномовного тексту на основі визначення функціональних особливостей інтелектуального аналізу текстового потоку. Здійснено загальний аналіз проблеми аналізу україномовного тексту та визначення основних проблем опрацювання україномовного тексту дало можливість сформулювати основні етапи та вимоги до проекту типової комп'ютерної лінгвістичної системи розв'язку конкретної задачі опрацювання природної мови. Ідентифікація основних характеристик комп'ютерних лінгвістичних систем та обґрунтування реалізації проекту типової комп'ютерної лінгвістичної системи дало можливість визначити очікувані ефекти від відповідної реалізації проекту. 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** розроблено методи аналізу та синтезу комп'ютерної лінгвістичної системи на основі створення загальної типової структури системи опрацювання текстового контенту українською мовою через підтримку модульності, моделювання взаємодії основних процесів і компонентів, що дало можливість розширити колекцію розв'язків різних типових задач опрацювання природної мови шляхом реалізації типового програмного забезпечення таких систем. 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** Удосконалено 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** здобувач навела одержані в роботі результати, відзначила новизну та практичну значимість дисертаційних досліджень.

У **додатках** наведено основні регулярні вирази морфологічного українських іменників та дієслів, дерево закінчень слів в українській мові, критерії графемного вхідного тексту, правила класифікації графем у вигляді послідовності символів, класифікація алгоритмів стемінгу лексем природної мови, лінгвістичні характеристики деяких класів морфем основ дієслів, основні правила формування українських дієприкметників, морфонологічні правила, аналіз граматичних/морфологічних ознак української/ англійської мов, аналіз синтаксичних/семантичних ознак української/англійської мов, список за рейтингом частоти появи стійких словосполучень для 3 випадкових статей, відмінності методів за рейтинговим списком із 100 стійких словосполучень,

відмінності інших методів за рейтингом частоти появи стійких словосполучень, абсолютні та відносні частоти появи стопових слів в уривку та еталоні, результат роботи алгоритму аналізу стилю автора публікації, список публікацій здобувача, а також подано результати впровадження дисертаційних досліджень.

Недоліки і зауваження

Дисертаційна робота не лишена певних недоліків. У якості зауважень слід зазначити:.

1. При розгляду процесів інтеграції інформаційних ресурсів, потоків та процесів (враховуючи семантичні) не визначено достатньо функціональний вплив на ці процеси таксономій та онтологій. Не враховано також широко застосовуваний на сьогодення при інтеграції різних документів «принцип відкритих таксономій»

2. При визначенні критеріїв оцінки ефективності КЛС (розділ 1.6) та ефектів реалізації проекту типової КЛС (підрозділ 2.3.3) не враховано характеристики швидкості обробки одиниці інформації та питомих витрат на обробку цієї одиниці інформації.

3. В тексті вступу та висновках присутнє скорочення ОПМ, а по тексту дисертації NLP, хоча це одне і те ж саме двома різними мовами.

4. Частина рисунків (скринів) поганої якості, зокрема 1.4, 2.9-2.10, 4.6-4.16 тощо.

5. В тексті роботи присутньо багато скорочень, абревіатур, англомовних термінів та комп'ютерного сленгу, що ускладнює процес розуміння змісту.

6. Розділ 4 «Архітектура комп'ютерної лінгвістичної системи опрацювання контенту українською мовою» некоректно названий. Розділ більше присвячений не архітектурі системи, а модифікованому процесу опрацювання природньої мови, адаптованому для української мови через реалізацію конвеєра аналізу інформаційного потоку на основі машинного навчання. Архітектура подібних систем була описана більш детально в третьому розділі. Аналогічно в назві розділу 6 треба було додати інформаційна технологія замість технологія (бо існує наприклад виробнича технологія).

7. Формули (4.52) та (4.53) є за змістом ідентичними.

8. Розроблений метод ідентифікації ключових слів україномовного контенту на основі аналізу появи основ слів певної частини мови, а не самих слів. Доречно було в навести в розділі 5 порівняння результатів аналізу текстів науково-технічних статей при ідентифікації ключових слів на основі основ слів та на основі слів для наочності подання експериментальної апробації.

9. В розділі 6 не зрозуміло, чи змінюються коефіцієнти персоналізованих ознак авторського стилю в залежності від тематики статті

конкретного автора та часу написання. Аналогічно, чи суттєво впливає зміна тематики та часу написання статті конкретного автора на динаміку розподілу N-грам літер, основ слів, стійких словосполучень, службових слів тощо.

10. Рисунки 6.29-6.35 є змістовно подібними на ті, що наведені в розділах 3.2.5 «Синтаксичний аналіз української мови» та 4.3 «Метод морфологічного аналізу української мови».

11. Текст роботи містить незначні граматичні та стилістичні помилки.

Проте, вказані зауваження жодною мірою не знижують високого рівня наукових результатів отриманих здобувачкою і викладених нею у дисертаційної роботі, абсолютно не впливають на їх якість та практичну цінність, й носять вузький локальний характер.

Висновок про відповідність дисертації встановленим вимогам. Дисертація Висоцької В.А. «Аналіз та синтез комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту» є завершеною науковою працею, в якій отримані нові науково обґрунтовані результати, що в сукупності вирішують поставлену 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, що відображає її високий теоретичний та практичний рівень реалізації. Вона містить не захищені раніше наукові положення і нові науково обґрунтовані результати в галузі інтелектуальних систем оброблення інформації. Плагіат відсутній.

Тема роботи є актуальною, результати, що отримані є новими та відповідають поставленим цілям та завданням й мають безсумнівне наукове та практичне значення. Оформлення дисертації та реферату акуратне і наочне та відповідає вимогам МОН України. Стиль викладання результатів є послідовним, лаконічним та логічним. Реферат у повній мірі відображає зміст, результати і висновки дисертаційного дослідження. Публікації автора, що наведені у дисертації та рефераті, досить повно висвітлюють основні наукові результати дисертаційного дослідження.

Дисертаційна робота Висоцької В.А. відповідає паспорту спеціальності 10.02.21 – структурна, прикладна і математична лінгвістика, 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 (філософія, семіотика, математика, інформатика, логіка, кібернетика,

ергономіка, акустика, біологія, психологія, соціологія, нейрофізіологія, педагогіка), структурне моделювання і формалізація рівнів, одиниць та відношень у мові й мовленні, теоретико-множинні моделі в мовознавстві.

Обсяг дисертації, її структура, зміст та оформлення повністю 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.

Враховуючі новизну одержаних наукових результатів, їх теоретичне та практичне значення, вважаю, що дисертаційна робота «Аналіз та синтез комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту» відповідає вимогам щодо докторських дисертаційних робіт, а її авторка Висоцька Вікторія Анатоліївна, заслуговує на присудження наукового ступеня доктора технічних наук за спеціальністю 10.02.21 – структурна, прикладна і математична лінгвістика.

Офіційний опонент
заступник директора з наукової роботи
Національного центру «Мала академія наук України»
доктор технічних наук, професор О. Є. Стрижак