ВІДГУК ОФІЦІЙНОГО ОПОНЕНТА

на дисертаційну роботу ***Висоцької Вікторії Анатоліївни*** на тему: ***«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

Актуальність теми дослідження.

Проблема розроблення комп'ютерної лінгвістичної системи (КЛС) опрацювання природної мови (Natural-Language Processing, NLP) 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.

На сьогодення реалізовано та впроваджено багато КЛС різного призначення, навіть для опрацювання україномовного текстового контенту. Але це зазвичай комерційні проекти закритого типу (немає ні публікацій ні доступу до адміністративної частини) та найчастіше це є іноземні проекти. Публікацій ніби багато для розуміння як в загальному відбувається процес опрацювання природної мови, особливо для англійських текстів. Але застосувати ці моделі, методи, алгоритми та технології напряму для україномовного текстового контенту не приводить майже ні до якого позитивного результату. Вже саме на рівні морфологічного аналізу виникає суттєвий конфлікт між розробленими методами та вхідним українським текстом – на виході не коректний результат. Наприклад для простого алгоритму Портера (стемінг) без відповідної модифікації не коректне буде відокремлення основи слова від флексії, що призведе до некоректної ідентифікації ключових слів текстів, щ в свою чергу впливає на будь-яку NLP-задачу, де необхідно швидко ідентифікувати множину ключових слів (рубрикація, пошук, анотування тощо). Визначення основних процесів та особливостей лінгвістичного аналізу україномовних текстів значно полегшить етапи опрацювання текстового

потоку контенту як інтеграція, супровід та управління контентом. В свою чергу адаптація процесів інтелектуального аналізу текстового контенту з ідентифікацією функціональних вимог до відповідних модулів КЛС призведе до можливості розробити типову архітектуру подібних систем на принципі модульності (додавання компонентів в залежності від змісту NLP-задачі та призначення КЛС).

Застосування вказаних технологій/методів/моделей в типовій архітектурі КЛС, адаптованих для будь-якої NLP-задачі опрацювання україномовного текстового контенту, є необхідною передумовою успішної реалізації проекту комп'ютерної лінгвістичної системи для розв'язку конкретної NLP-задачі, який вимагає застосування відповідної множини стандартних бібліотек, утиліт та програмного забезпечення з відкритим кодом, що вирішуватимуть спеціалізовані задачі проекту згідно потреб кінцевого користувача.

Дисертаційна робота Висоцької В.А. присвячена вирішенню актуальної науково-прикладної проблеми 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 для розв'язку різних NLP-задач.

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

Наукові положення, висновки та рекомендації, сформульовані в дисертації, повною мірою обґрунтовані, оскільки вони логічно випливають із результатів, отриманих за допомогою чітких математичних викладок з коректним використанням коректних методів моделювання логічних лінгвістичних структур україномовного текстового контенту та розроблення процедур його опрацювання. Теоретичний аналіз, проведений у дисертаційній роботі, ґрунтується на сучасних методах опрацювання та аналізу текстового контенту, теорії формальних граматик, теорії множин, теорії моделей даних та знань, теорії ймовірності і математичної статистики, теорії моделей, теорії алгоритмів та логіко-лінгвістичних числень, теорії інформації, об'єктно-орієнтованого програмування. Достовірність отриманих результатів зумовлена коректністю виконаних досліджень, математичних моделей та розрахунків, проведених за допомогою сучасних прикладних програмних пакетів.

Наукова новизна дисертаційної роботи полягає в наступному.

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;

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;

Одержала 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.

Значення одержаних результатів для науки і практики, отриманих у дисертаційній роботі, полягає у розробленні нової методики побудови комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту для розв'язку різних NLP-задач на основі застосування інтелектуального

аналізу текстового потоку з інформаційних ресурсів. Це стало можливим завдяки поєднанню адаптованих до української мови методів лінгвістичного аналізу, вдосконаленої інформаційної технології опрацювання інформаційних ресурсів, технології машинного навчання та множини метрик оцінювання ефективності функціонування комп'ютерних лінгвістичних систем. Особливість побудови таких КЛС полягає на основному принципу модульності системи (наявності/відсутності основних та додаткових модулів), що полегшує розроблення конкретних модулів згідно вимог щодо реалізації відповідних процесів для розв'язку конкретної NLP-задачі. Автор застосував принципово новий підхід до визначення авторства текстового контенту, який ґрунтується на аналізі множини ключових слів, множини стійких словосполучень, колекції показників лінгвометрії, стилеметрії, а також аналізі результатів N-грам. Розроблені методи та засоби дали можливість будувати комп'ютерні лінгвістичні системи опрацювання україномовного текстового контенту для розв'язку конкретної NLP-задачі згідно потреб постійної/потенційної цільової аудиторії на основі аналізу історії їх дій на Web-ресурсі КЛС.

Практичну цінність роботи підтверджує впровадження її результатів в держбюджетні теми та навчальний процес Національного університету «Львівська політехніка, а також в навчальний процес Національного технічного університету «Харківський політехнічний інститут», а саме за рахунок використання навчальних посібників: Математична лінгвістика. Книга 2. Комбінаторна лінгвістика: навчальний посібник / В. В. Пасічник, Ю. М. Щербина, В. А. Висоцька, Т. В. Шестакевич. Львів: Вид-во Львів. політехніки, 2019. 250 с.; Литвин В. В. Глибинне навчання: навч. посіб. / В. В. Литвин, Р. М. Пелещак, В. А. Висоцька. – Львів: Видавництво Львівської політехніки, 2021. – 264 с.; Чисельні методи в комп'ютерних науках / В. А. Андруник, В. А. Висоцька, В. В. Пасічник, Л. Б. Чирун, Л. В. Чирун. Львів: Новий Світ – 2000, 2017. Т. 1. 470 с.; Чисельні методи в комп'ютерних науках / В. А. Андруник, В. А. Висоцька, В. В. Пасічник, Л. Б. Чирун, Л. В. Чирун. Львів: Новий Світ – 2000, 2017. Т. 2. 536 с.; Ришковець Ю. В., Висоцька В. А. Алгоритмізація та програмування. Ч. 1: навч. посіб. Львів: Новий Світ – 200, 2018. 337 с.; Ришковець Ю. В., Висоцька В. А. Алгоритмізація та програмування. Ч. 2: навч. посіб. Львів: Новий Світ – 2000, 2018. 316 с.; Висоцька В. А., Литвин В. В., Лозинська О. В. Дискретна математика: практикум: навч. посіб. Львів: Новий Світ – 2000, 2019. 575 с.; Висоцька В. А., Оборська О. В. Python: алгоритмізація та програмування: навчальний посібник. Львів: Новий Світ – 2000, 2020. 516 с. Результати дисертаційної роботи Висоцької В.А., які опубліковано у багатьох статтях, використовуються у роботі зі студентами при підготовці наукових публікацій, дипломних магістерських робіт.

Рекомендації щодо використання результатів дисертації.

Запропонований підхід, методи, моделі та програмні засоби можуть бути використані розробниками комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту для розв'язку різних NLP-задач, що має важливе значення для технічних, соціальних комунікацій, філологічних психологічних, національної безпеки та юридичних наук.

Повнота викладу в опублікованих працях.

Основні результати дисертації опубліковано у 254 наукових публікаціях, із них 71 стаття у наукових фахових виданнях України (зокрема, 26 із них включено до Scopus або Web of Science), 72 статті у наукових періодичних виданнях інших держав (зокрема, 59 із них включено до Scopus або Web of Science, з них 4 статті опубліковано в журналах з Q2), 100 тез доповідей та матеріалів конференцій (зокрема, 64 із них включено до Scopus або Web of Science), 9 монографій та 2 розділи монографії, які включено до міжнародних наукометричних баз. Зокрема 50 статей у фахових наукових виданнях України та 31 стаття у наукових періодичних виданнях інших держав відповідають вимозі МОН України щодо публікації в одному виданні.

Оцінка змісту дисертаційної роботи, її завершеність.

Дисертаційна робота є завершеною науковою працею, яка складається з анотацій, вступу, шести розділів, висновків, списку використаних джерел з 1044 назв на 52 сторінках та 6 додатків на 82 сторінках. Загальний обсяг дисертації – 480 сторінок, з них: 306 сторінок основного тексту, 179 рисунків, 62 таблиці.

У **вступі** обґрунтовується актуальність теми, формуються мета та основні задачі досліджень, подається короткий зміст роботи.

У **першому розділі** проведено аналіз сучасного стану та перспективи розвитку ІТ опрацювання україномовного текстового контенту. Визначено поняття КЛС та наведена загальна їх класифікація. Проведений детальний аналіз відомих КЛС, що дало можливість вдосконалити загальну класифікацію відповідних систем. Визначені основні задачі NLP комп'ютерних лінгвістичних систем, на основі яких наведені приклади та порівняльний аналіз відомих сучасних КЛС. Це дало можливість сформувати загальні напрями дослідження. Проведено аналіз специфіки побудови КЛС шляхом систематизації процесів реалізації та функціонування, що забезпечить можливість виділити клас систем, функціональні властивості яких дозволяють виконувати кількісне оцінювання очікуваних ефектів впровадження типової КЛС опрацювання україномовного текстового контенту для розв'язку різних задач NLP. Описана та проаналізована основна загальна схема процесу лінгвістичного аналізу тексту природньою мовою засобами КЛС. Визначені основні стани та властивості КЛС, їх класифікація та особливості. Проаналізовано відомі класичні підходи та напрями NLP. Наведена загальна класифікація основних підходів NLP, напрямів та додаткових методів лінгвістичного дослідження для задач NLP. Також визначені основні методи дослідження когнітивної лінгвістики. Проведено аналіз існуючих основних методів та методики NLP засобами машинного навчання (ML). Проведена їх класифікація та визначені типові проблеми методів ML для опрацювання україномовних текстів. Зроблений огляд відомих ІТ розроблення КЛС на основі особливостей інтелектуального аналізу потоку україномовного контенту. Визначені основні вимоги до оцінювання ефективності КЛС на основі технології ML та аналізу великих даних зі сховищ даних україномовного текстового контенту. Розглянуті основні методи ML для аналізу великих даних з множини текстових потоків контенту. Визначені вимоги до кластеризації текстового контенту при неконтрольованому ML. Подано загальний огляд проблеми побудови КЛС

опрацювання україномовного текстового контенту. Визначені обмеження існуючих методів опрацювання україномовного текстового контенту та розроблення комп'ютерних лінгвістичних систем для розв'язку різних NLP-задач дали змогу сформулювати науково-прикладну проблему.

У **другому розділі** розроблена ІТ опрацювання україномовного текстового контенту на відміну від існуючих підтримує принцип модульності типової архітектури КЛС для розв'язку конкретної задачі NLP та аналізу множини параметрів та метрик ефективності функціонування системи відповідно до поведінки цільової аудиторії. Розроблено загальну структуру КЛС для опрацювання текстового контенту українською мовою та концептуальну схему/модель функціонування типової КЛС на основі моделювання взаємодії основних процесів і компонентів системи, що дало змогу вдосконалити ІТ інтелектуального аналізу текстового потоку на основі опрацювання інформаційних ресурсів. 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 NLP-задачі. Зроблена та конкретизована постановка проблеми опрацювання україномовного тексту на основі визначення функціональних особливостей інтелектуального аналізу текстового потоку. Загальний аналіз проблеми аналізу україномовного тексту та визначення основних проблем опрацювання україномовного тексту дало можливість сформулювати основні етапи та вимоги до проекту типової КЛС розв'язку конкретної NLP-задачі. Ідентифікація основних характеристик КЛС та обґрунтування реалізації проекту типової КЛС дало можливість визначити очікувані ефекти від відповідної реалізації проекту. На основі аналізу вхідних/вихідних потоків контенту комп'ютерної лінгвістичної системи визначені та сформульовані функціональні вимоги до проекту типової КЛС, її програмних модулів, мережних, програмних та технічних інструментів програмної реалізації ІС.

У **третьому розділі** розроблена загальна структура КЛС опрацювання текстового контенту українською мовою та концептуальна схема/модель функціонування типової КЛС на основі моделювання взаємодії основних процесів та компонентів ІС. Здійснено моделювання основних NLP-процесів КЛС за рахунок взаємодії основних процесів/компонентів ІС та адаптованих до української мови методів лінгвістичного опрацювання текстового контенту на основі графемного, морфологічного, лексичного, синтаксичного, семантичного, структурного, онтологічного та прагматичного аналізу дозволила вдосконалити ІТ інтелектуального аналізу текстового потоку для розв'язку конкретної задачі NLP. Це забезпечило адаптацію процесів NLP для аналізу україномовного текстового контенту. Розроблена та описана формальна модель комп'ютерної лінгвістичної системи для опрацювання україномовного текстового контенту, що дало змогу визначити основні структурні елементи та оператори опрацювання природної мови на кожному рівні аналізу тексту як графемного/фонологічного, морфологічного, синтаксичного, семантичного, референційного, структурного, онтологічного та прагматичного. У зв'язку зі складністю морфології української мови детальна увага приділена саме опису моделі морфологічного аналізу текстового контенту.

У **четвертому розділі** розроблена загальна архітектура комп'ютерних лінгвістичних систем на основі основних процесів опрацювання інформаційних ресурсів як інтеграція, супровід та управління контентом, а також з застосуванням методів інтелектуального та лінгвістичного аналізу текстового потоку з використанням технології машинного навчання. Удосконалено ІТ 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 NLP та підвищити ефективність функціонування КЛС на 6-9%. Це стало можливим завдяки поєднанню адаптованих до української мови методів лінгвістичного аналізу, вдосконаленої ІТ опрацювання інформаційних ресурсів, ML та множини метрик оцінювання ефективності функціонування КЛС. Основний принцип побудови таких КЛС полягає на модульності, що полегшує їх побудову згідно вимог щодо наявності відповідних процесів для розв'язку конкретної задачі NLP. Описано основні NLP-методи на основі регулярних виразів узгодження з шаблонами при графемному та морфологічному аналізах україномовних текстів.

Удосконалено методи ОПМ на основі регулярних виразів узгодження з шаблонами, що дало змогу адаптувати методи токенізації та нормалізації тексту каскадами простих підстановок регулярних виразів та кінцевих автоматів. Визначені основні допустимі операції регулярних виразів як об'єднання та диз'юнкція символів/ланцюжків/виразів, оператори лічби та прецедентності, а також анкори як спец-символи ідентифікації присутності/відсутності символів в RE. Визначені основні етапи токенізації та нормалізації українського тексту каскадами простих підстановок регулярних виразів та кінцевих автоматів.

Удосконалено метод МА 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%.

Реалізовані та описані алгоритми сегментації та нормування слова, сегментації речення та модифікований стеммінг Портера як ефективний спосіб ідентифікації афіксів лем для можливості розмічування аналізованого слова. На відмінну від класичного алгоритму Портера (не має високої точності навіть для англомовних текстів) модифікований є адаптованим саме для української мови та дає точний результат в межах 85-93% випадків в залежності від якості, стилю, жанру тексту та відповідно наповнення словників КЛС. Описано алгоритм мінімальної редакційної відстані рядків українських текстів як мінімальна кількість операцій, необхідних для перетворення одного в інший.

У **п'ятому розділі** розроблено метод синтаксичного аналізу україномовного текстового контенту, спрямованого на автоматичне виявлення значущих ключових слів вхідних текстів. Визначено роль і формальні ознаки синтаксичного аналізатора в процесі виявлення ключових слів тематики контенту, проведено декомпозицію процедур запропонованого методу на 4-х етапах. Порівняно з відомими синтаксичними аналізаторами, запропонований метод забезпечує самовдосконалення та самонавчання автоматизованої системи визначення ключових

слів за рахунок механізму ідентифікації значущих статистичних параметрів у визначених модератором межах. Експериментальне дослідження підтвердило достовірність методу – для різних методик опрацювання первинного тексту середній збіг списків виявлених ключовиків 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% згідно із етапами аналізу текстів статей. Точність збігу ключових слів порівняно зі всіма знайденими системою коливається в проміжку 34,3-71,9% відповідно до етапів аналізу текстів статей.

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%. Метод полягає у використанні закону Зіпфа при формуванні стійких словосполучень як ключових з врахуванням наступних правил попереднього лінгвістичного опрацювання тексту: вилучення всіх стових слів; біграми формувати лише в межах знаків пунктуації; дієслово та займенник вважати знаками пунктуації; дієслова визначати за їх флексіями; біграми формувати на основі їх основ без врахування їх флексій; визначення прикметників за їх флексіями та вважати, що прикметники мають бути лише на першому місці у біграмі з україномовних текстів. Розроблено програми комплекс для визначення стійких словосполучень як ключових. Запропоновано підхід до розроблення ПЗ лінгвістичного контент-аналізу для визначення стійких словосполучень при ідентифікації ключових слів текстового україномовного та англомовного контенту. Особливість підходу полягає у адаптації лінгвостатистичного аналізу лексичних одиниць до особливостей конструкцій україномовних та англомовних слів/текстів. Досліджено результати експериментальної апробації запропонованого методу контент-аналізу англомовних та україномовних текстів для визначення стійких словосполучень при ідентифікації ключових слів технічних текстів.

У **шостому розділі** розроблено метод визначення автора в україномовних текстах на основі аналізу коефіцієнтів лексичного авторського мовлення в еталонному уривку авторського тексту. Метод ґрунтується на аналізі колекції ключових слів, стійких словосполучень, показників лінгвометрії, стилометрії, а також результатів аналізу N-грам. Впровадження методу забезпечило можливість визначити множину потенційних авторів публікацій та розробити метод ідентифікації авторського стилю. Метод полягає в порівняльному аналізі авторської атрибуції в статистично опрацьованому доробку автора (еталоні) з довільним аналізованим уривком. Метод оцінює ступінь приналежності тексту до шаблону авторського стилю із аналізом відповідних коефіцієнтів лексичного авторського мовлення. Причому метод працює при умові, що шаблон авторського стилю згенерований на достовірних даних. Для атрибуції використано аналіз опорних слів. Отримані результати подано у вигляді коефіцієнтів кореляції.

У розділі також розглянуто алгоритм ідентифікації службових слів на основі лінгвістичного аналізу текстового контенту. Для кожного з уривків проаналізовані та порівняні із еталонним значеннями абсолютні та відносні частоти появи стоп-

слів. Таким чином, застосування методу опорних слів дозволяє знаходити серед досліджуваних уривків фрагмент, що найбільш ймовірно належить до еталону. Інші результати підтверджують дієвість методу опорних слів у авторській атрибуції текстів. Висунуте припущення про незначущість впливу частки як параметра методу на результати привело до зменшення коефіцієнтів кореляції. Проте, для підтвердження чи спростування того факту, що частки не є визначальним фактором в авторському стилі необхідно виконати додаткові ґрунтові дослідження.

У розділі наведено алгоритм лексичного аналізу україномовних текстів та алгоритм синтаксичного аналізатора текстового контенту. Особливостями алгоритмів є адаптація морфологічного та синтаксичного аналізу словоформ до своєрідності побудови україномовних слів/текстів. При цьому авторка враховувала частини мови та відмінювання в межах цієї частини мови на основі аналізу флексій та основ слів за регулярними виразами.

Наведено порівняння результатів контент-моніторингу на множині 300 одноосібних робіт технічного спрямування біля 100 різних авторів за період 2001–2021 рр. для визначення чи змінюються і як змінюються коефіцієнти різноманітності тексту авторів в різні проміжки часу. Також проведено декомпозицію методу ідентифікації потенційного автора на основі аналізу таких параметрів стилю мовлення як зв'язність мовлення, ступень синтаксичної складності, лексична різноманітність, ступінь концентрації та винятковості. Проаналізовані також таки ознаки авторського стилю, як загальний обсяг слів тексту, обсяг унікальних слів, обсяг сполучників/прийменників, обсяг речень, обсяг слів із частотою 1 та ≥10.

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. Розроблено КЛС на інформаційному ресурсі http://victana.lviv.ua засобами CMS Joomla! (для розроблення е-каркасу статей), PHP (для реалізації методів опрацювання текстового контенту), HTML (для реалізації розмітки сторінок), CSS (для опису стилів сторінок), MySQL (для зберігання даних та словників). Експериментальне дослідження підтвердило достовірність методу визначення ключових слів.

У **додатках** наведено основні регулярні вирази МА українських іменників та дієслів, дерево закінчень слів в українській мові, критерії ГА 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 наведені акти впровадження результатів дисертаційної роботи.

Відповідність дисертації паспорту спеціальності.

Дисертація Висоцької В.А. на тему «Аналіз та синтез комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту» є завершеною науковою працею та відповідає паспорту спеціальності 10.02.21 – структурна, прикладна і математична лінгвістика, а саме за такими розділами: «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», «Обчислювальна, статистична та квантитативна лінгвістика», «Автоматизовані та автоматичні системи маркування лінгвістичних структур текстів (графемний, морфологічний, лексичний, семантичний, синтаксичний, концептологічний аналіз тощо)», «Лінгвістичне забезпечення інформаційних систем».

Зауваження до дисертаційної роботи такі:

1. В пункті 4.1 «Загальна архітектура комп'ютерних лінгвістичних систем» мало уваги приділено опису моделей машинного навчання конвеєра опрацювання україномовного текстового контенту, формуванню навчальних датасетів та використаних методів машинного навчання, в тому числі глибинного, наприклад, на Transformer або Recurrent Neural Networks.

2. Метод ідентифікації ключових слів україномовного контенту пункту 5.1 мав би бути поданий з урахуванням описаного у пунктах 4.2-4.5 регулярних виразів та продукційних правила лінгвістичного аналізу тексту та в пункті 4.1 конвеєра опрацювання текстового контенту на основі машинного навчання.

3. Пункт 6.3 «Лінгвометричний аналіз визначення автора контенту на основі статистичних параметрів різноманітності мовлення» в основному побудований на експериментах. Варто було приділити більшу увагу його теоретичному обґрунтуванню.

4. Пункт 6.4 «Метод кількісної оцінки визначення авторства текстового контенту на основі статистичного аналізу розподілу N-грам» мав би бути краще структурований, оскільки включає аналіз табличних даних, графіків, зображень, код програм, скринів, опис алгоритмів тощо.

5. В пункті 6.5 «Аналіз розробленого методу кількісної оцінки ідентифікації потенційного автора науково-технічної публікації» не описано як впливає на результат час написання шаблону (уривку еталона) авторської роботи при визначенні ступеня вірогідності автора науково-технічної публікації з віддаленого хоча б понад десятиліття від часу написання оригіналу.

6.Текст дисертаційної роботи містить незначні стилістичні, граматичні, пунктуаційні та поліграфічні огріхи, та іноді залучає не чіткі малюнки.

7. Подання матеріалу дисертаційної роботи є незвичним, оскільки практична імплементація кожного методу наводиться в тому ж розділі, що і теорія, а не в окремому розділі експериментів та аналізу результатів.

Зазначені недоліки не знижують значущість та якість одержаних результатів.

Незважаючи на зауваження, дисертаційна робота справляє гарне враження за рівнем аналізу об'єкта досліджень та різноманітністю використаного математичного апарату.

Висновок.

В загальному дисертаційна робота **Висоцької Вікторії Анатоліївни** «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'ютерної лінгвістичної системи для розв'язання різноманітних задач опрацювання україномовного текстового контенту на основі розроблення нових та удосконаленні відомих моделей, методів та засобів NLP.

За актуальністю обраної теми, обсягом та рівнем виконаних досліджень, повнотою охоплення наукової проблеми, новизною і ступенем обґрунтованості отриманих результатів, практичних висновків та рекомендацій робота задовольняє вимогам, що пред'являються до докторських дисертацій. Таким чином, дисертаційна робота «Аналіз та синтез комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту» відповідає вимогам МОН України, які висуваються до робіт на здобуття наукового ступеня доктора наук, зокрема 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 їй наукового ступеня доктора технічних наук за спеціальністю 10.02.21 – структурна, прикладна і математична лінгвістика.

Офіційний опонент,
професор кафедри інтелектуальних
комп'ютерних систем
Національного технічного університету
«Харківський політехнічний інститут»
Міністерства освіти і науки України,
доктор технічних наук, професор _____ Ніна ХАЙРОВА

Підпис професора Ніни Хайрової засвідчую

Підпис проф. Ніни Хайрової
ЗАСВІДЧУЮ:
ВЧЕНИЙ СЕКРЕТАР
НАЦІОНАЛЬНОГО ТЕХНІЧНОГО УНІВЕРСИТЕТУ
"ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ"
"28" 08 ЗАЙЦЕВ Ю. І.

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ * Національний технічний університет "Харківський політехнічний інститут" * код 02071180