Міністерство освіти і науки України

Національний університет «Львівська політехніка»

ВИСОЦЬКА Вікторія Анатоліївна

УДК 004.82:004.89:004.91

АНАЛІЗ ТА СИНТЕЗ КОМП'ЮТЕРНИХ ЛІНГВІСТИЧНИХ СИСТЕМ ОПРАЦЮВАННЯ УКРАЇНОМОВНОГО ТЕКСТОВОГО КОНТЕНТУ

10.02.21 – структурна, прикладна і математична лінгвістика

РЕФЕРАТ

дисертації на здобуття наукового ступеня
доктора технічних наук

Львів – 2023

Дисертацією є рукопис

Робота виконана у Національному університеті «Львівська політехніка», Міністерства освіти і науки України

Науковий консультант: доктор технічних наук, професор, **Литвин Василь Володимирович**, Національний університет «Львівська політехніка» Міністерства освіти і науки України, завідувач кафедри інформаційних систем та мереж

Офіційні опоненти: доктор технічних наук, професор, **Хайрова Ніна Феліксівна**, Національний технічний університет «Харківський політехнічний інститут» Міністерства освіти і науки України, професор кафедри інтелектуальних комп'ютерних систем;

доктор технічних наук, старший науковий співробітник **Стрижак Олександр Євгенійович**, Національний центр «Мала академія наук України», заступник директора з наукової роботи;

доктор технічних наук, професор, **Шинкаренко Віктор Іванович**, Український державний університет науки і технологій МОН України, професор кафедри комп'ютерних інформаційних технологій факультету комп'ютерних технологій і систем.

Захист відбудеться «14» вересня 2023р. о 13.00 год. на засіданні докторської ради Д 35.052.05 у Національному університеті «Львівська політехніка» за адресою: 79013, м. Львів, вул. С. Бандери, 12, аудиторія 226 головного корпусу.

З дисертацією можна ознайомитися в науково-технічній бібліотеці Національного університету «Львівська політехніка» за адресою: 79013, м. Львів, вул. Професорська, 1.

Реферат розісланий «14» серпня 2023 р.

Вчений секретар
спеціалізованої вченої ради Р.А. Бунь
Д 35.052.05

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. На сьогодення активний розвиток інформаційних технологій (ІТ) знаходиться на перетині глобалізації та інформатизації. Швидкі темпи зростання інформатизації суспільства напряму пов'язані з темпами розвитку та впровадженням комп'ютерних лінгвістичних систем (КЛС), розроблення яких базується на моделях та методах опрацювання природної мови (ОПМ). Складність розроблення моделей, методів та засобів ОПМ полягає не лише в розв'язку не типових задач ОПМ, але й в адаптації цих моделей, методів та засобів для конкретної природної мови. Кожна природня мова є унікальною, зі своїм колоритом правил, історії, граматики, виключень та особливостей генерування лінгвістичних одиниць для передачі сенсу, що ускладнює процес розроблення КЛС.

Зазвичай кожний успішний проект розроблення КЛС призначений під конкретну задачу (наприклад, машинний переклад, ідентифікація плагіату/рерайту, рубрикація тексту, аналіз атрибуції тексту, інформаційний пошук, реферування, голосові помічники, інтелектуальні чат-боти тощо) 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. Рідкісні випадки, коли до таких проектів КЛС розробники надають відкритий доступ та можливість ознайомитися з їх структурою та змістом. Створення будь-якого прикладного додатку ОПМ для довільної природної мови із понад 7000 мов та діалектів базується на дослідженні великих текстових одномовних/паралельних корпусів цієї мови, який містить понад сотень мільйонів слів та лінгвістичних ресурсів. Лише близько для 20 природних мов (англійська, китайська, західноєвропейські мови, японська тощо) відомі результати досліджень таких корпусів, що дає змогу для цих мов розробляти КЛС різної складності. 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 прикладних додатків ОПМ. Такі прикладні додатки використовуються для побудови КЛС в кібербезпеці (виявлення фейків та пропаганди, так званих тролів/ботів в соціальних мережах), соціології (аналіз динаміки зміни громадської думки на тематичні питання), філології (автоматичне дослідження великих масивів даних різного тематичного спрямування та різних часових періодів), психології (аналіз психологічного портрету особи, ідентифікація посттравматичного стресового розладу учасників бойових дій або окупації), національній безпеці (інформаційна війна), юриспруденції (криміналістика та судова справа), соціальних комунікаціях (аналіз дописів спільнот в соціальних мережах) та в інших важливих галузях сучасної України. Означене обумовлює актуальність теми дисертаційного дослідження.

Наукові дослідження N. Chomsky, В.М. Глушкова, А.В. Гладкого, Д.В. Ланде, В.А. Широкова, Н.В. Шаронової, Н.Ф. Хайрової, О.П. Левченко, О.В. Бісікала, С.Н. Бук, Н.П. Дарчук, З.В. Партика, А.В. Анісімова, Ю.Д. Апресяна, О.О. Марченка, І.М. Кульчицького, А.О. Никоненка, М. Гросса, А. Лантена, V.H. Yngve, S. Sharoff,

Ю.А. Шрейдера, D. Jurafsky, B. Bengfort, J.H. Martin, L. Tesniere, T. Ojeda, P.M. Postal, D.G. Hays, T.A. van Dijk, S. Marcus, J. Lyons, L.W. Tosh, Y. Bar-Hillel, D.G. Bobrow, G. Lakoff, R. Bilbro, N. Kotsyba, А.Ю. Берка, Ю.М. Щербини, В.Ю. Величка, В.Ф. Старка та багатьох інших дають змогу зрозуміти основні принципи лінгвістичного опрацювання тексту в залежності від особливостей конкретної природньої мови. Більше 80% таких досліджень стосуються опрацювання англомовних текстів. Суттєво менше досліджень для слов'янських мов, зокрема, для низькоресурсної української мови. Зокрема відсутні публікації щодо рекомендацій розроблення, функціональних вимог, загальної структури та типової архітектури КЛС опрацювання україномовного текстового контенту. Напряму застосувати моделі, методи, алгоритми та ІТ опрацювання англійської мови для україномовного текстового контенту не приводить до позитивних результатів. Вже на рівні морфологічного аналізу виникає суттєвий конфлікт між розробленими методами для англомовного тексту та їх використанням для україномовного тексту. Наприклад, для простого алгоритму Портера (стемінг) без відповідної модифікації не коректним є відокремлення основи слова від флексії, що призводить до неточності ідентифікації ключових слів, що, в свою чергу, впливає на розв'язок будь-якої задачі ОПМ, де необхідно швидко ідентифікувати множину ключових слів (рубрикація, пошук, анотування тощо). Визначення основних особливостей та процесів лінгвістичного аналізу українськомовних текстів значно полегшить етапи опрацювання текстового потоку інформації 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 її типову архітектуру на принципі модульності (додавання компонентів в залежності від змісту задачі ОПМ та призначення КЛС).

Наведене свідчить про актуальність досліджень під час вирішення важливої науково-прикладної проблеми 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 ОПМ.

Зв'язок роботи з науковими програмами, планами, темами. Тема дисертації відповідає науковому напряму «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.).

Мета і завдання дослідження. Метою роботи є розроблення моделей, методів, засобів аналізу та синтезу комп'ютерних лінгвістичних систем на базі нових та удосконалення відомих методів опрацювання україномовного текстового контенту для розв'язання задач опрацювання природньої мови. Метою дисертаційної роботи визначено необхідність виконання таких завдань:

1) провести аналіз специфіки побудови КЛС шляхом систематизації процесів їх 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 ОПМ;
2) розробити інформаційну технологію побудови КЛС опрацювання україномовного тексту, що дасть змогу визначити їх базову структуру, функціональні вимоги, послідовність налаштування та навчання системи, загальні засади проектування;
3) запропонувати ІТ опрацювання інформаційних ресурсів як інтеграція, управління та супровід українськомовного контенту на основі вдосконалення лінгвістичного аналізу текстового контенту для розроблення метрик оцінювання ефективності функціонування КЛС для розв'язку різних задач ОПМ;
4) розробити методи опрацювання україномовного текстового контенту для розв'язку різних задач ОПМ для підвищення точності отриманих результатів;
5) розробити методи та засоби інтелектуального аналізу текстового контенту для підвищення ефективності розв'язку різних задач ОПМ;
6) створити програмні модулі опрацювання україномовного текстового контенту для розв'язку різних задач ОПМ та проведення експериментів;
7) провести апробацію отриманих результатів шляхом побудови та впровадження прикладних КЛС опрацювання україномовного текстового контенту.

Об'єктом дослідження є процеси аналізу та синтезу комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту.

Предметом дослідження є моделі, методи та засоби опрацювання україномовного текстового контенту для розв'язку різних задач ОПМ.

Методи дослідження. 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 різних задач ОПМ; методи об'єктно-орієнтованого та системного аналізу і проектування – для проектування та розроблення КЛС; теорію реляційних баз даних, методи штучного інтелекту, об'єктно-орієнтоване програмування – для програмної реалізації КЛС опрацювання україномовного текстового контенту для розв'язку різних задач ОПМ.

Наукова новизна одержаних результатів полягає у вирішенні важливої науково-прикладної проблеми аналізу та синтезу КЛС для розв'язання різних задач опрацювання україномовного текстового контенту на основі розроблення нових та удосконалення відомих моделей, методів та засобів ОПМ. 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 ОПМ шляхом реалізації типового програмного забезпечення таких систем;

одержали 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 ОПМ;
- 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 ОПМ;

удосконалено

- методи ОПМ, 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 ОПМ.

Практичне значення одержаних результатів полягає у тому, що їх можна використати для побудови прикладних КЛС опрацювання україномовного текстового контенту. Зокрема, 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 на 6-9% та виділити з тексту тематичні терми для подальшої рубрикації публікації;
– 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 на 6-9% в залежності від розв'язку конкретної задачі ОПМ;
– застосування методу обчислення ступеня верифікації автора україномовного тексту на основі аналізу стилів потенційних авторів дозволило підвищити точність ідентифікації на 6-12% 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 до [9;34]% із загальної кількості учасників проекту в залежності від тематики та часового діапазону написання науково-технічної публікацій, а також частоти публікацій цього автора в цей проміжок на конкретну тематику;
– 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.

Особистий внесок здобувача. Усі наукові результати, подані у дисертації, одержані здобувачем особисто. Роботи [5-6, 54, 73, 75-76, 80] опубліковано без співавторів. У друкованих працях, опублікованих у співавторстві, особисто здобувачу належать такі результати: [41, 43, 46, 48] – метод класифікації текстових документів; [2, 7, 23-26, 33, 40, 89] – вдосконалений метод управління контентом інформаційних ресурсів; [21, 29, 62, 70] – метод інтеграції контенту інформаційних ресурсів; [4, 57] – аналіз часової залежності морфології вихідного сигналу на основі машинного навчання та нейронних мереж; [34-37, 49, 96] – метод супроводу контенту інформаційного ресурсу; [38, 52, 55-56, 84] – вдосконалені лінгвістичні методи опрацювання тексту; [15, 20, 22, 27, 85-88] – метод визначення автора текстового україномовного контенту; [47, 60-61, 68-69] – аналіз ігрових методів опрацювання інформації; [1, 53, 65] – вдосконалений метод семантичного аналізу текстового україномовного контенту; [3, 16, 63, 67, 82, 97-100] – аналіз процесів опрацювання контенту в різних предметних областях на основі машинного

навчання та аналізу великих даних; [1, 65, 81, 95] – вдосконалений метод морфологічного аналізу текстового україномовного контенту; [32, 39, 66, 93] – метод інтелектуального пошуку текстового контенту; [8-9, 11, 13, 18, 83] – метод визначення ключових слів текстового україномовного контенту; [50-51, 90] – метод опрацювання службового контенту; [71-72, 74, 77-79] – метод опрацювання інформаційних ресурсів; [14, 28, 30, 31, 42, 92, 94] – онтологічний підхід для опрацювання текстового контенту; [19, 65] – вдосконалений метод синтаксичного аналізу текстового україномовного контенту; [10, 17, 64, 91] – метод контент-аналізу для опрацювання текстових масивів даних; [12, 44-45, 58-59] – метод аналізу психологічного стану особистості на основі ОПМ.

Апробація результатів дисертації. Основні результати дисертаційної роботи доповідалися на міжнародних, українських та міжвузівських конференціях та семінарах, зокрема: Міжнародна конференція «Computational Linguistics and Intelligent Systems» (CoLInS, Lviv-Kharkiv, 2017-2021); Міжнародна конференція «Modern Machine Learning Technology» (MoMLeT, Shatsk, 2019-2021); IEEE Міжнародна конференція «Smart Information Systems and Technologies» (SIST, Nur-Sultan, 2021); Міжнародна конференція «Intelligent data acquisition and advanced computing systems: technology and applications» (IDAACS, Бухарест, 2017; Metz, 2019; Cracow, 2021); IEEE Міжнародна конференція «Advanced information and communication technologies» (AICT, Lviv, 2019, 2021); Міжнародна конференція «Академічна доброчесність: виклики сучасності» (Warszawa, 2020); IEEE International Conference: Modern problems of radio engineering, telecommunications and computer science TCSET (Lviv-Slavske, 2016, 2022); Міжнародна конференція «Computer Science and Information Technologies» (CSIT, Lviv, 2015-2021); Міжнародна наукова конференція “Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту ISDMIT (Залізний Порт, 2015-2019); International scientific and practical conference “Scientific Research Priorities: theoretical and practical value” (Nowy Sanz, 2017); Міжнародний симпозіум «Intelligent data acquisition and advanced computing systems» (IDAACS-SWS, Львів, 2018); Науково-технічна конференція «Обчислювальні методи і системи перетворення інформації» (Львів, 2018); International conference of System analysis and information technology SAIT (Kyiv, 2017); IEEE Міжнародна конференція «Data Stream Mining and Processing» (DSMP, Lviv, 2016, 2018, 2020); Міжнародна конференція «Математика. Інформаційні технології. Освіта» (Луцьк, 2016); International conference of perspective technologies and methods in MEMS Design MEMSTECH (Lviv-Polyana, 2016); Всеукраїнська науково-практична конференція «Інтелектуальні системи та прикладна лінгвістика» (Харків, 2016); IEEE International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics CADSM (Lviv-Polyana, 2015); Міжнародна конференція «Обробка сигналів і негаусівських процесів» (Черкаси, 2015); Міжнародна конференція «Інформація, комунікація, суспільство 2015» (Славське, 2015). Результати дисертаційних досліджень регулярно доповідалися на наукових семінарах кафедри «Інформаційні системи та мережі» Національного університету «Львівська політехніка» (2015 р. – 2022 р.).

Публікації. Основні результати дисертаційного дослідження опубліковано у 254 наукових публікаціях, серед яких 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 квартилем 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.

Структура та обсяг роботи. Дисертаційна робота 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.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі обґрунтовано актуальність теми, сформульовано мету та основні завдання досліджень, показано зв'язок із науковими програмами, планами, темами, описано наукову новизну. Розглянуто практичну цінність, реалізацію та впровадження результатів роботи. Наведено дані про особистий внесок здобувача, апробацію роботи та публікації, подано 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 ОПМ 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 ОПМ. 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 ОПМ. Наведена загальна класифікація основних підходів ОПМ, напрямів та додаткових методів лінгвістичного дослідження для задач ОПМ. Також визначені основні методи дослідження когнітивної лінгвістики. Проведено аналіз існуючих основних методів та методики ОПМ засобами машинного навчання (МН). Проведена їх класифікація та визначені типові проблеми методів МН 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 МН та аналізу великих даних зі сховищ даних (СД) україномовного текстового контенту. Розглянуті основні методи МН для аналізу великих даних з множини текстових потоків контенту. Визначені вимоги до кластеризації текстового контенту при неконтрольованому МН. Подано загальний огляд проблеми побудови КЛС

опрацювання україномовного текстового контенту. Визначення основних процесів та особливостей лінгвістичного аналізу українськомовних текстів значно полегшить етапи опрацювання текстового потоку контенту як інтеграція, супровід та управління контентом (рис. 1). Адаптація процесів інтелектуального аналізу текстового контенту з ідентифікацією функціональних вимог до відповідних модулів КЛС призведе до можливості розробити типову структуру подібних систем на принципі модульності (додавання компонентів в залежності від змісту задачі ОПМ та призначення КЛС).

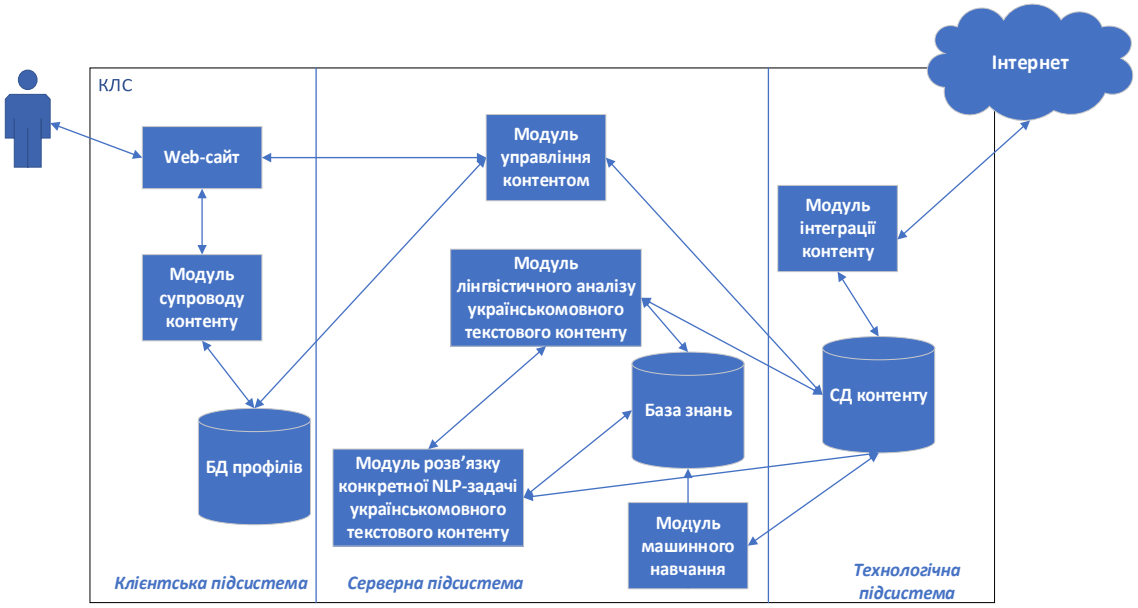


Рис. 1. Узагальнена структура комп'ютерної лінгвістичної системи

Застосування вказаних ІТ/методів/моделей в типовій структурі КЛС, адаптованих для будь-якого процесу опрацювання україномовного текстового контенту, є необхідною передумовою успішної реалізації проекту КЛС для розв'язку конкретної задачі ОПМ, який вимагає застосування відповідної множини стандартних бібліотек, утиліт та програмного забезпечення (ПЗ) з відкритим кодом, що вирішуватимуть спеціалізовані задачі проекту згідно потреб кінцевого користувача. Стан КЛС визначається кортежом головних властивостей в конкретний момент часу або активності відповідного процесу ОПМ: $s_i = (p_{i1}, p_{i2}, \dots, p_{im})$, $i = \overline{1,n}$, де s_i – відповідний i-ий стан в конкретний момент часу t_l з множини з потужністю $|S|=n$, p_{ij} – відповідна ij-та властивість стану з множини з потужністю $|P|=m$, яка визначає поведінку КЛС як $p_j = (r_{ij1}, r_{ij2}, \dots, r_{ijv})$, $j = \overline{1,m}$, де r_{ijk} – відповідний параметр конкретної властивості p_{ij} для стану s_i.

Для будь-якої КЛС станом s_i є один із процесів ОПМ, наприклад, ідентифікація ключових слів і/або стійких словосполучень для наступного стану s_{i+1} системи як рубрикація текстового масиву даних. Відповідно властивостями для стану s_i є морфологічна p_{i1}, лексична p_{i2} та синтаксична p_{i3}, в окремих ОПМ задачах може бути і семантична тощо. Тоді для властивості p_j визначається множина параметрів для відповідного аналізу тексту в залежності від конкретної задачі ОПМ. За цими параметрами уточнюють стратегію функціонування КЛС в момент часу t_l для:

– морфологічної властивості p_{i1} параметрами є N-грами та морфеми: корені r_{i11}, закінчення r_{i12}, афікси r_{i13}; граматичні категорії різних частин мови r_{i14}, довжина слова r_{i15}, місцерозташування слова в реченні r_{i16}, кількість складів у слові r_{i17}, кількість змісту слова r_{i18}, співвідношення приголосних та голосних r_{i19}, тощо;
– лексичної властивості p_{i2} параметрами є місцезнаходження речення в тесті r_{i21}, місцезнаходження слова в реченні r_{i22}, вага слова r_{i23}, вага речення r_{i24}, основа слова r_{i25}, флексія слова r_{i26} тощо;
– синтаксичної властивості p_{i3} параметрами є глибина слова в дереві залежностей речення r_{i31}, місцерозташування слова в реченні r_{i32}, кількість змісту слова r_{i33}, кількість слів на речення r_{i34}, кількість слів r_{i35} та речень r_{i36}, чи є слово з великої літери r_{i37}/ з дефісом r_{i38}/ складеним r_{i39} тощо;
– семантичної властивості p_{i4} параметрами є кількість змісту слова r_{i41}, глибина слова в дереві залежностей r_{i42}, розмір абзаців r_{i43}, розміщення абзаців r_{i45} тощо;

В залежності від кортежу $p_j \in s_i$ визначається поведінка КЛС, тобто реалізація множини правил (активація дій або подій) реалізації конкретного процесу ОПМ в залежності від вхідних текстових даних. Відповідно подією o_l є зміна однієї властивості на іншу $p_{ij} \rightarrow p_{ik}$ або $o_l: p_i \rightarrow p_j$ згідно виконання певних умов U для вхідного аналізованого тексту X та проміжного опрацьованого тексту $C: p_i = o_l(p_j, U, X, C)$. Дія d_g є процесом активації події o_l іншою подією o_v в КЛС: $C' = d_g(o_l \circ o_v)$. Чим складніша мова (морфологія, синтаксис тощо), тим складніше реалізувати опрацювання відповідних текстів природною мовою. Крім того, для таких низько-ресурсних мов як українська не стандартизовані правила та словники опрацювання текстів природною мовою для розв'язку відповідних задач ОПМ. Багато наукових лінгвістичних шкіл та ІТ-фахівців працюють над створенням україномовних словників, корпусів текстів та правил для опрацювання українських текстів. Але зазвичай це лінгвісти та філологи, які не ознайомлені з особливостями конкретних сучасних інструментів, як мови програмування, методи МН, аналізу великих даних тощо. Існує колосальна прогалина між результатами дослідження філологів та прикладних лінгвістів з одного боку, та ІТ-фахівцями з іншого для опрацювання україномовних тестів. Сьогодні досить мало реалізовано та впроваджено для загального доступу інструментів ОПМ як української.

На основі проведеного аналізу виокремлені раніше невирішені проблеми сформульовані задачі дослідження.

У другому розділі визначено основні особливості та методи аналізу та синтезу КЛС на основі визначення основних етапів як графемний, морфологічний, лексичний, синтаксичний семантичний аналіз/синтез україномовного тексту для розв'язку конкретної задачі ОПМ. На основі структурних, функціональних та інформаційних методів аналізу систем та аналізу вхідних/вихідних потоків контенту визначено основні характеристики КЛС, що дали змогу сформулювати функціональні вимоги до КЛС, її програмних модулів, мережних, програмних та технічних інструментів програмної реалізації системи. На основі формалізованих методів синтезу системи розроблено метрики очікуваних ефектів від реалізації та впровадження КЛС.

Розроблена типова структура КЛС S_{wtm} складається з модулів розв'язку конкретної задачі ОПМ M_{dis}, супроводу контенту M_{dmr}, інтеграції контенту M_{dcp},

управління контентом M_{dvm}, лінгвістичного M_{lat} та інтелектуального аналізу текстових потоків контенту (ІАТПК) M_{was}:

$$S_{wtm} =< M_{dis}, M_{dmr}, M_{dcp}, M_{dvm}, M_{lat}, M_{was} >. \quad (1)$$

Відповідно, модуль розв'язку конкретної задачі ОПМ M_{dis}:

$$M_{dis} =< N_{wvr}, S_{gcc}, S_{gco}, S_{gcv}, S_{gro}, P_{wnv}, I_{wnv} >, \quad (2)$$

де S_{gcc} – середній коефіцієнт конверсії, S_{gco} – середня вартість замовлень, S_{gcv} – середня вартість або корисність мети відвідування, S_{gro} – середня P_{ROI} або середнє повернення на інвестиції, P_{wiv} – відсоток (%) прибутку від нових відвідувачів, I_{wnv} – індекс нових покупців/замовників при першому відвідуванні.

Наявність модуля супроводу текстового контенту M_{dmr} скорочує затрати на модераторів/аналітиків, які здійснюють збір/аналіз статистичних даних динаміки функціонування КЛС, активності постійної цільової аудиторії як реакції на зміни контенту веб-сайту, формування правил аналізу інформаційних портретів користувачів та тематичних сюжетів контенту:

$$M_{dmr} =< I_{gyk}, K_{gvb}, P_{wap}, P_{wvk}, S_{grk}, I_{gck}, P_{wck}, P_{wvk}, K_{wcz}, P_{wvz} >, \quad (3)$$

де I_{gyk} – індекс якості реклами; K_{gvb} – коефіцієнт впізнання бренду; P_{wap} і P_{wvk} – % нових/повторних замовників і користувачів; S_{grk} – середній P_{ROI} за типом реклами; I_{gck} і P_{wck} – індекс та % конверсії цілей за типом реклами; P_{wvk} і P_{wvz} – % відвідувань за типом реклами засобу; K_{wcz} – коефіцієнт конверсії цілей за типом засобу.

$$I_{gyk}(w) = \frac{P_{wcv}(w)}{P_{wvk}(w)}, \qquad K_{gvb} = \frac{N_{ubq}+N_{utv}}{N_{uaq}+N_{utv}}, \quad (4)$$

де $P_{wvk}(w)$ – функція визначення % відвідувань від реклами w; $P_{wcv}(w)$ – функція визначення % конверсії цілей для відвідувань від w; $I_{gyk}(w)$ – функція визначення індексу якості реклами w; N_{uaq} – загальна кількість (к-ть) користувацьких запитів інтелектуально-інформаційного пошуку (ІІП) за ключовими словами; N_{utv} – к-ть прямих відвідувань веб-сайту; N_{ubq} – к-ть запитів ІІП із назвою бренду.

Наявність модуля інтеграції текстового контенту M_{dcp} скорочує затрати на КЛС-модераторів та авторів контенту, автоматизуючи/реалізуючи деякі їх роботи/функції як збір контенту з множини різних достовірних джерел, його розпізнавання, фільтрування, збереження, форматування, аналіз, анотування, класифікація тощо:

$$M_{dcp} =< P_{glt}, P_{gst}, P_{ght}, K_{gvb}, K_{uzv}, P_{uav}, P_{uzv}, S_{gnc}, P_{wvv}, S_{gpv}, S_{gtp} >, \quad (5)$$

де P_{glt}, P_{gst} та P_{gst} – % повторних відвідувань користувача з попереднього відвідування $>t_2$, в межах $[t_1; t_2]$ при $t_1<t_2$ та $<t_1$ днів відповідно; K_{gvb} – коефіцієнт впізнавання бренду; P_{uav} та P_{uzv} – % нових/повторних та зацікавленості відвідувачів; S_{gnc} – середня к-ть кліків на рекламі за N_{wvr} відвідувань; P_{wvv} – показник відмов для однієї веб-сторінки; S_{gpv} – середня к-ть переглядів веб-сторінки за відвідування; S_{gtp} – середня тривалість перебування на веб-сторінці.

$$P_{wvv} = \frac{N_{vnp}}{N_{inp}}, \quad S_{gnc} = \frac{N_{wcr}}{N_{wav}} \cdot N_{wvr}, \quad K_{uzv} = \frac{N_{wad}}{N_{wav}}, \quad P_{uzv} = \frac{N_{wzv}}{N_{wvk}}. \quad (6)$$

де N_{inp} – к-ть відвідувань веб-сторінки напряму; N_{vnp} – к-ть односторінкових відвідувань для веб-сторінки; N_{wvr} – к-ть відвідувань для аналізу; N_{wav} – загальна к-ть відвідувань; N_{wcr} – середня к-ть кліків на рекламі; N_{wad} – загальна к-ть дій на сторінці; N_{wvk} і N_{wzv} – загальна к-ть всіх та зацікавлених користувачів.

Наявність модуля управління текстовим контентом скорочує затрати на модераторів/адміністраторів, які оновлюють веб-сайт та формують правила кешування/пошуку популярних інформаційних блоків:

$$M_{dvm} =< K_{wis}, P_{wep}, P_{gum}, P_{gup}, P_{gur}, P_{gus}, P_{gub}, P_{gul}, P_{wep}, K_{wdu}, S_{wdu} >, \quad (7)$$

де K_{wis} – показник внутрішнього ІІП; P_{wep} – % видання сторінки з помилкою; P_{gum} і P_{gup} – % мобільних користувачів та з високошвидкісним підключенням до Інтернет; P_{gur} і P_{gus} – % користувачів з низькою/середньою/високою роздільною здатністю дисплею і з конкретною операційною системою; P_{gub} і P_{gul} – % користувачів з конкретним браузером та з підтримкою англійської і/або української мови; K_{wdu} – показник кількості користувачів, переглядів та відвідувань сторінки. Показник S_{wdu} є базовим модуля керування контентом:

$$S_{wdu} =< N_{svt}, N_{sut}, N_{spt}, N_{spv} >, \quad (8)$$

де N_{spv} і N_{spt} – середня к-ть переглядів сторінки за відвідування та за конкретний Δt час; N_{sut} – середня к-ть унікальних користувачів за конкретний Δt час; N_{svt} – середня к-ть відвідувань за конкретний Δt час. Показник внутрішнього пошуку по сайту:

$$K_{wis} =< N_{nns}, P_{uts}, P_{ksp}, P_{bus}, P_{cuss}, P_{pop}, P_{ucs}, S_{vrs}, P_{uos}, P_{uns}, P_{unr}, P_{uur}, S_{nup}, T_{svs}, P_{uis}, P_{nrp}, K_{wps} >,$$

де N_{nns} – к-ть нульових результатів пошуку; P_{uts} і P_{ksp} – % користувачів, що перебували на сторінці $> t$ часу та переглянути $> k$ сторінок після здійсненого пошуку; P_{bus} і P_{cus} – % здійснених покупок та % покупців серед користувачів, що використовують пошук; P_{pop} – % відмов після відвідування однієї сторінки як результату пошуку; P_{ucs} – % конверсії від користувачів, що використовують пошук; P_{unr} і P_{uur} – % користувачів, які не використовують і використовують пошук; S_{nup} – середня к-ть сторінок, переглянутих відвідувачами після пошуку; T_{svs} – середній час перебування на сайті для відвідування після пошуку; P_{uns} і P_{uos} – % відвідувачів, які проводять декілька пошуків на протязі відвідування та які покинули сайт після перегляду результатів пошуку; S_{vrs} – середня к-ть результатів пошуку; P_{uis} – % відвідувань з пошуком; P_{nrp} – % нульових результатів пошуку, зокрема,

$$P_{wep} = \frac{N_{wep}}{N_{wpp}}, \quad P_{nrp} = \frac{N_{nps}}{N_{vps}}, \quad K_{wps} = \frac{N_{wsv}}{N_{wns}}, \quad (9)$$

де N_{wpp}, N_{wep} і N_{vps} – к-ть всіх переглянутих сторінок, виданих з помилкою та переглянутих сторінок з пошуком відповідно; N_{nps} – к-ть нульових результатів пошуку; N_{wns} і N_{wsv} – відвідування без пошуку та із пошуком.

Наявність модуля інтелектуального аналізу текстових потоків контенту скорочує час/затрати/персонал/ресурси на своєчасне оперативне отримання релевантного унікального актуального контенту, що призводить до зростання обсягів цільової аудиторії КЛС, зокрема сприяє зростанню економічного ефекту від впровадження:

$$M_{was} =< S_{wcc}, S_{wtv}, S_{wnv}, P_{wuv}, P_{wnv} >, \quad (10)$$

де S_{wcc} – середній коефіцієнт конверсії; S_{wtv} – середня тривалість відвідування; S_{wnv} – середня к-ть переглядів за відвідування; P_{wuv} – % унікальних замовників/відвідувачів/користувачів; P_{wnv} – % нових замовників веб-сайту.

Згідно відстеження подій K_{as} та взаємодії з сайтом K_{du} аналізують:

$$K_{usa} = \alpha(K_{wdu}, K_{was}) =< P_{vcu}, P_{sau}, P_{siu}, I_{wdx} >, \qquad I_{wdx} = \frac{R_{wcv}+R_{wec}}{N_{upv}}, \quad (11)$$

де P_{siu} – % взаємодії з сайтом (наприклад, коментування, голосування, реєстрація, авторизація, підписка тощо); P_{sau} – % користувачів, які активізують різні події (наприклад, клік на рекламу, запуск функції, пауза тощо); P_{vcu} – % користувачів, взаємодіючих з різними типами подання контенту (перегляд наступного спілкування, панорамування, масштабування, тощо); I_{wdx} – значення міри корисності відповідно сторінки/сайту/КЛС/контенту; N_{upv} – к-ть унікальних переглядів сторінки; R_{wec} – прибуток від е-бізнесу; R_{wcv} – значення міри корисності відвідування користувачів (на основі транзакцій) та мети відвідування користувачів (на основі корисності цілей).

Аналіз успішності/результативності/оперативності пошуку по сайту:

$$K_{iip} =< P_{wuv}, R_{ecc}, S_{wcv}, P_{wip}, P_{wcv}, N_{wvt}, R_{wcv}, R_{wec}, N_{wtr}, N_{wcv}, I_{ssp} >, \quad (12)$$

де P_{wuv} – значення корисності відвідування P_{wuv} сайту/сторінки; R_{ecc} – рейтинг конверсії в е-бізнесі для КЛС відповідної задачі ОПМ; S_{wcv} – значення середньої корисності; P_{wip} – значення прибутку е-бізнесу для КЛС відповідної задачі ОПМ; P_{cv} – значення досягнутої конверсії відвідувань сайту/сторінки КЛС:

$P_{wuv} = \frac{R_{wcv}+R_{wec}}{N_{wvt}}$, $R_{ecc} = \frac{N_{wtr}}{N_{wvt}} \cdot 100\%$, $S_{wcv} = \frac{R_{wcv}+R_{wec}}{N_{wcv}+N_{wtr}}$, $P_{wip} = R_{wcv} + R_{wec}$, $P_{wcv} = \frac{N_{wcv}}{N_{wvt}} \cdot 100\%$,

де N_{wvt} – к-ть відвідувань; R_{wec} – корисність е-бізнесу; R_{wcv} – корисність мети; N_{wtr} – к-ть транзакцій; N_{wcv} – к-ть конверсії.

Для залучення нових відвідувачів та зростання обсягів постійної цільової аудиторії застосовують розрахунок впливу на прибуток ІІП по сайту I_{ssp}:

$$I_{ssp} = (R_{ssv} - R_{snv}) \cdot N_{ssv}, \quad (13)$$

де N_{ssv} – к-ть відвідувань з ІІП; R_{snv} і R_{ssv} – корисність відвідування без і з ІІП.

Тематика множини ключових слів є одним із основних показником ІІП для ідентифікації конкретного контенту сторінки. Для множин ключових слів, що збільшують значення конверсії, оптимізують інвестиції. Значення повернення на інвестиції (P_{ROI}) має бути позитивним ($N_{Inc} > N_{Exp}$), тобто:

$$P_{\text{ROI}} = \frac{N_{Inc}-N_{Exp}}{N_{Exp}} \cdot 100\% > 0, \quad P_{\text{ROIvp}} = \frac{(N_{Inc} \cdot A_{Inc})/100-N_{Exp}}{N_{Exp}} \cdot 100\%, \quad (14)$$

де N_{Exp}– витрати; N_{Inc} – прибуток; A_{Inc} – розмір прибутку. Тоді знаходять на скільки $>q\%$ коштів без ризику отримати $P_{\text{ROI}}<0$ можна затратити на конкретне ключове слово в рекламі. Для розрахунку обсягів коштів на залучення користувачів застосовують:

$$C_{amax} = \frac{\frac{N_{Inc} \cdot A_{Inc}}{100}}{\frac{P_{\text{ROIvp}}}{100}+1}, \qquad C_{cmax} = C_{amax} \cdot \frac{R_{ecc}}{100}. \quad (15)$$

Метод визначення ефективності/якості сайту КЛС для розв’язку задачі ОПМ:

Етап 1. Формулювання та ідентифікація корисності відповідно до цілей цільової аудиторії згідно вхідних даних з кортежу X.

Етап 2. Активація звітів функціонування КЛС з кортежу Y вихідних даних:

Крок 1. Визначити необмежене число цілей (≈4 цілі на кожний профіль цільової аудиторії).

Крок 2. Ідентифікувати оптимальний обсяг відвідувань/часу кінцевого користувача/замовника для успішної конверсії.

Крок 3. Проаналізувати обсяги вкладу кожної цілі в загальний прибуток.

Крок 4. Поєднати цілі за категоріями/напрямами/видами.

Крок 5. Сформувати окремі множини транзакцій як відповідних цілям.

Етап 3. Підтримка різних маркетингових кампаній/замовників через M_{dmr}.

Етап 4. Підтримка опрацювання службового контенту сайту модулем M_{dvm}.

Етап 5. Оновлення профілів цільової аудиторії згідно підтримки зворотного зв'язку через модуль M_{dmr}, та аналіз дій користувачів через модуль M_{dvm}.

Етап 6. Інтегрування контенту з різних джерел через M_{dcp} згідно досягнутих цілей та опрацювання його через модуль M_{was}.

Етап 7. Періодична перевірка, чи досягаються цілі та зростає прибуток згідно поставлених цілей. Якщо він спадає, перехід до етапу 1, інакше до етапу 2.

Класифікований список вхідного потоку контенту X з множиною відповідних властивостей розмежовує учасників проекту через їх типізацію та обмеження прав доступу в залежності від контенту: постійні користувачі, потенційні відвідувачі, лінгвісти, аналітики статистики, адміністратори, модератори контенту/правил, автори унікального контенту, інформаційний ресурс як джерело контенту тощо. Типізована структура шаблону вхідного потоку контенту з множиною відповідних властивостей сприяє визначити основні функціональні вимоги до сайту/КЛС та її типової структури та чітко окреслити нефункціональні можливості, класифікувати джерела, розрахувати частоти та відповідні обмеження/умови інтеграції з типового джерела:

$$X =< X_a, X_s, X_q\ , X_f, X_s, X_w, X_b, X_d, X_k, X_v, X_u, X_r, X_t, X_o >, \quad (16)$$

де X_a – URL-адреси джерел для баз даних (БД) фільтрів КЛС; X_s – контент як результат інтеграції з різних за наперед визначеним списком URL-адрес джерел X_a без наперед визначеної структури згідно релевантних тематичних запитів; X_q – тематичні запити відвідувачів/користувачів сайту КЛС у вигляді множини ключових слів або стійких словосполучень; X_f – фактичні дані постійних користувачів/профілів та множина правил дозволених дій в межах відповідного типу користувача КЛС; X_s – статистичні дані дій/подій/явищ суб'єктів/об'єктів КЛС розв'язку відповідної задачі ОПМ та правила збору/збереження/аналізу статистики в певних проміжках часу функціонування КЛС; X_w – статистичні дані функціонування КЛС; X_b – вміст БД/СД контенту/правила/фільтрів/анотацій тощо КЛС; X_d – різного виду лінгвістичні словники в залежності від призначення КЛС для розв'язку конкретної задачі ОПМ; X_k – множина персоналізованих/анонімних відгуків і коментарів користувачів до відповідного контенту КЛС; X_v – кортеж результатів персоналізованих/анонімних голосувань постійних/потенційних користувачів щодо контенту КЛС; X_u – статистичні персоналізовані індивідуальні дії користувачів КЛС; X_r – множина зовнішньої/внутрішньої реклами тематичного контенту; X_t – тематичні стікери інформаційного контенту (курси валют, анонси, дайджести, погода, анекдоти, гороскоп тощо); X_o – кортеж опцій налаштування та зміни конфігурацій КЛС/сайту.

Наповнення кортежу вихідного потоку даних Y згідно призначення КЛС для розв'язку конкретної задачі ОПМ напряму залежить від змісту вхідного класифікованого потоку контенту X з наперед визначеною множиною властивостей в залежності від взаємодії із сайтом відповідних типів учасників проекту:

$$Y =< Y_c, Y_q, Y_a, Y_v, Y_s, Y_p, Y_t, Y_r, Y_o, Y_k >, \quad (17)$$

де Y_c – текстовий контент як інформаційний продукт або результат надання відповідної інформаційної послуги розв'язку конкретної задачі ОПМ на сайті КЛС; Y_q – множина змістовно згенерованих/кешованих сторінок як результат тематичних

запитів/ІІП користувачів/відвідувачів сайту КЛС; Y_a – анотації/дайджести/реферати на текстовий тематичний контент; Y_v – кортеж статистики взаємодії користувачів/ відвідувачів з сайтом; Y_s – кортеж змісту профілів постійних користувачів КЛС згідно персоналізованої статистики Y_v для відповідного генерування індивідуального портрету користувача/аудиторії в певні проміжки часу; Y_p – кортеж змістовного рекомендованого контенту сайту, персоналізованого під конкретного постійного користувача згідно профіля/дій/взаємозв'язку із КЛС в певні проміжки часу; Y_t – множина тем/рубрик контенту з можливістю поновлення згідно результатів останніх ІІП/запитів від постійних користувачів сайту; Y_o – схема взаємозв'язків текстового тематичного контенту за відповідною класифікацією (актуального, релевантного, авторського, застарілого, популярного, подібного, останньо-переглянутого, часто-переглянутого, послідовно за певним найчастіше переглянутого, довше переглянутого, найчастіше переглянутого з пошукових систем або внутрішнього ІІП, переглянутого типовою групою користувачів тощо); Y_r – множина результату рейтингування контенту за наперед визначеною шкалою у межах відповідної класифікації ранжування; Y_k – множина маркованого оцінювання та рейтингування коментарів користувачів як ступінь дозволу опублікування на сайті/сторінці при необхідності з позначкою заборони для конкретного дописувача писати подальші коментарі та ранжування за ступенем довіри всіх дописувачів. Список вихідного потоку контенту, його основні ознаки та відповідна класифікація, ІТ генерування/ підтримка/аналіз сприяє визначенню чітких загальних функціональних вимог реалізації КЛС для розв'язку будь-якої задачі ОПМ.

Отже розроблено модель КЛС опрацювання україномовного тексту та визначено функціональні вимоги та загальні засади проектування подібних систем. Запропоновано вдосконалену ІТ опрацювання інформаційних ресурсів як інтеграція, управління та супровід контенту на основі розроблених метрик оцінювання ефективності функціонування КЛС для розв'язку різних задач ОПМ.

У третьому розділі розроблено загальну структуру КЛС опрацювання текстового контенту українською мовою та концептуальну схему/модель функціонування типової КЛС на основі моделювання взаємодії основних процесів та компонентів ІС на основі визначених в попередньому розділі колекцій вхідних та вихідних даних моделі системи. Здійснено моделювання основних процесів ОПМ в КЛС. Розроблена та описана формальна модель КЛС для опрацювання україномовного текстового контенту, що дало змогу визначити основні структурні елементи та оператори опрацювання природної мови на кожному рівні аналізу тексту як графемного (ГА), морфологічного (МА), лексичного (ЛА), синтаксичного (СА), семантичного (СЕМ), референційного, структурного, онтологічного та прагматичного (ПА). У зв'язку зі складністю морфології української мови детальна увага приділена вдосконаленню моделей та методів графемного та морфологічного аналізів україномовного текстового контенту на основі продукційних правил, граматики Хомські, скінчених автоматів та теорії множин. Розроблені продукційні правила ідентифікації/генерування різних форм українських дієприкметників на основі формальної породжувальної граматики Хомські. Також вдосконалені методи лексичного та синтаксичного аналізів україномовного тексту на основі логіки

предикатів, продукційних правил, теорії дерев та граматики Хомські. Наведені приклади моделювання процесів розв'язку типових задач ОПМ як КЛС ідентифікації вірусних заголовків новин та виправлення граматичних та стилістичних помилок.

Модель процесу лінгвістичного аналізу україномовного тексту M_{lat} подаємо як:

$$M_{lat} =< X, W, C, K, Y, D, S_{IAC}, S_{LA}, S_{\nu}, S_{\varpi_1}, S_{\varpi_2}, S_{\rho_1}, S_{\rho_2}, S_{\rho_3}, S_{\rho_4}, S_{\upsilon}, \nu, \varpi_1, \varpi_2, \rho_1, \rho_2, \rho_3, \rho_4, \upsilon >,$$

де X – вхідні дані в КЛС з різних джерел інформації W; Y – вихідний релевантний контент з КЛС як результат ІІП згідно запитів користувачів/відвідувачів; S_{LA} – процес лінгвістичного аналізу контенту як складової ІАТПК-підсистеми S_{IAC}; S_{ν} – процес генерування/модифікації правил функціонування всіх модулів від модератора КЛС; S_{ϖ_1} – процес наповнення неструктурованої БД інтегрованим контентом X; S_{ϖ_2} – модуль наповнення структурованої БД на основі опрацьованого інтегрованого контенту C; S_{ρ_1} і S_{ρ_2} – процеси генерування результатів згідно запитів відвідувачів і користувачів; S_{ρ_3} – процес опрацювання кешу для формування звітів на популярні запити від користувачів КЛС; S_{ρ_4} – процес наповнення/модифікації кешу; S_{υ} – процес генерування статистичних результатів функціонування КЛС/модулів та діяльності користувачів D; ν – оператор генерування/модифікації правил функціонування всіх модулів від модератора КЛС; ϖ_1 – оператор наповнення неструктурованої БД інтегрованим контентом X; ϖ_2 – оператор наповнення структурованої БД на основі опрацьованого інтегрованого контенту C; ρ_1 і ρ_2 – оператори генерування результатів згідно запитів відвідувачів і користувачів; ρ_3 – оператор опрацювання кешу для формування звітів Y на популярні запити від користувачів; ρ_4 – оператор наповнення/модифікації кешу даними K; υ – оператор генерування статистичних результатів функціонування КЛС/модулів та діяльності користувачів:

$$S_{LA} =< X, Y, C, D, R, \alpha, \beta, \gamma, \delta, \lambda, o, \iota, \varsigma, \mu >, \quad Y = \mu \circ o \circ \varsigma \circ \iota \circ \lambda \circ \delta \circ \gamma \circ \beta \circ \alpha, \qquad (18)$$

де X – вхідний текстовий масив даних; Y – кортеж вихідного опрацьованого тексту згідно призначення КЛС; C – множина проміжного контенту, який опрацьовується на відповідному рівні в КЛС; D – допоміжні словники; R – множина правил опрацювання; α – оператор ГА; β – оператор МА; γ – оператор ЛА; δ – оператор СА; λ – оператор СЕМ; o – оператор онтологічного аналізу; ι – оператор референційного аналізу; ς – оператор структурного аналізу; μ – оператор ПА.

Основний процес лінгвістичного аналізу текстового контенту подано:

$$Y = \mu(C_\mu, D_\mu, R_\mu, o(C_o, D_o, R_o, \varsigma(C_\varsigma, D_\varsigma, R_\varsigma, \iota(C_\iota, D_\iota, R_\iota, , \lambda(C_\lambda, D_\lambda, R_\lambda, \delta(C_\delta, D_\delta, R_\delta, \gamma(C_\gamma, D_\gamma, R_\gamma, \beta(C_\beta, D_\beta, R_\beta, \alpha(C_\alpha, D_\alpha, R_\alpha, X))))))))), \qquad (19)$$

де множини контенту $C = \{C_\mu, C_o, C_\varsigma, C_\iota, C_\lambda, C_\delta, C_\gamma, C_\beta, C_\alpha\}$, лінгвістичних словників $D = \{D_\mu, D_o, D_\varsigma, D_\iota, D_\lambda, D_\delta, D_\gamma, D_\beta, D_\alpha, \}$ та множини продукційних/асоціативних правил $R = \{R_\mu, R_\iota, R_o, R_\varsigma, R_\lambda, R_\delta, R_\gamma, R_\beta, R_\alpha\}$.

Основний лінгвістичний процес опрацювання текстової україномовної інформації для розв'язку конкретної задачі ОПМ складається з дев'яти етапів:

Eman 1. Графемний аналіз α текстової україномовної інформації X:

$$C_\alpha = \alpha(X, D_\alpha, R_\alpha), \quad C_\alpha = \alpha_7 \circ \alpha_6 \circ \alpha_5 \circ \alpha_4 \circ \alpha_3 \circ \alpha_2 \circ \alpha_1, \qquad (20)$$

де X – вхідний текстовий масив даних; α – оператор ГА; C_α – графемна структура вхідного тексту; D_α – графемні словники та бібліотеки; R_α – правила ГА; α_1 –

оператор оптичного розпізнавання символів; α_2 – оператор графемного розбору вхідного тексту X на розділи, абзаци та речення; α_3 – оператор графемного розбору лінгвістичних ланцюжків на окремі слова; α_4 – оператор формування множини нерозпізнаних ланцюжків; α_5 – оператор ідентифікації та маркування нерозпізнаних ланцюжків як числа, дати, незмінних зворотів, скорочень, власних та географічних назв тощо; α_6 – оператор маркування нетекстових ланцюжків як спецсимволи, формули, рисунки, таблиці тощо; α_7 – оператор генерування маркованої лінійної послідовності слів C_α із службовими знаками та зв'язками.

Eman 2. Морфологічний аналіз β текстового контенту C_β полягає в ідентифікації, аналізі та визначенні форми і структури слів, зокрема:

$$C_\beta = \beta(C_\alpha, D_\beta, R_\beta), \quad C_\beta = \beta_3 \circ \beta_2 \circ \beta_1 \text{ або } C_\beta = \beta_3 \circ \beta_4 \circ \beta_1, \qquad (21)$$

де β_1 – оператор морфологічної сегментації графемно-розпізнаного ланцюжка символів (слів/лексем); β_2 – оператор лематизації лексем; β_3 – оператор розмічування частин мови для сегментованих слів; β_4 – оператор стемінгу слів.

Продукційні правила ідентифікації/генерування українських дієприкметників:

I. Формування граматичних значень: $\{D_K \to D_K(x,y)\}$, де $x = (act/pas)$; $y = (pres/past)$, наприклад, $\{D_K \to D_K(pas,pres); D_K \to D_K(act,pres), \ldots\}$.

II. Аналіз морфем: $\{D_K(act,pres) \to O'(t,\bar{d},a_3)C(act,pres,a_3)\Phi; \quad D_K(act,past) \to O'(\bar{t},d,a_3)C(act,past,a_3)\Phi; \quad D_K(pas,pres) \to O'(t,d-\bar{d},a_3)C(pas,pres,a_3)\Phi; \quad D_K(pas,past) \to O'(t,d-\bar{d},a_3)C(pas,past,a_3)\Phi\}$, де O, C, Φ – позначення різних морфем без опису.

III. Розкладання дієслівної основи: $\{O'(\overline{atem}) \to O(\overline{atem})T,; \quad O'(\overline{d},\overline{\emptyset})C(x,y) \to O(\overline{d},\overline{\emptyset})C_dC(x,y,I); O'(atem) \to O(atem)\}$, де T – тематичний елемент (ТЕ) *-u(i,ї)-/-a(я)-/-ол(р)о-*; $\overline{atem}$ –значення ознаки a_4, відмінне від $atem$, тобто $(a/i/\tilde{a}/\tilde{\imath}/o)$, C_d – суфікс дієслова; $\overline{\emptyset}$ – будь-яке значення ознаки, відмінне від $\emptyset$; x та y повинні задовольняти наступній умові: при $x = pas$ необхідно, щоб $y = pres$.

IV. Ідентифікація ТЕ: $\{(\tilde{a})T\alpha \to O(\tilde{a})\zeta; \quad O(\tilde{\imath})T\alpha \to O(\tilde{\imath})\zeta; \quad O(a)T \to O(a)a+; \quad O(i)T \to O(i)i+; \quad O(o)T \to O(o)o+,; \quad O(\overline{d},II,a)TC(act,pres) \to O(\overline{d},II,a)a + C(act,pres); \quad O(d-\overline{d},I,a)TC(pas,pres) \to O(d-\overline{d},I,a)a + C(pas,pres); \quad O(d-\overline{d},I,i)TC(pas,pres) \to O(d-\overline{d},I,i) + C(pas,pres); \quad (\tilde{a},II)T\beta \to O(\tilde{a},II)a + \xi; \quad O(\tilde{\imath},I)T\beta \to O(\tilde{\imath},I) + \xi\}$, де ζ та ξ – довільна голосна та приголосна; + – межа між морфемами.

V. Утворення дієслів відповідною морфемою: $\{O(\text{I},y)C_d \to O(\text{I},y)$ува +; $O(\text{I},y)C_d \to O(\text{I},y)$ова +; $O(\tilde{y})C_d \to O(\tilde{y}); \quad O(\overline{t},d,\text{н})C_d \to O(\overline{t},d,\text{н}) + C(pas,past); \quad O(t,d,\text{н})C_dC(pas,pres) \to O(t,d,\text{н})$ну $+ C(pas,pres)\}$.

VI. Ідентифікація суфікса: $\{C(act,past,I-II) \to$ л +; $O(atem)C(act,pres,I) \to$ уч +; $O(\overline{atem})YC(act,pres,I) \to$ юч +; $O(atem)C(act,pres,II) \to$ ач +; $O(\overline{atem})YC(act,pres,II) \to$ яч +; $C(pas,pres/past,I-II) \to$ н +; $C(pas,pres/past,I-II) \to$ т +; $O(atem)C(pas,pres/past,I-II) \to$ ен +; $O(\overline{atem})YC(pas,pres/past,I-II) \to$ єн +; $O(atem)C(pas,pres/past,I-II) \to$ ува +; $O(\overline{atem})YC(pas,pres/past,I-II) \to$ юва +; $C(pas,pres/past,I-II) \to$ овува +; $O(atem)C(pas,pres/past,I-II) \to$ ова +; $O(\overline{atem})YC(pas,pres/past,I-II) \to$ йова +; $O(\overline{atem})X'C(pas,pres/past,I-II) \to X'$ьова+$\}$, де Y – будь-який суфікс/ТЕ; X' – м'яка приголосна, X – довільної приголосна.

VII. Вибір форми дієприкметника $(f/\bar{f})$ і флексії: $\{\Phi \to \Phi(f);\ \Phi(f,s) \to$ ого, им, ому; $\Phi(f,m) \to$ ий; $\Phi(f,w) \to$ ою, ої; $\Phi(f,\bar{s}) \to$ им, ими, их; $\Phi \to \Phi(\overline{f});\ \Phi(\overline{f},w) \to$ а, $y;\ \Phi(\overline{f},k) \to$ е; $\Phi(\overline{f},\bar{s}) \to i,;\ C(pas)\Phi(\overline{f}) \to$ о$\}$.

VIII. Ідентифікація основи на основі словника: $\{O(t-\bar{t},\ d-\overline{d},\ I,\ atem,$ у$) \to$ автоматиз+, буд+, мал′+, …; $O(t-\bar{t},\ \overline{d},\ I,\ atem,\ \emptyset) \to$ вес+, …; $O(t,\ d-\overline{d},\ II,\ \tilde{\imath},\ \emptyset) \to$ втрач+, …; $O(\bar{t},\ \overline{d},\ I,\ a,\ \emptyset) \to$ втруч+, …; $O(t,\ d-\overline{d},\ I,\ \tilde{\imath},\ \tilde{y}) \to$ дослідж+, …; $O(\bar{t},\ d,\ I,\ \tilde{\imath},\ \tilde{y}) \to$ запізн+, …; $O(t,\ \overline{d},\ I,\ a,\ \emptyset) \to$ кох+, …; $O(t,\ \overline{d},\ II,$ ĩ, $\emptyset) \to$ люб+, …; $O(t,\ \overline{d},\ I,\ atem,\ \emptyset) \to$ нес+, …; $O(t,\ d,\ I,\ atem,$ у$) \to$ побуд+, розфарб+, …; $O(t,\ d,\ II,$ ĩ, $\emptyset) \to$ поділ+, …; $O(t,\ d,\ I,\ atem,\ \emptyset) \to$ привес+, …; $O(\bar{t},\ \overline{d},\ I,\ \tilde{a},\ \emptyset) \to$ сміj+, стогн+, …; $O(t,\ \overline{d},\ I,\ a,\ \emptyset) \to$ спит+, …; $O(\bar{t},\ d,\ I,\ atem,$ н$) \to$ усміх+, …; $O(t,\ \overline{d},\ I,\ atem,\ y) \to$ фарб+, …; $O(t,\ d,\ I,$ о, $\emptyset) \to$ мол+, …; $O(\bar{t},\ d,\ I,$ і, $\emptyset) \to$ змарн+, …; …$\}$.

IX. Основні морфонологічні правила: $\{\alpha_1 + \to \alpha_1 + j\alpha_2;\ j +$ и $\to$ і; о$Z + C(pas,pres) + \Phi \to aZ + C(act,pres) + \Phi$; с' + W → ш + W; в' + W → вл' + W; б' + W → бл' + W; д' + W → дж' + W; т'+ W → ч + W; …; д + W → д' + W; с + W → с' + W; …; нн + Ф → н + о$\}$, де α_1 та α_2 – довільні голосні; j – позначення звуку [j] (*йот*); Z – довільна послідовність не довша за 3 символи; W = *-е*(*є*)*н-*, *-у*(*ю*)*ва-*, *-ова-*, *-овува-*.

X. Графічно-орфографічні правила: $\{j + a \to$ я, $ja \to$ я; $j +$ у → ю, jу → ю; $j +$ е → є, jе → є; …; X′ + $a \to$ X + я; X′ + у → X + ю; X′ + и → X + і; X′ + і → X +; X′ + е → X + є$\}$.

XI. Стирання показника межі між морфемами: $\{A + B \to AB\}$, де A і B – будь-які морфеми, що до $A + B$ непридатне жодне з правил груп IX-X.

Eman 3. Лексичний аналіз γ текстового контенту C_γ у проміжному етапі аналізу послідовності лексем для генерування дерева розбору на рівні СА:

$$C_\gamma = \gamma(C_\beta, D_\gamma, R_\gamma),\quad C'_\gamma = \gamma_2 \circ \gamma_1,\ C'_\gamma = \gamma_5 \circ \gamma_4 \circ \gamma_3 \text{ або } C'_\gamma = \gamma_5 \circ \gamma_4, \qquad (22)$$

де γ_1 – оператор сегментації мовлення для ідентифікації/уточнення слів/словосполучень/лексем після МА; γ_2 – оператор розпізнавання мовлення або мовлення-у-текст; γ_3 – оператор оптичного розпізнавання символів як друга частина після ГА та МА для уточнення некоректних моментів розпізнавання з врахуванням розпізнаних сусідніх лексем; γ_4 – оператор токенізації/сегментації слів як підготовка даних для побудови дерева розбору при СА; γ_5 – текст-у-мовлення.

Eman 4. Синтаксичний аналіз δ текстового контенту C_δ полягає в побудові дерева розбору залежностей слів (рис. 2) в послідовності лексем на основі їх категорій:

$$C_\delta = \delta(C_\gamma, D_\delta, R_\delta),\quad C_\delta = \delta_3 \circ \delta_2 \circ \delta_1, \qquad (23)$$

де δ_1 – оператор реалізації індукції граматики; δ_2 – оператор ідентифікації/ліквідації неоднозначності меж або порушення речення; δ_3 – оператор синтаксичного парсингу фраз/речень для побудови дерева СА. Правила формулювання україномовних фраз:

I. Вибір структури: $\{R \to \#\tilde{S}_{x,y,н,w}\ \tilde{V}_{y,теnер,w}\#\}$, де $\tilde{V}$ – дієслівна група, $\tilde{S}$ – іменна група, x – рід, y – однина/*од*, або множина/*мн*; z – відмінок, w – особа.

II. Іменна група: $\{\tilde{V}_{x,y,z,3} \to \tilde{S}_{x,y,z,3}\ \tilde{S}_{x',y',р,w};\quad \tilde{S}_{x,y,z,3} \to A_{x,y,z}\ \tilde{S}_{x,y,z,3};\ K_1\tilde{S}_{x,y,z,w}\ K_2 \to K_1 S^{займ}_{x,y,z,w} K_2,\ K_1 \neq A_{x,y,z},\ K_2 \neq \tilde{S}_{z'};\ \tilde{S}_{x,y,z,3} \to S_{x,y,z}\}$.

III. Дієслівна група: $\{\widetilde{V}_{y,теnер,w} \to V_{y,}$тепер$_{,w}\ \tilde{S}_{x',y',}$зн$_{,w'}\ \tilde{S}_{x'',y'',}$ор$_{,w''};\ \tilde{V}_{y,теnер,w} \to V_{y,теnер,w}\ \tilde{S}_{x',y',ор,w'}\ \tilde{S}_{x'',y'',зн,w''};\quad \tilde{V}_{y,теnер,w} \to V_{y,теnер,w}\ \tilde{S}_{x',y',}$зн$_{,w'}$; $\tilde{V}_{y,теnер,w} \to V_{y,теnер,w}\ \tilde{S}_{x',y',ор,w'}\}$.

IV. Підстановка слів: $\{S_{\text{ч},y,z} \to \text{син}_{y,z}, \ldots;\ S_{\text{ж},y,z} \to \text{посм}i\text{шка}_{y,z}, \ldots;\ S_{сер,y,z} \to щастя_{y,z}, \ldots;\ S^{\text{займ}}_{\text{х,од,z,1}} \to \text{я}_z;\ S^{займ}_{x,од,z,2} \to ти_z;\ V_{y,тепер,w} \to наповнити_{y,тепер,w}, \ldots;\ A_{x,y,z} \to веселий_{x,y,z}, безмежний_{x,y,z}, мій_{x,y,z}, твій_{x,y,z}, \ldots\}$.

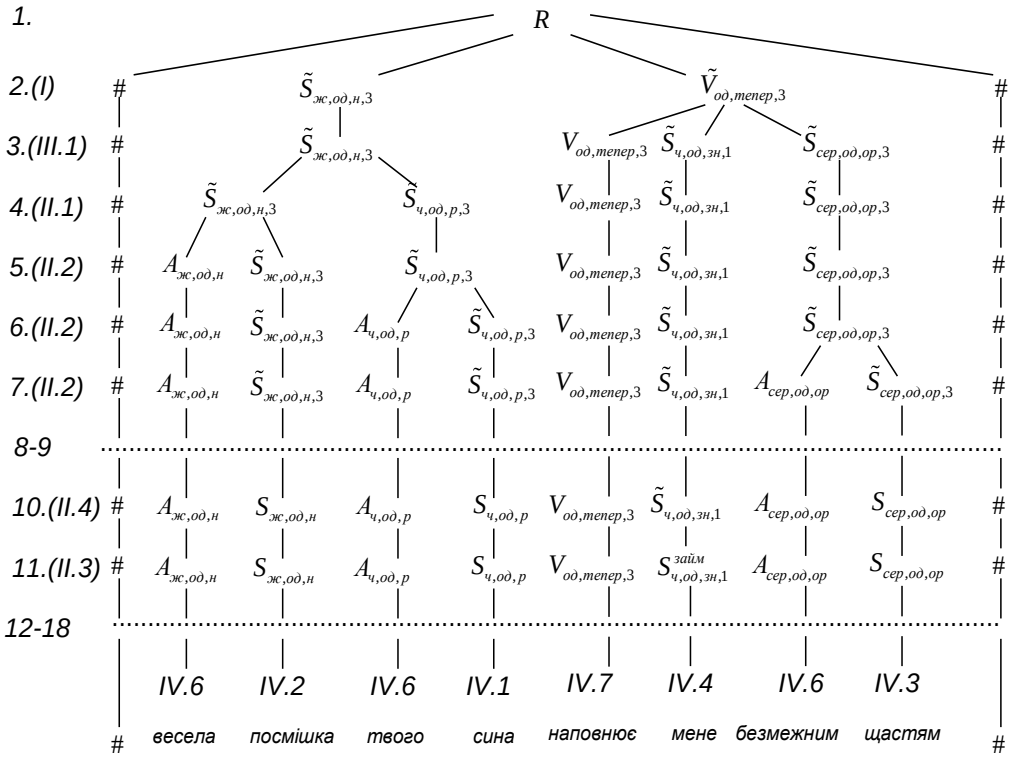


Рис. 2. Приклад побудови дерева розбору залежностей слів речення

Етап 5. Семантичний аналіз λ україномовного тексту C_λ полягає

$$C_\lambda = \lambda(C_\delta, D_\lambda, R_\lambda), \quad C_\lambda = \lambda_2 \circ \lambda_1, \tag{24}$$

де λ_1 – оператор ідентифікації лексичної семантики з генеруванням колекції значень кожної лексеми тексту; λ_2 – оператор ідентифікації реляційної семантики взаємозалежностей змісту лексем тексту.

Етап 6. Референційний аналіз ι ідентифікації міжфразових єдностей C_ι.

$$C_\iota = \iota(C_\lambda, D_\iota, R_\iota). \tag{25}$$

Референційний аналіз часто є частиною СЕМ. Для українських текстів при аналізі великих корпусів текстів найкраще виносити як окремий етап (наприклад для аналізу переписки соціальної групи/спільноти в соціальних мережах або інших діалогів для ідентифікації логічних змістовних зав'язків між дописами різних учасників із-за суб'єктивізму мовлення кожного.

Етап 7. Структурний аналіз ς україномовного тексту C_ς на основі ступеня збігу лексичних термінологічних одиниць єдності фрагментів тексту. Часто є частиною СЕМ для коротких текстів/повідомлень, або взагалі не використовують. Для великих корпусів текстів як додатковий етап ліквідації маркованої неточності при СЕМ.

$$C_\varsigma = \varsigma(C_\iota, D_\varsigma, R_\varsigma) \text{ або } C_\varsigma = \varsigma(C_\lambda, D_\varsigma, R_\varsigma). \tag{26}$$

Етап 8. Онтологічний аналіз о текстового контенту $C_\text{о}$ на основі або частиною результатів СЕМ та референційного/структурного аналізів при потребі:

$$C_\text{о} = \text{о}(C_\varsigma, D_\text{о}, R_\text{о}),\ C_\text{о} = \text{о}(C_\iota, D_\text{о}, R_\text{о}) \text{ або } C_\text{о} = \text{о}(C_\lambda, D_\text{о}, R_\text{о}). \tag{27}$$

Eman 9. Прагматичний аналіз μ текстового контенту C_μ застосовують для визначення структури тексту з врахуванням контексту речень при формуванні абзаців, розділів та діалогів. ПА є суттєвим доповненням СЕМ, референційного та структурного аналізів, якщо вони не сприяли ліквідації маркованої неточності.

$$Y = \mu(C_\mu, D_\mu, R_\mu, C_\lambda, [C_o, C_\varsigma, C_\iota],), Y = \mu_2 \circ \mu_1, \quad (28)$$

де μ_1 – оператор ідентифікації семантики поза окремими реченнями/фразами; μ_2 – оператор опрацювання текстів через вищого рівня додатки ОПМ, наприклад, для імітування розумної поведінки та очевидного розуміння природної мови.

Отже, розроблено модель лінгвістичного аналізу україномовного тексту через вдосконалення графемного/морфологічного/лексичного/синтаксичного аналізів.

Четвертий розділ присвячено розробленню загальної схеми/моделі конвеєра функціонування КЛС на основі вдосконалених методів опрацювання інформаційних ресурсів як інтеграція, супровід та управління контентом, а також розробленням із вдосконалених методів інтелектуального та лінгвістичного аналізу текстового потоку з використанням технології машинного навчання (рис. 3).

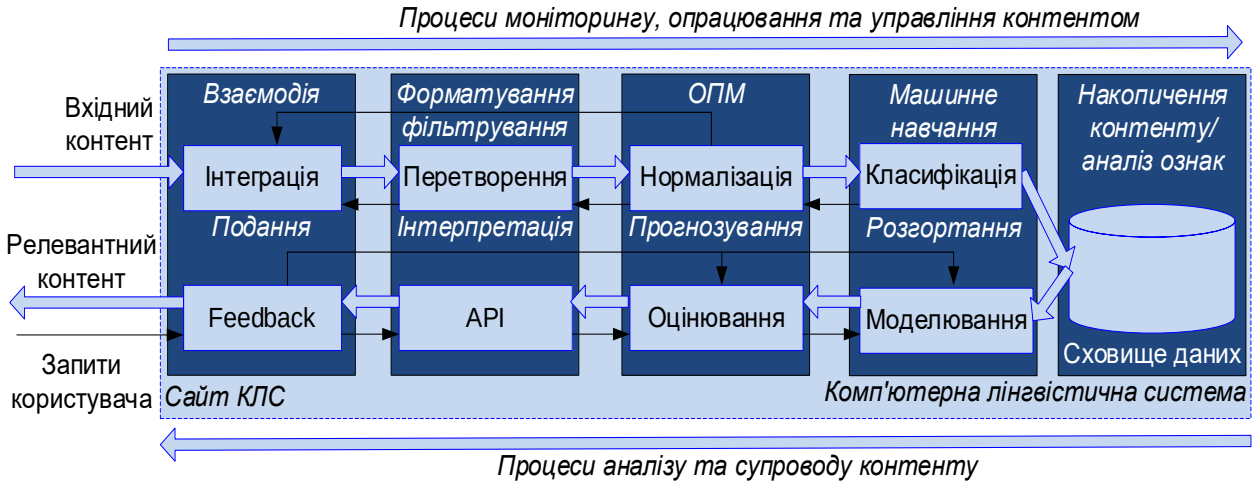


Рис. 3. Схема конвеєра функціонування КЛС

На основі зворотного зв’язку із користувача та вихідних даних моделі МН цільова аудиторія взаємодіє з КЛС, що сприяє адаптації обраної моделі навчання. П’ять стадій відповідних процесів визначають основні архітектурні принципи побудови типових КЛС. Для процесів моніторингу, опрацювання та управління контентом це – взаємодія, форматування/фільтрування, ОПМ, МН та накопичення даних в СД. Для процесів аналізу та супроводу контенту відповідно це – аналіз ознак, розгортання, прогнозування, інтерпретація та подання контенту/результату. На стадії взаємодії розроблено набір правил інтеграції контенту з множини достовірних джерел в певні часові проміжки. Також паралельно розроблено набір правил перевірки даних введених від користувача КЛС як попередній етап для стадії форматування/ фільтрування згідно колекції наперед закладених модератором правил та контенту з СД. Наступна стадія ОПМ є проміжним етапом для МН та накопичення даних. Стадія МН реалізована через SQL-запити та модулі. Процес супроводу більш простіший для реалізації, ніж етап управління, особливо при аналізі результатів ОПМ, при якому створено додаткові лексичні ресурси та артефакти (словники, перекладачі, регулярні вирази тощо), від яких напряму залежить ефективність функціонування КЛС (рис. 4).

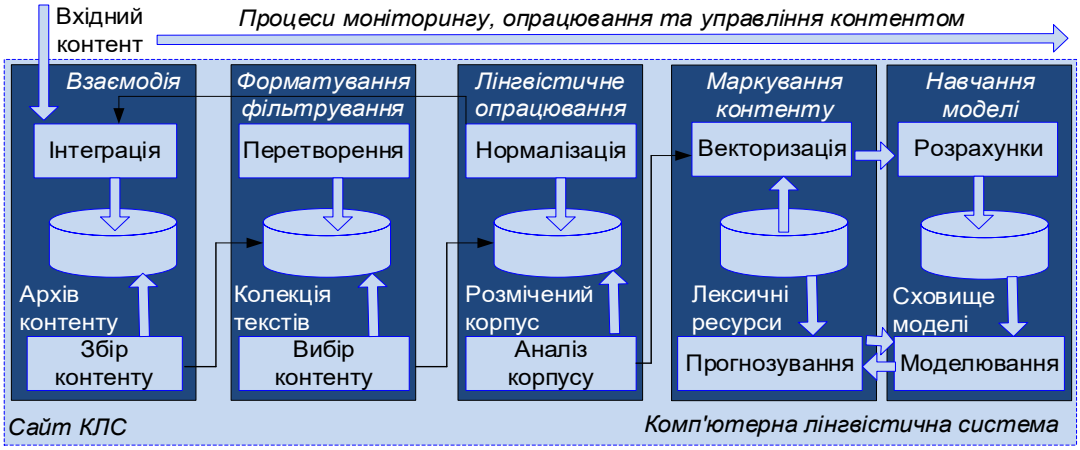


Рис. 4. Схема конвеєра опрацювання україномовного текстового контенту

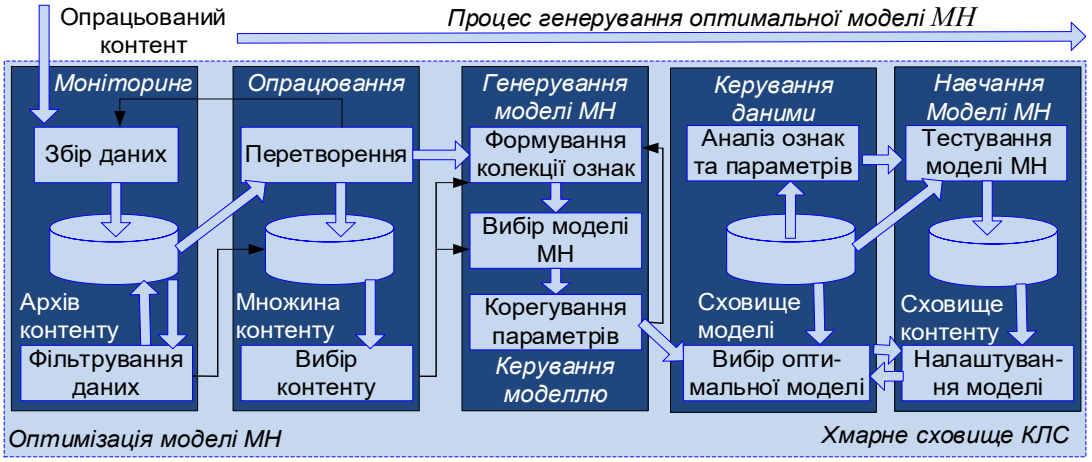


Рис. 5. Процес конвеєра машинного навчання

Процес переходу від неопрацьованого тексту до розгорнутої моделі МН складається з послідовності додаткових перетворень контенту. По-перше, вхідний текстовий контент перетворюється в вхідний корпус як колекція текстів, накопичується і зберігається в СД. Вхідний контент далі групують, фільтрують, форматують, лінгвістично опрацьовують, маркують, нормалізують та перетворюють у вектори для подальшого опрацювання. При остаточному перетворенні моделі (рис. 5) тренують на векторному корпусі, створюють узагальнене подання вихідного контенту для подальшого застосування при розв'язку конкретної задачі ОПМ.

Вдосконалено методи ОПМ на основі розроблених 82 регулярних виразів (РВ) узгодження з шаблонами при ГА та понад 2000 РВ морфологічного аналізу україномовних текстів. Визначені основні допустимі операції РВ як об'єднання та диз'юнкція символів/ланцюжків/виразів, оператори лічби та прецедентності, а також анкери присутності/відсутності символів в регулярних виразах. Визначені основні етапи токенізації та нормалізації українського тексту каскадами простих підстановок РВ та кінцевих автоматів. Реалізовані та описані алгоритми сегментації та нормування слова, сегментації речення та модифікований стемінг Портера як ефективний спосіб ідентифікації афіксів лем для можливості розмічування аналізованого слова. Модифікований алгоритм стемінгу Портера оснований на пошуку/перевірки отриманих проміжних результатів з деревом флексій (щоб не

перебирати всі можливі флексії) та з вмістом тематичних словників основ з множиною РВ-правил ідентифікації ознак (класифікація за частинами мови).

Етап 1. Ідентифікувати наступну лексему як слово w_i ($w_s = w_i$).

Етап 2. Перевірити з словником стоп-слів $D_{w_{sw}}$ чи w_s є службовим словом. Якщо так, то $i = i + 1$ та перейти до етапу 1, інакше – до етапу 3.

Етап 3. Перейти до кінця слова w_s. Розпізнати флексію f_1^i в w_s із всіх можливих (обирається найдовша, наприклад, у w_s=*текстова* обираємо закінчення f_1^i=*ова*, а не $f_1^i \neq a$) з РВ типу слів $R_{adjectival}$, R_{noun}або R_{verb} та при наявності видалення флексії f_1^i.

Етап 4. Збереження флексії f_1^i у тегу слова w_i.

Етап 5. Маркувати w_i. як тип $m_{adjectival}^{w_i}$, $m_{noun}^{w_i}$або $m_{verb}^{w_i}$ відповідно.

Етап 6. Знаходження видаленої флексії f_1^i в дереві флексій $T_{flection}$ (обирається найдовша). Перевірка вмісту піддререва $T_{flection}^{f_1}$ з наявним закінченням слова f_2^i ($f = f_2^i + f_1^i$). Якщо w_s. закінчується на f_2^i та має відповідник в $T_{flection}^{f_1}$, то зберігаємо в $f_i = f$ та видаляємо в w_s.

Етап 7. Отриману основу w_s початкового слова w_i перевіряємо із змістом словника основ D_{w_s} слів української мови. При відсутності відповідника зберігаємо $< w_i, w_s >$ в додатковому тимчасовому проміжному словнику $D_{<w_i,w_s>}$ для модератора та перехід до етапу 1, інакше перехід до етапу 4.

Етап 8. Аналіз флексії та наявності/відсутності чергування літер в основі/флексіях слів $< w_i, w_s >$ і аналогу основи слова в D_{w_s} згідно відповідного РВ-правила МА для ідентифікації додаткових ознак аналізованого слова w_i.

Етап 9. Дописування ідентифікованих лінгвістичних ознак розпізнаної частини мови до тегу слова w_i типу $m_{adjectival}^{w_i}$, $m_{noun}^{w_i}$або $m_{verb}^{w_i}$ відповідно. Збереження результатів у відповідний словник D_{w_i} аналізованого тексту.

На відмінну від класичного алгоритму Портера модифікований є адаптованим саме для української мови та дає точний результат в межах 85-93% випадках в залежності від якості, стилю, жанру тексту та відповідно наповнення словників КЛС. Всього для МА україномовних іменників реалізовано біля 1300 правил опрацювання суфіксів та закінчень з врахування чергувань літер, прикметників – 99 РВ-правил, а дієслів - понад 800 РВ-правил. Описано алгоритм мінімальної редакційної відстані рядків українських текстів як мінімальна кількість операцій, необхідних для перетворення одного в інший. Також для ідентифікації стійких словосполучень як ключових слів розроблено алгоритм обчислення метрики максимальної правдоподібності для моделі 2-грам та 3-грам на основі аналізу основ слів. Для прогнозу умовної ймовірності наступної основи слова застосовуємо Марківське припущення (ймовірність слова залежить від попереднього). Причому якщо ключовими словами є множина іменників або прикметник з іменником, тоді інші слова як дієслова, дієприкметники тощо вважатимуться додатковими розділювачами як інші знаки пунктуації, які розмежовують стійкі словосполучення як потенційно ключові слова. Порядок основ є не важливим для української мови.

Етап 1. Вхідний текст опрацювати та розбити на окремі фрази (речення) $R_1 R_2 \ldots R_m$, маркуючи кожний початок-закінчення відповідним тегом .

Ліквідувати всі не алфавітні символи. Великі літери перевести в малі. Видалити службові слова при необхідності (для певних задач ОПМ).

Етап 2. Застосувати стемінг Портера для отримання відповідно послідовності основ слів $x_{i1}x_{i2}\ldots x_{in_i}$ основ слів $\forall R_i$ з врахуванням нормалізації слів.

Етап 3. Отримати на вхід запити $Q_1Q_2\ldots Q_k$ як послідовності слів шуканих даних. Знайти $\forall Q_j$ для кожного слова $y_{j1}y_{j2}\ldots y_{jk_j}$ основу через стемінг.

Наприклад для фрази пошукового запиту Q_j:

Методи та засоби опрацювання інформаційних ресурсів систем електронної контент комерції

y_{j1}	y_{j2}	y_{j3}	y_{j4}	y_{j5}	y_{j6}	y_{j7}	y_{j8}	y_{j9}	y_{j10}
метод	*та*	*засіб*	*опрац*	*інформ*	*ресурс*	*систем*	*електрон*	*контент*	*комерц*
58	190	25	62	122	83	170	89	408	300

Етап 4. Провести статистичний аналіз входження основ слів та послідовностей основ слів запиту у аналізований текст.

Основи слів аналізованого тексту		x_{i1} *метод*	x_{i2} *та*	x_{i3} *засіб*	x_{i4} *опрац*	x_{i5} *інформ*	x_{i6} *ресурс*	x_{i7} *систем*	x_{i8} *електрон*	x_{i9} *контент*	x_{i10} *комерц*
x_{i1}	*метод*	0	8	0	6	0	0	0	0	1	0
x_{i2}	*та*	2	0	5	1	7	0	2	0	0	1
x_{i3}	*засіб*	0	2	0	14	0	0	0	0	0	0
x_{i4}	*опрац*	0	0	0	0	46	0	0	1	3	4
x_{i5}	*інформ*	0	0	0	0	0	64	9	0	0	0
x_{i6}	*ресурс*	0	7	0	0	0	0	0	1	0	0
x_{i7}	*систем*	0	8	0	1	0	0	0	21	0	0
x_{i8}	*електрон*	0	0	0	0	0	0	0	0	72	10
x_{i9}	*контент*	0	10	0	0	0	0	0	0	0	73
x_{i10}	*комерц*	0	6	0	0	0	0	0	0	176	0

Етап 5. Знайти вірогідності появи 2-грам в аналізованому тексті. В кожному рядку значення ділимо на y_{ji}, де i номер рядка після нормалізації.

Основи слів тексту		x_{i1} *метод*	x_{i2} *та*	x_{i3} *засіб*	x_{i4} *опрац*	x_{i5} *інформ*	x_{i6} *ресурс*	x_{i7} *систем*	x_{i8} *електрон*	x_{i9} *контент*	x_{i10} *комерц*	y_{ji}
x_{i1}	*метод*	0	0,18	0	0,1	0	0	0	0	0,02	0	58
x_{i2}	*та*	0,01	0	0,03	0,005	0,035	0	0,01	0	0	0,005	190
x_{i3}	*засіб*	0	0,08	0	0,16	0	0	0	0	0	0	25
x_{i4}	*опрац*	0	0	0	0	0,74	0	0	0,016	0,048	0,064	62
x_{i5}	*інформ*	0	0	0	0	0	0,52	0,074	0	0	0	122
x_{i6}	*ресурс*	0	0,084	0	0	0	0	0	0,012	0	0	83
x_{i7}	*систем*	0	0,047	0	0,006	0	0	0	0,124	0	0	170
x_{i8}	*електрон*	0	0	0	0	0	0	0	0	0,81	0,112	89
x_{i9}	*контент*	0	0,025	0	0	0	0	0	0	0	0,179	408
x_{i10}	*комерц*	0	0,02	0	0	0	0	0	0	0,053	0	300

Отримані матриці в більшості випадків будуть розрідженими. Фраза та різні варіації (множина/однина та відмінки) P(система електронної контент комерції):

P(електрон|систем)P(контент|електрон) P(комерц|контент) =0,124×0,81×0,179=0,01797876.

Вдосконалено метод СЕМ на основі таксономії концептів, що задає синтаксис української мови як кореневий концепт онтології: $Concepts_{\mu}: < R_{Snt} > \rightarrow C'_{\mu}$.

При СЕМ для ідентифікації множини семів відповідного україномовного тексту та їх взаємозв'язку спочатку на основі результатів СА будують семантичний граф відношень лінгвістичних одиниць з врахування частин мови слів:

$$C'_{\mu} = \lambda(C_{\lambda}, D_{\lambda}, R_{\lambda}, Concepts_{\mu}),\ Concepts_{\mu} =< C_{WrdCmb}, C_{SntCmb} >,$$

де C_{WrdCmb} – кортеж концептів утворення словосполучень; C_{SntCmb} – кортеж концептів генерування речень в українській мові. Кортеж C_{WrdCmb} подано як:

$$C_{WrdCmb} =< Sgn_1^{Wrd}, Sgn_2^{Wrd}, Sgn_3^{Wrd}, Sgn_4^{Wrd} >,$$

де Sgn_i^{Wrd} – кортеж властивостей генерування словосполучень:

$$Sgn_1^{Wrd} =< Sgn_{Lxc}^{I}, Sgn_{Snt}^{I} >,$$

$$Sgn_2^{Wrd} =< Sgn_{Nou}^{II}, Sgn_{Adc}^{II}, Sgn_{Nmr}^{II}, Sgn_{Prn}^{II}, Sgn_{Vrb}^{II}, Sgn_{Adv}^{II} >,$$

$$Sgn_3^{Wrd} =< Sgn_{Crd}^{III}, Sgn_{Inf}^{III} >, Sgn_4^{Wrd} =< Sgn_{SmWd}^{IV}, Sgn_{CmWd}^{IV} >,$$

де Sgn_{Lxc}^{I} – кортеж лексичних ознак генерування словосполучень; Sgn_{Snt}^{I} – кортеж синтаксичних ознак генерування словосполучень; Sgn_{Nou}^{II} – кортеж іменних властивостей; Sgn_{Adc}^{II} – кортеж прикметникових властивостей; Sgn_{Nmr}^{II} – кортеж властивостей числівників; Sgn_{Prn}^{II} – кортеж займенникових властивостей; Sgn_{Vrb}^{II} – кортеж дієслівних властивостей; Sgn_{Adv}^{II} – кортеж прислівних властивостей; Sgn_{Crd}^{III} – кортеж сурядних та Sgn_{Inf}^{III} – кортеж підрядних властивостей; Sgn_{SmWd}^{IV} – кортеж простих та Sgn_{CmWd}^{IV} – кортеж складних властивостей.

Кортеж Sgn_{Crd}^{III} описує складові властивості речення зв'язку:

$$Sgn_{Crd}^{III} =< Sgn_{AdCm}^{Crd}, Sgn_{CnCm}^{Crd}, Sgn_{DvCm}^{Crd} >,$$

де Sgn_{AdCm}^{Crd} – кортеж властивостей розділового, Sgn_{CnCm}^{Crd} – кортеж властивостей єднального та Sgn_{DvCm}^{Crd} – кортеж властивостей протиставного зв'язків.

$$Sgn_{Inf}^{III} =< Sgn_{CtCm}^{Inf}, Sgn_{MgCm}^{Inf}, Sgn_{AgCm}^{Inf} >,$$

де Sgn_{CtCm}^{Inf} – кортеж властивостей узгодження; Sgn_{MgCm}^{Inf} – кортеж властивостей керування; Sgn_{AgCm}^{Inf} – кортеж властивостей прилягання. Кортеж концептів генерування речень: $C_{SntCmb} =< Sgn_1^{Snt}, Sgn_2^{Snt}, Sgn_3^{Snt}, Sgn_{SnMb}^{Snt} >$, де властивості генерування речень згруповані у Sgn_i^{Snt} – кортежі властивостей генерування речень; Sgn_{SnMb}^{Snt} – кортеж властивостей ідентифікації членів речення;

$$Sgn_1^{Snt} =< Sgn_{NrSn}^{I}, Sgn_{PrSn}^{I}, Sgn_{InSn}^{I} >, Sgn_2^{Snt} =< Sgn_{EmNt}^{II}, Sgn_{EmCl}^{II} >,$$

$$Sgn_3^{Snt} =< Sgn_{SlSt}^{III}, Sgn_{ClSt}^{III} >, Sgn_{SnMb}^{Snt} =< Sgn_{MnStMb}^{SnMb}, Sgn_{SdStMb}^{SnMb} >,$$

де Sgn_{NrSn}^{I} – кортеж властивостей генерування розповідних речень; Sgn_{PrSn}^{I} – кортеж властивостей генерування питальних речень; Sgn_{InSn}^{I} – кортеж властивостей генерування спонукальних речень; Sgn_{EmNt}^{II} – кортеж властивостей генерування емоційно-нейтральних речень; Sgn_{EmCl}^{II} – кортеж властивостей генерування емоційно-забарвлених речень; кортеж концептів утворення Sgn_{SlSt}^{III} простих та Sgn_{ClSt}^{III} складних речень; Sgn_{MnStMb}^{SnMb} – кортеж властивостей ідентифікації головних членів речення; Sgn_{SdStMb}^{SnMb} – кортеж властивостей ідентифікації другорядних членів речення; $Sgn_{NrSn}^{I} =< Sgn_{AfSt}^{NrSn}, Sgn_{NgSt}^{NrSn} >$; Sgn_{AfSt}^{NrSn} – кортеж властивостей генерування стверджувальних речень; Sgn_{NgSt}^{NrSn} – кортеж властивостей генерування заперечних речень. Для генерування простого речення Sgn_{SlSt}^{III} аналізують ознаки:

$$Sgn_{SlSt}^{III} =< Sgn_1^{SlSt}, Sgn_2^{SlSt}, Sgn_3^{SlSt}, Sgn_4^{SlSt}, Sgn_5^{SlSt}, Sgn_6^{SlSt}, Sgn_7^{SlSt}, Sgn_8^{SlSt} >.$$

де Sgn_i^{SlSt} – кортеж властивостей генерування простих речень.

Отже, вдосконалено методи ОПМ на основі розроблених регулярних виразів графемного та морфологічного аналізу україномовного тексту, модифікованого

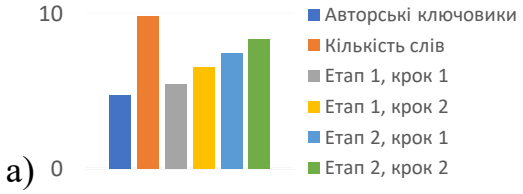
алгоритму стемінгу Портера та метод обчислення метрики максимальної правдоподібності для моделі 2-грам та 3-грам на основі аналізу основ слів.

У п'ятому розділі проаналізовані результати експериментальної апробації розроблених методів та засобів лінгвістичного інтелектуального аналізу текстів українською мовою на основі розроблення методів ідентифікації ключових слів, визначення стійких словосполучень, тематичної класифікації тексту та виявлення дублювання тексту. Розглянуто особливості методу синтаксичного аналізу україномовного текстового контенту, спрямованого на 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 2-х етапах (табл. 1), де A (всього визначених ключових слів при заданій вазі слова), B (утворених значущих слів без займенника та дієслів), C (співпадіння слів зі авторським списком), D (точність співпадіння ідентифікованих ключових слів з авторським), E (додатково визначені ключові слова, але не визначені автором публікації). На етапі 1 дослідження для кроку 1 (аналіз повних статей) та кроку 2 (статті без метаданих як анотація, авторські ключові слова та список літератури) провадилося без застосування МН, а на етапі 2 – з МН. Найкращих результатів за критерієм щільності досягає метод аналізу статті без метаданих. Автор статті часто визначає більшу кількість слів (A_2) та меншу кількість ключових слів (A_1), ніж реально присутні в тексті науково-технічної публікації (рис. 6). На відміну від відомих синтаксичних аналізаторів, запропонований метод забезпечує самовдосконалення та самонавчання модуля визначення ключових слів за рахунок механізму ідентифікації значущих статистичних параметрів у визначених модератором межах. Розроблено систему на сайті victana з можливістю обрання зі списку мов аналізованого тексту (http://victana.lviv.ua/index.php/kliuchovi-slova).

Таблиця 1. Статистичні дані дослідження змісту науково-технічних публікацій

Назва	Вага слова	Етап 1					Етап 2				
		A	B	C	D	E	A	B	C	D	E
Крок 1	≥ 1	5,46	3,92	2,51	2,08	1,74	7,43	7,03	3,27	3	4,18
	≥ 2	1,08	0,88	0,63	0,59	0,26	2,67	2,64	1,65	1,54	1,12
	≥ 3	0,41	0,38	0,22	0,21	0,16	1,21	1,2	0,85	0,79	0,41
	≥ 4	0,15	0,13	0,09	0,09	0,04	0,46	0,45	0,33	0,31	0,15
	≥ 5	0	0	0	0	0	0	0	0	0	0
Крок 2	≥ 1	6,51	5,02	2,68	2,23	2,37	8,35	7,78	3,25	2,91	4,99
	≥ 2	1,34	1,11	0,74	0,72	0,39	3,12	3,07	1,81	1,67	1,43
	≥ 3	0,51	0,45	0,29	0,27	0,17	1,42	1,4	0,93	0,85	0,54
	≥ 4	0,19	0,17	0,12	0,12	0,05	0,73	0,72	0,45	0,42	0,31
	≥ 5	0,11	0,1	0,06	0,06	0,04	0,33	0,32	0,25	0,23	0,1

Значення A_3 відмінне за значення A_1 на 0,69 (за кількістю, але не за змістом); A_4 від A_1 на 1,74; A_5 від A_1 на 2,66; A_6 від A_1 на 3,58. Значення A_2 відмінне за значення A_3 на 4,36; відповідно A_2 від A_4 на 3,31; A_2 від A_5 на 2,39; A_2 від A_6 на 1,47. Адаптивна зміна параметрів/правил модуля збільшує колекцію ідентифікованих ключових слів майже вдвічі (наприклад, значення A_1 за A_3 більше в 1,144654; A_6 – в 1,750524; A_5 – в 1,557652; A_4 – в 1,36478). Загальний приріст значення, отриманий в залежності від модерації словників складає відповідно для A_3 – 14,46541; A_4 – 36,47799; A_5 – 55,7652; A_6 – 75,05241. При порівнянні A_2 більше за $A_3 \div A_6$ та маємо ланцюг таких значень як 1,7985; 1,5084; 1,3217; 1,176.



а)

б)

Назва	Стовпець	Середньоарифметична к-ь ключових слів	
A_1	Авторські	визначених автором	4,77
A_2	Кількість слів	містять авторські	9,82
A_3	Етап 1, крок 1	ймовірних ключових слів, знайдених модулем на етапі X та кроку Y (рис. 7-8)	5,46
A_4	Етап 1, крок 2		6,51
A_5	Етап 2, крок 1		7,43
A_6	Етап 2, крок 2		8,35

Рис. 6. Результати аналізу понад 300 науково-технічних публікацій

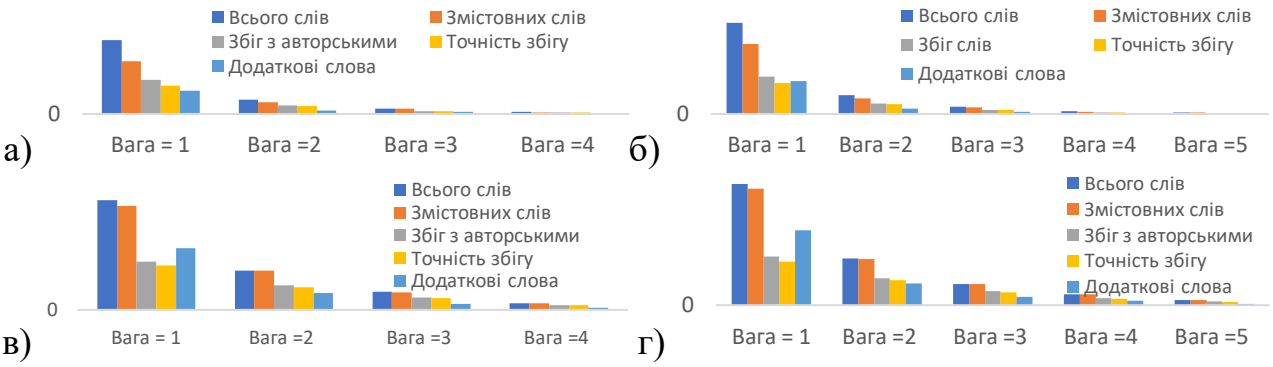


Рис. 7. Отримання значущих слів на етапі: а) 1.1, б) 1.2, в) 2.1 та г) 2.2

Для різних етапів та кроків експерименту 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. Для A_3 найчастіше модуль ідентифікував кількість ключових слів {5, 7, 3} (≥10), хоча розподіл знайдених ключових слів в межах [1;18] слів (окрім 17).

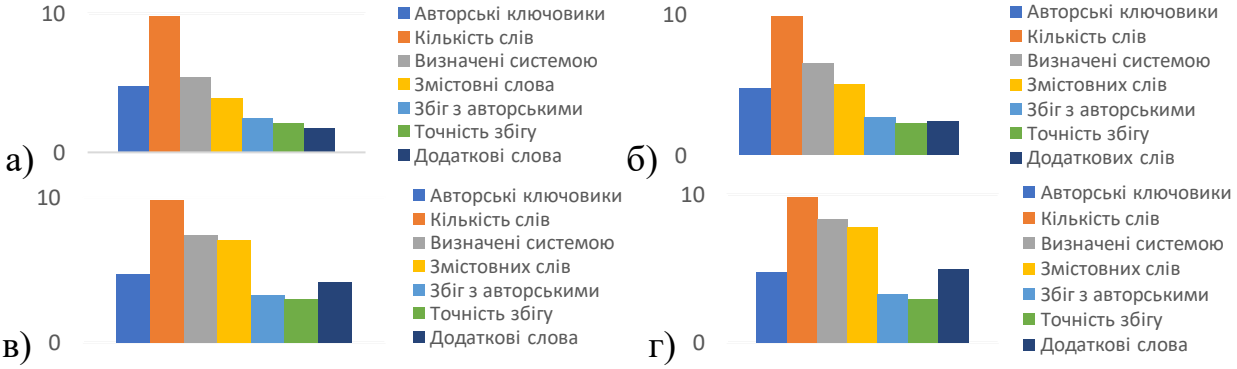


Рис. 8. Середньоарифметична поява слів на етапі: а) 1.1, б) 1.2, в) 2.1 та г) 2.2

Для A_4 найчастіше модуль визначив кількість ключових слів також {5, 7, 3}, хоча розподіл знайдених ключових слів є в межах [1;18] (окрім 17), але збільшилась кількість ідентифікованих слів та досягнуто найбільшого показника надійності. Для A_5 найчастіше модуль ідентифікував кількість ключових слів {7, 6, 5, 10, 8}, хоча розподіл знайдених ключових слів в межах [2;14] (значно звузився діапазон). Для A_6 найчастіше модуль ідентифікував кількість ключових слів {8, 5, 7, 10}, розподіл ідентифікованих ключових слів в межах [3;16] (покращилась точність). Точність визначення ключових слів збільшується в процесі модерації словників та модуля МН. Різниця між кількістю ключових слів, визначених автором та ідентифікованих

модулем, при A_3 складає 44,39919 % (різниця у %). Точність покращується при A_4 – 33,70672 %, значно покращується при A_5 – 24,33809 %, а при A_6 складає 14,96945 %.

Проведений аналіз для відфільтрованих текстів без мета-даних та невідфільтрованих текстів. Отримані середні значення для відфільтрованих текстів $\overline{Per_f} = 0{,}28$ та невідфільтрованих $\overline{Per_0} = 0{,}19$ показують, що фільтрація наукових статей покращує щільність ключових слів у 1,48 раз або на 47,83 % (рис. 9а).

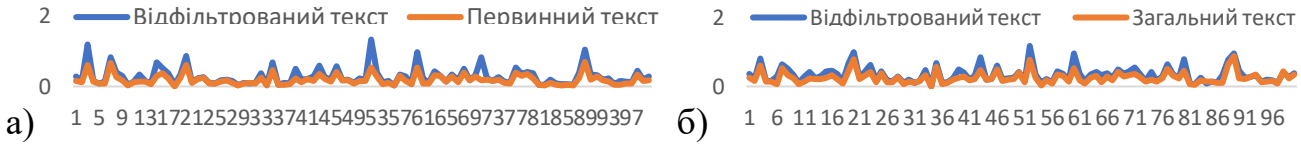


Рис. 9. Результати перевірки статей без уточнення тематичного словника

Отримані значення для текстів $\overline{Per_f^v} = 0{,}34$ та $\overline{Per_0^v} = 0{,}25$ з врахуванням уточнення тематичного словника через МН та поповнення заблокованих слів показують, що фільтрація з одночасною модерацією тематичного словника покращує щільність ключових слів у 1,35 раз або на 35,44 % (рис. 9б). Порівняння значень в первинному авторському тексті $\overline{Per_0} = 0{,}19$ та $\overline{Per_0^v} = 0{,}25$ без/з уточнення тематичного словника відповідно демонструє ефективність модерації тематичного словника у початковому тексті – щільність ключових слів збільшується у 1,34 раз або на 34,33 % (рис. 10а). Порівняння значень в відфільтрованому авторському тексті $\overline{Per_f} = 0{,}28$ та $\overline{Per_f^v} = 0{,}34$ без/з уточнення тематичного словника відповідно демонструє ефективність модерації тематичного словника у відфільтрованому тексті – щільність ключових слів збільшується у 1,23 раз або на 23,14 % (рис. 10б).

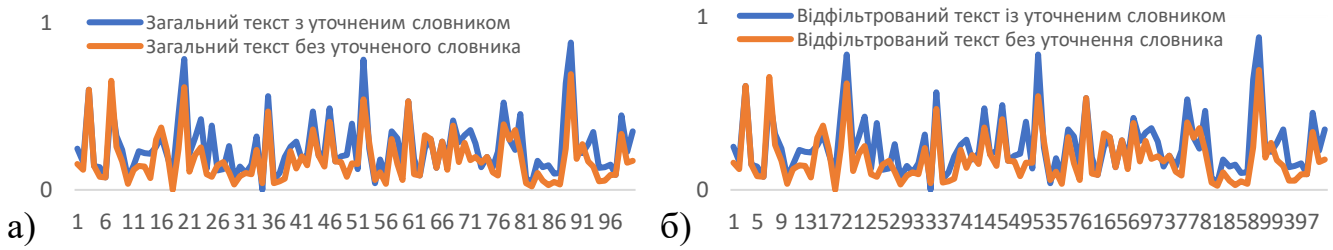


Рис. 10. Результати аналізу статей з різними словниками

Отже, експериментальне дослідження підтвердило достовірність методу – для різних етапів опрацювання первинного тексту середній збіг списків виявлених ключових слів з авторськими змінюється у проміжку 52,6-68,5% (на 9%). 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. Розроблено метод визначення стійких словосполучень при ідентифікації ключових слів текстового контенту в еталонних уривках авторського тексту. Метод полягає у використанні закону Зіпфа при формуванні стійких словосполучень як ключових з врахуванням наступних правил попереднього лінгвістичного опрацювання тексту: вилучення всіх стопових слів; біграми формувати лише в межах знаків пунктуації та слів, які не є дієслово або

займенник (останні вважати знаками пунктуації); дієслова визначати за флексіями; біграми формувати на основі їх основ без врахування їх флексій; визначення прикметників за флексіями та вважати, що прикметники мають бути лише на першому місці у біграмі з україномовних текстів. Розроблено модуль для визначення стійких словосполучень як ключових слів в текстовому контенті. 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 ідентифікації стилю автора тексту на основі аналізу коефіцієнтів лексичного мовлення в еталоні. Метод полягає в порівняльному аналізі авторської атрибуції в статистично опрацьованому доробку автора (еталоні) з довільним аналізованим уривком. Метод оцінює ймовірність приналежності тексту статті автору еталону із аналізом відповідних коефіцієнтів лексичного мовлення як концентрації тексту I_{kt}, зв'язності мовлення K_z, винятковості тексту I_{wt}, синтаксичної складності мовлення K_s та лексичної різноманітності мовлення K_l. Ступень зв'язності мовлення K_z не суттєво зменшується. В 2001 р. змінюється в межах [0,5;1,2], а в 2021 р. – в межах [0,4; 0,9] (рис. 11). Причому метод працює при умові, що авторський еталон вже досліджений – задача ОПМ для формування авторського частотного словника, в тому числі службових/стопових слів.

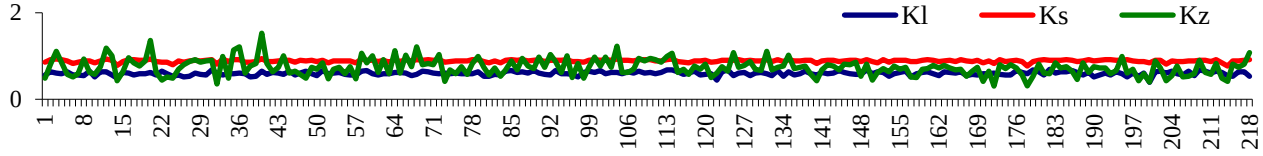


Рис. 11. Аналіз дистрибуції параметрів стилю мовлення K_l, K_s та K_z

Розроблено алгоритм визначення стопових слів текстового контенту на основі лінгвістичного аналізу текстового контенту. Для індивідуального стилю авторського тексту маркерами є службові/стопові слова (наприклад, частки, сполучники, прийменники, слова-паразити, суржик, сленг тощо), які не пов'язані з тематикою статті. Для кожного з уривків проаналізовані та порівняні із еталонним значеннями абсолютні та відносні частоти появи стопових слів. Отже, застосування методу опорних слів дає такі результати: знаходження серед досліджуваних уривків того, що найбільш ймовірно належить до еталону. Інші результати також 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, але розташувало ймовірність приналежності до еталону для уривків у вірному порядку (табл. 2). Більш ймовірно Уривок 4 належить автору шаблону (хоча між результатами 4 та 2 є не суттєва різниця, але вони все таки якщо написані в одному проміжку часу, не належать автору шаблону, якщо в різні проміжки з шаблоном – ймовірність приналежності цьому автору зростає).

Таблиця 2. Коефіцієнти кореляції для стоп-слів

Нова нумерація	Номер статті	R_{e-U}	Частка	Сполучник	Прийменник	R'_{e-U}
1	**4**	**0,7326**	**0,9594**	**0,9544**	**0,5639**	**0,6905**
2	2	0,7066	0,9580	0,5714	0,4928	0,4913
3	1	0,6076	1	0,79	0,72	0,6900
-	3	0,2810	0,8800	0,1624	0,1517	0,2254

Розроблено алгоритм лексичного аналізу україномовних текстів та алгоритм синтаксичного аналізатора текстового контенту. Особливостями алгоритму є адаптація морфологічного та синтаксичного аналізу лексичних одиниць до особливостей конструкцій україномовних слів/текстів. Алгоритми апробованого при ідентифікації значущих стопових слів україномовного тексту на основі регулярних виразів. При парсингу слів враховувалась належність до частини мови та відмінювання в межах цієї частини мови. Для цього реалізовано аналіз флексій слів для класифікації, виділення основи та формування відповідних алфавітно-частотних словників. Наповнення словників в подальшому враховувалися на наступних кроках визначення авторства тексту як розрахунок параметрів та коефіцієнтів авторського мовлення. Програмна реалізація для розв'язку деяких задач ОПМ, як дослідження:

– ключових слів (https://victana.lviv.ua/kliuchovi-slova);
– стійких словосполучень (https://victana.lviv.ua/nlp/stiiki-slovospoluchennia);
– рубрикації текстового контенту (https://victana.lviv.ua/kliuchovi-slova);
– кількісних оцінок мовлення (https://victana.lviv.ua/nlp/linhvometriia);
– стилю автора на основі розрахунків коефіцієнтів стилеметрії та їх порівняння з відповідними коефіцієнтами в еталоні (https://victana.lviv.ua/nlp/stylemetriia);
– відмінностей текстових ознак (https://victana.lviv.ua/nlp/hlotokhronolohiia);
– особливостей стилю текстів на основі N-грам (https://victana.lviv.ua/nlp/n-grams).

Досліджено результати експериментальної апробації запропонованого методу контент-моніторингу для визначення автора в україномовних наукових текстах технічного профілю. Проведено порівняння результатів на множині понад 300 одноосібних робіт технічного спрямування біля 100 різних авторів за період 2001–2021 рр. для визначення чи змінюються і як коефіцієнти різноманітності тексту цих авторів в різні проміжки часу. Розроблено метод ідентифікації потенційного (ймовірного) автора україномовного тексту на основі аналізу коефіцієнтів лексичного авторського мовлення в еталонному уривку авторського тексту. Проведено декомпозицію методу визначення автора на основі аналізу таких коефіцієнтів мовлення як зв’язність мовлення, ступінь синтаксичної складності, лексична різноманітність, індекси концентрації та винятковості тексту. Паралельно проаналізовані такі параметри авторського стилю, як кількість слів у певному тексті, загальна кількість слів цього тексту, кількість речень, кількість прийменників, кількість сполучників, кількість слів із частотою 1, та кількість слів із частотою 10 та більше, а також ключові слова та 3-грами. Для прикладу проаналізовано 3-грами 3-х статей. Для найчастіше вживаних літер частота появи 3-грам з такими початковим літерами буде розподіл майже однаковий (пікові значення на рис. 12а), а для інших літер – ні. Тому доцільно досліджувати лише 3-грами для початкових літер, що рідше зустрічаються в текстах конкретної мови для визначення ступеня належності тексту відповідному автору (наприклад, рис.12б).

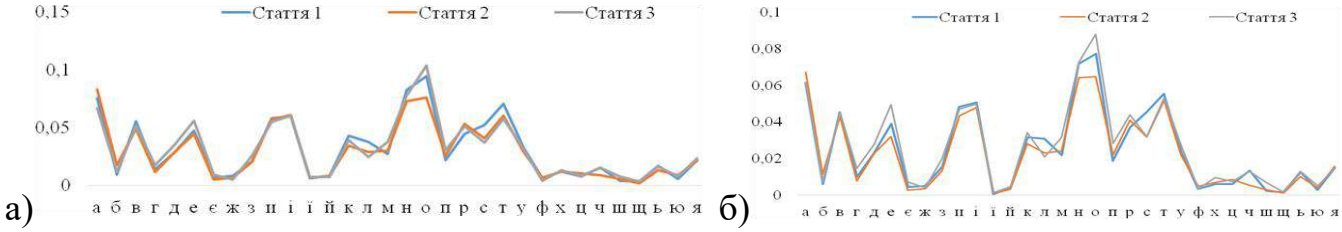


Рис. 12. Графік розподілу частот появи 1-грами та 3-грам в Статтях 1–3

Згідно цих графіків випливає, що Статті (1,2) ймовірніше написані одним автором, хоча Статті (1,3) також могли бути написані одним автором (але це не є істиною). А ось Статті (2,3) точно написані різними авторами. Застосування лінгвостатистичного аналізу 3-грам до множини статей дозволяє сформувати підмножину подібних за лінгвістичними характеристиками публікацій. Накладання додаткових умов у вигляді проведення лінгвостатистичних аналізів (множини ключових слів, стійких словосполучень (табл. 3), стилеметричного, лігвометричного тощо) дозволить значно скоротити підмножину, уточнивши список ймовірніших авторських робіт. Так, аналіз змісту та частоти появи лише службових слів відокремлює Статті (1,3) в різні підмножини, Статті (1,2) залишить в одній. Для Статті 1 проаналізовано 78,4814 % 3-грам, для Статті 2 – 72,6332 % та для Статті 3 – 84,1271 %. Різниця вживання відповідних 3-грам між Статтями (1,2) є R_{12}=56,5254 %, між Статтями (2,3) – R_{23}=69,4271 %, між Статтями (1,3) – R_{13}=62.9839 %. Відповідно Статті (1,2) більш подібні на [6-12]%(R_{23}>R_{12} на 12.9017 %, R_{23} > R_{13} на 6.4432 %, R_{13}> R_{12} на 6.4585 %, тобто R_{23}>R_{13}>R_{12}), ніж Статті (1,3) та (2,3). Чим менше R_{ij}, тим більша ступінь, що статті написані одним і тим же автором. Тоді в випадку Стаття (1,2) більш ймовірніше написана одним автором/колективом, ніж Статті (2,3) та (1,3) відповідно.

Таблиця 3. Список за рейтингом частоти появи стійких словосполучень для Статті 1

FREG			t-тест		LR		X2	
Словосполучення	АЧ	ВЧ	Словосполучення	t	Словосполучення	logL	Словосполучення	X2
система електронний	4	0.088889	система електронний	1.822222	інформаційний технологія	5.03e–1	прийняття рішення	45.000000
інформаційний система	4	0.088889	електронний контент-комерція	1.578091	інтелектуальний система	2.13e–1	система електронний	45.000000
електронний контент-комерція	3	0.066667	розділ науковий	1.319933	інформаційний система	8.36e–2	електронний контент-комерція	32.946429
розділ науковий	2	0.044444	інформаційний система	1.222222	портал науковий	5.58e–2	розділ науковий	29.302326
портал науковий	1	0.022222	прийняття рішення	0.977778	курс технологія	3.31e–2	курс технологія	21.988636
інтелектуальний система	1	0.022222	курс технологія	0.955556	сховище дані	3.31e–2	сховище дані	21.988636
прийняття рішення	1	0.022222	сховище дані	0.955556	прийняття рішення	8.27e–3	портал науковий	14.318182
курс технологія	1	0.022222	портал науковий	0.933333	розділ науковий	1.89e–3	інформаційний система	5.848550
сховище дані	1	0.022222	інтелектуальний система	0.777778	електронний контент-комерція	1.55e–4	інтелектуальний система	3.579545
інформаційний технологія	1	0.022222	інформаційний технологія	0.688889	система електронний	1.37e–6	інформаційний технологія	1.890409

При ідентифікації автора тексту передбачається, що текст відображає індивідуальну манеру письма автора, яка дозволяє відрізнити його від інших. Щоб порівнювати тексти між собою необхідно зіставити тексту деяку числову характеристику, яка була б наближена для текстів одного і того ж автора, і суттєво різнилася б для творів різних авторів. Такою характеристикою може бути щільність розподілу літеросполучень з трьох поспіль символів (3-грам). При експериментальній апробації на основі розроблених різних 4 алгоритмів обчислення ступеня верифікації

автора україномовного тексту із множини можливих отримані значення, які підтверджують, що стиль авторів під номерами x та y досить наближений (понад 90%) на стиль колективних робіт 1–4 відповідно. Також значно зменшили кількість авторів (з 42,02% до 34,04% із загальної кількості 100 учасників проекту з понад 300 статей) з подібністю в стилі мовлення. На рис. 13 подані графіки отриманих результатів при застосуванні алгоритмів для аналізу розробленого методу визначення стилю автора.

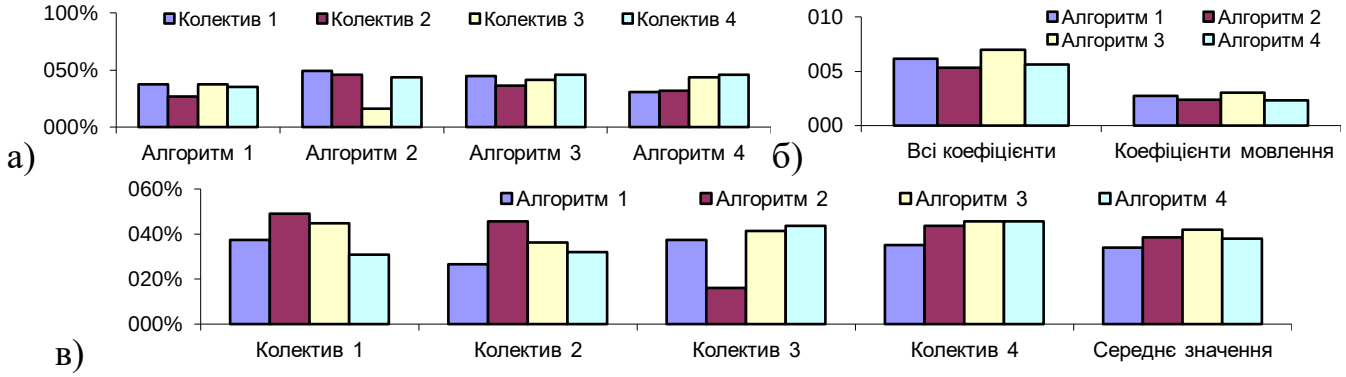


Рис. 13. Аналіз стилю: а – за розробленими алгоритмами; б – з врахуванням ознак мовлення; в – для аналізованих колективних робіт

Далі для визначення стилю автора використано аналіз стопових слів та ключових слів творів авторів, як потрапили до тих 34,04%. Кожна особистість має власний словниковий запас для передачі думки, в тому числі так званих «паразитичних» (тобто, отже, хоча тощо) та службових слів (і, та, й, але, хоч би тощо). На рис. 14 поданий приклад аналізу стилю автора на другому етапі – через аналіз частоти появи службових та ключових слів з врахування різних фільтрів.

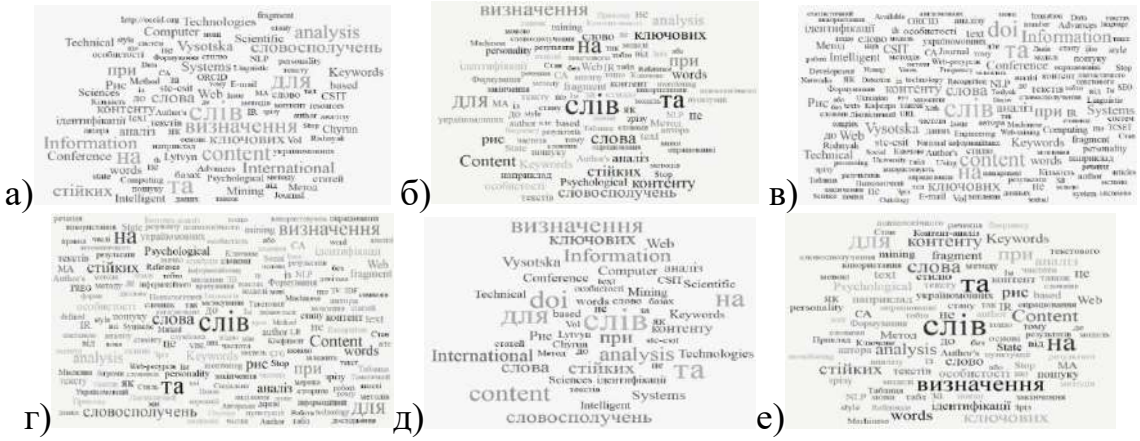


Рис. 14. Дослідження стилю на етапі 2 для тексту з побудовою частотного словника: а – повного зі 100 слів; б – основного зі 100 слів; в – повного з 200 слів; г – основного з 200 слів; д – повного з 50 слів; е – основного з 50 слів

Отже, 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 на 6%. Також 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 на 7%.

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, інформацію про апробацію результатів дисертаційної роботи та впровадження.

ВИСНОВКИ

У дисертаційній роботі вирішено важливу науково-прикладну проблему аналізу та синтезу КЛС для розв’язання різних задач опрацювання україномовного текстового контенту на основі розроблення нових та удосконаленні відомих моделей, методів та засобів ОПМ:

1. Проведено аналіз сучасного стану та перспективи розвитку ІТ опрацювання природної мови, що дало змогу визначити проблему та задачі дослідження, а також сформувати загальні напрями дослідження при відсутності некомерційних КЛС з відкритим кодом для опрацювання україномовного текстового контенту та стандартизованого підходу проектування.
2. Обґрунтовано актуальність розв’язання проблеми аналізу та синтезу КЛС на основі розроблення загальної структури системи опрацювання україномовного текстового контенту, яка 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 ОПМ. Це забезпечило адаптацію процесів ОПМ для аналізу україномовного текстового контенту та на їх основі підвищити точність отриманих результатів на 6-48% в залежності від конкретної задачі ОПМ. Наприклад, для задачі ОПМ визначення ключових слів україномовного тексту щільність ключових слів збільшується в діапазоні [1,23; 1,48] раз або на [23,14; 47,83]% в залежності від якості/точності поповнення тематичного словника через машинне навчання.
3. 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 КЛС для до розв’язку різних задач ОПМ. Розроблені методи та засоби дають можливість будувати КЛС опрацювання україномовного текстового контенту згідно потреб постійної/потенційної цільової аудиторії на основі аналізу історії дій користувачів веб-сайту.

4. 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.
5. 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%.
6. 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 ОПМ 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, МН 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 ОПМ.
7. 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.
8. 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%.
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. 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 – 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.

СПИСОК ОСНОВНИХ ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Статті у періодичних виданнях, індексованих у Scopus та Web of Science

1. Lytvyn V., Pukach P., Vysotska V., Vovk M., Kholodna N. Identification and correction of grammatical errors in Ukrainian texts based on machine learning technology. Mathematics. 2023. Vol. 11. 904. ISSN 2227-7390. (квартиль Q2 відповідно до SCImago Journal).
2. Bisikalo O., Danylchuk O., Kovtun V., Kovtun O., Nikitenko O., Vysotska V. Modeling of operation of information system for critical use in the conditions of influence of a complex certain negative factor. International Journal of Control, Automation and Systems. 2022. Vol. 20. P. 904–1913. Print ISSN 1598-6446. (квартиль Q2 відповідно до SCImago Journal).
3. Bublyk M., Kowalska-Styczeń A., Lytvyn V., Vysotska V. The Ukrainian economy transformation into the circular based on fuzzy-logic cluster analysis. Energies. 2021. Vol. 14(18). Art. 5951. ISSN:1996-1073. (квартиль Q2 відповідно до SCImago Journal).
4. Lytvyn V., Vysotska V., Peleshchak I., Rishnyak I., Peleshchak R. Time dependence of the output signal morphology for nonlinear oscillator neuron based on Van der Pol model. International Journal of Intelligent Systems and Applications. 2018.Vol. 10(4). P. 8–17. ISSN: 2074-904X. (квартиль Q2 відповідно до SCImago Journal).
5. Висоцька В. Метод авторифікації тексту науково-технічних публікацій на основі лінгвістичного аналізу коефіцієнтів мовної різноманітності. Радіоелектроніка. Інформатика. Управління. 2020. № 1(52). С. 108–124.
6. Висоцька В. Інформаційна технологія просування інтернет-ресурсів в пошукових системах на основі контент-аналізу ключових слів web-сторінок. Радіоелектроніка, інформатика, управління. 2021 № 3 (58). С. 133-151.
7. Алєксєєва К. А., Берко А. Ю., Висоцька В. А. Технологія управління комерційним web-ресурсом на основі нечіткої логіки. Радіоелектроніка. Інформатика. Управління. 2015. № 3 (34). С. 71–79.
8. Бісікало О. В., Висоцька В. А. Виявлення ключових слів на основі методу контент-моніторингу україномовних текстів. Радіоелектроніка. Інформатика. Управління. 2016. № 1 (36). С. 74–83.
9. Бісікало О. В., Висоцька В. А. Застосування методу синтаксичного аналізу речень для визначення ключових слів україномовного тексту. Радіоелектроніка. Інформатика. Управління. 2016. № 3 (38). С. 54–65.
10. Lytvyn V., Pukach P., Bobyk I., Vysotska V. The method of formation of the status of personality understanding based on the content analysis. Eastern-European Journal of Enterprise Technologies. 2016. Vol. 5. P. 4–12.
11. Литвин В. В., Бобик І. О., Висоцька В. А. Застосування системи алгоритмічних алгебр для граматичного аналізу символьних обчислень виразів логіки висловлювань. Радіоелектроніка. Інформатика. Управління. 2016. № 4 (39). С. 77–89.
12. Lytvyn V., Vysotska V., Pukach P., Bobyk I., Pakholok B. A method for constructing recruitment rules based on the analysis of a specialist's competences. Eastern-European Journal of Enterprise Technologies. 2016. Vol. 6/2 (84). P. 4–14.
13. Lytvyn V., Vysotska V., Pukach P., Brodyak O., Ugryn D. Development of a method for determining the keywords in the Slavic language texts based on the technology of web mining. Eastern-European Journal of Enterprise Technologies. 2017. Vol. 2/2 (86). P. 14–23.
14. Lytvyn V., Vysotska V., Pukach P., Vovk M., Ugryn D. Method of functioning of intelligent

agents, designed to solve action planning problems based on ontological approach. Eastern-European Journal of Enterprise Technologies. 2017. Vol. 3/2 (87). P. 11–17.
15. Lytvyn V., Vysotska V., Pukach P., Bobyk I., Uhryn D. Development of a method for the recognition of author's style in the Ukrainian language texts based on linguometry, stylemetry. Eastern-European Journal of Enterprise Technologies. 2017. Vol. 4/2 (88). P. 10–18.
16. Коробчинський М. В., Чирун Л. Б., Висоцька В. А., Нич М. О. Особливості прогнозування результатів матчів у кіберспорті. Радіоелектроніка. Інформатика. Управління. 2017. № 3 (42). С. 95–105.
17. Коробчинський М. В., Чирун Л. Б., Висоцька В. А., Кондратьєв Є. О. Особливості формування та аналізу контенту інтернет-газети музичних новин. Радіоелектроніка. Інформатика. Управління. 2017. № 4. С. 139–150.
18. Lytvyn V., Vysotska V., Uhryn D., Hrendus M., Naum O. Analysis of statistical methods for stable combinations determination of keywords identification. Eastern-European Journal of Enterprise Technologies. 2018. Vol. 2/2 (92). P. 23–37.
19. Lytvyn V., Vysotska V., Maria H. Method of data expression from the Ukrainian content based on the ontological approach. Радіоелектроніка. Інформатика. Управління. 2018. № 3 (46). P. 144–157.
20. Lytvyn V., Vysotska V., Pukach P., Nytrebych Z., Demkiv I., Kovalchuk R., Huzyk N. Development of the linguometric method for automatic identification of the author of text content based on statistical analysis of language diversity coefficients. Eastern-European Journal of Enterprise Technologies. 2018. Vol. 5/2 (95). P. 16–28.
21. Lytvyn V., Vysotska V., Kuchkovskiy V., Pelekh I., Bobyk I., Malanchuk O., Ryshkovets Y., Brodyak O., Bobrivetc V., Panasyuk V. Development of the system to integrate and generate content considering the cryptocurrent needs of users. Eastern-European Journal of Enterprise Technologies. 2019. Vol. 1/2(97). P. 18–39.
22. Lytvyn V., Vysotska V., Pukach P., Nytrebych Z., Demkiv I., Senyk A., Malanchuk O., Sachenko S., Kovalchuk R., Huzyk N. Analysis of the developed quantitative method for automatic attribution of scientific and technical text content written in Ukrainian. Eastern-European Journal of Enterprise Technologies. 2018. Vol. 6/2 (96). P. 19–31.
23. Berko A., Vysotska V., Lytvyn V., Naum O. Planning the activities of intellectual agents in the electronic commerce systems. Радіоелектроніка. Інформатика. Управління. 2018. № 4. С. 143–158.
24. Lytvyn V., Vysotska V., Demchuk A., Demkiv I., Ukhans'ka O., Hladun V., Kovalchuk R., Petruchenko O., Dzyubyk L., Sokulska N. Design of the architecture of an intelligent system for distributing commercial content in the internet space based on SEO-technologies, neural networks, and machine learning. Eastern-European Journal of Enterprise Technologies. 2019. Vol. 2/2(98). P. 15–34.
25. Lytvyn V., Vysotska V., Shatskykh V., Kohut I., Petruchenko O., Dzyubyk L., Bobrivetc V., Panasyuk V., Sachenko S., Komar M. Design of a recommendation system based on collaborative filtering and machine learning considering personal needs of the user. Eastern-European Journal of Enterprise Technologies. 2019. Vol. 4/2 (100). P. 6–28.
26. Vysotska V., Demchuk A., Lytvyn V. Features of the architecture for Internet commercial content management system based on methods of Machine Learning, Web mining and SEO technologies. Радіоелектроніка. Інформатика. Управління. 2019. № 4. С. 121–135.
27. Lytvyn V., Vysotska V., Budz I., Pelekh Y., Sokulska N., Kovalchuk R., Dzyubyk L., Tereshchuk O., Komar M. Development of the quantitative method for automated text content authorship attribution based on the statistical analysis of N-grams distribution. Eastern-European Journal of Enterprise Technologies. 2019. Vol. 6/2 (102). P. 28–51.
28. Кравець П., Литвин В., Висоцька В. Ігрова модель онтологічної підтримки проектів.

Радіоелектроніка, інформатика, управління. 2021. № 1(56). С. 172–183.
29. Литвин В. В., Бублик М. І., Висоцька В. А., Мацелюх Ю. Р. Технологія візуальної симуляції пасажиропотоків у сфері громадського транспорту smart city. Радіоелектроніка, інформатика, управління. 2021 № 4 (59). С. 106-121.
30. Кравець П. О., Литвин В. В., Висоцька В. А. Моделювання ігрової задачі призначення персоналу для виконання ІТ-проектів на основі онтологій. Радіоелектроніка, інформатика, управління. 2022. № 1 (60). С. 130–145.
31. Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. Classification methods of text documents using ontology based approach. Advances in Intelligent Systems and Computing. 2017. Vol. 512. P. 229–240.
32. Lytvyn V., Vysotska V., Burov Y., Veres O., Rishnyak I. The contextual search method based on domain thesaurus. Advances in Intelligent Systems and Computing. 2018. Vol. 689. P. 310–319.
33. Kanishcheva O., Vysotska V., Chyrun L., Gozhyj A. Method of integration and content management of the information resources network. Advances in Intelligent Systems and Computing. 2018. Vol. 689. P. 204–216.
34. Vysotska V., Fernandes B. V., Emmerich M. Web content support method in electronic business systems. CEUR Workshop Proceedings. 2018. Vol. 2136. P. 20–41.
35. Lytvyn V., Vysotska V., Dosyn D., Burov Y. Method for ontology content and structure optimization, provided by a weighted conceptual graph. Webology. 2018. Vol. 15(2). P. 66–85.
36. Lytvyn V., Vysotska V., Osypov M., Slyusarchuk O., Slyusarchuk Y. Development of intellectual system for data de-duplication and distribution in cloud storage. Webology. 2019. Vol. 16. P. 1-42.
37. Vysotska V., Lytvyn V., Burov Y., Gozhyj A. Makara, S. The consolidated information web-resource about pharmacy networks in city. CEUR Workshop Proceedings. 2018. Vol. 2255. P. 239–255.
38. Rusyn B., Lytvyn V., Vysotska V., Emmerich M., Pohreliuk L. The virtual library system design and development. Advances in Intelligent Systems and Computing. 2019. Vol. 871. P. 328–349.
39. Vysotska V., Fernandes B. V., Lytvyn V., Emmerich M., Hirnyak M. Method for determining linguometric coefficient dynamics of Ukrainian text content authorship. Advances in Intelligent Systems and Computing. 2019. Vol. 871. P. 132–151.
40. Gozhyj A., Vysotska V., Yevseyeva I., Kalinina I., Gozhyj V. Web resources management method based on intelligent technologies. Advances in Intelligent Systems and Computing (AISC). 2019. Vol. 871. P. 206–221.
41. Vysotska V., Lytvyn V., Burov Y., Berezin P., Emmerich M., Fernandes B. V. Development of information system for textual content categorizing based on ontology. CEUR Workshop Proceedings. 2019. Vol. 2362. P. 53–70.
42. Burov Y., Vysotska V., Kravets P. Ontological approach to plot analysis and modeling. CEUR Workshop Proceedings. 2019. Vol. 2362. P. 22–31.
43. Lytvyn V., Vysotska V., Rusyn B., Pohreliuk L., Berezin P., Naum O. Textual content categorizing technology development based on ontology. CEUR Workshop Proceedings. 2019. Vol. 2386. P. 234–254.
44. Lytvyn V., Vysotska V., Rzheuskyi A. Technology for the psychological portraits formation of social networks users for the IT specialists recruitment based on Big Five, NLP and Big Data. CEUR Workshop Proceedings. 2019. M2392. P. 147–171.
45. Vysotska V., Burov Y., Lytvyn V., Oleshek O. Automated monitoring of changes in web resources. Advances in Intelligent Systems and Computing. 2020. Vol. 1020. P. 348–363.
46. Demchuk A., Lytvyn V., Vysotska V., Dilai M. Methods and means of web content personalization for commercial information products distribution. Advances in Intelligent Systems and Computing. 2020. Vol. 1020. P. 332–347.
47. Kravets P., Burov Y., Lytvyn V., Vysotska V. Gaming method of ontology clusterization. Webology. 2019. Vol. 16(1). P. 55–76.

48. Lytvyn V., Burov Y., Kravets P., Vysotska V., Demchuk A., Berko A., Ryshkovets Y., Shcherbak S., Naum O. Methods and models of intellectual processing of texts for building ontologies of software for medical terms identification in content classification. CEUR Workshop Proceedings. 2019. Vol. 2488. P. 354–368.
49. Lytvyn V., Gozhyj A., Kalinina I., Vysotska V., Shatskykh V., Chyrun L., Borzov Y. An intelligent system of the content relevance at the example of films according to user needs. CEUR Workshop Proceedings. 2019. Vol. 2516. P. 1–23.
50. Peleshko D., Rak T., Noennig J. R., Lytvyn V., Vysotska V. Drone monitoring system DROMOS of urban environmental dynamics. CEUR Workshop Proceedings. 2020. Vol. 2565. P. 178–193.
51. Krislata I., Katrenko A., Lytvyn V., Vysotska V., Burov Y. Traffic flows system development for smart city. CEUR Workshop Proceedings. 2020. Vol. 2565. P. 280–294.
52. Bisikalo O., Vysotska V. Linguistic analysis method of Ukrainian commercial textual content. CEUR Workshop Proceedings. 2020. Vol. 2608. P. 224–244.
53. Bisikalo O., Vysotska V., Burov Y., Kravets P. Conceptual model of process formation for the semantics of sentence in natural language. CEUR Workshop Proceedings. 2020. Vol. 2604. P. 151–177.
54. Vysotska V. Ukrainian participles formation by the generative grammars use. CEUR Workshop Proceedings. 2020. Vol. 2604. P. 407–427.
55. Batiuk T., Vysotska V., Lytvyn V. Intelligent system for socialization by personal interests on the basis of SEO technologies and methods of machine learning. CEUR Workshop Proceedings. 2020. Vol. 2604. P. 1237–1250.
56. Oliinyk V., Vysotska V., Burov Y., Mykich K., Fernandes V. B. Propaganda detection in text data based on NLP and machine learning. CEUR Workshop Proceedings. 2020. Vol. 2631. P. 132–144.
57. Kalinina I., Vysotska V., Sachenko S., Kovalchuk R., Gozhyj A. Qualitative and quantitative characteristics analysis for information security risk assessment in e-commerce systems. CEUR Workshop Proceedings. 2020. Vol. 2762. P. 177–190.
58. Lytvyn V., Hryhorovych A., Bublyk M., Hryhorovych V., Vysotska V., Chyrun L. Medical content processing in intelligent system of district therapist. CEUR Workshop Proceedings. 2020. Vol. 2753. P. 415–429.
59. Bublyk M., Lytvyn V., Vysotska V., Sokulska N., Chyrun L., Matseliukh Y. The decision tree usage for the results analysis of the psychophysiological testing. CEUR Workshop Proceedings.2020.Vol.2753.P.458–472.
60. Kravets P., Lytvyn V., Vysotska V., Burov Y., Andrusyak I. Game task of ontological project coverage. CEUR Workshop Proceedings. 2021. Vol. 2851. P. 344–355.
61. Kravets P., Burov Y., Oborska O., Vysotska V., Dzyubyk L., Lytvyn V. Stochastic Game Model of Data Clustering. CEUR Workshop Proceedings. 2021. Vol. 2853. P. 198–213.
62. Bublyk M., Vysotska V., Panasyuk V., Brodyak O., Chyrun L. Assessing security risks method in e-commerce system for IT portfolio management. CEUR Workshop Proceedings. 2021. Vol. 2853. P. 462–479.
63. Vysotska V., Lytvyn V., Danylyk V., Vyshemyrska S., Lurie I., Luchkevych M. Detecting items with the biggest weight based on neural network and machine learning methods. Communications in Computer and Information Science.2020.V.1158.P.383-396.
64. Tymoshenko K., Vysotska V., Kovtun O., Holoshchuk R., Holoshchuk S. Real-time Ukrainian text recognition and voicing. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 357–387.
65. Vysotska V., Holoshchuk S., Holoshchuk R. A comparative analysis for English and Ukrainian texts processing based on semantics and syntax approach. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 311–356.
66. Balush I., Vysotska V., Albota S. Recommendation System Development Based on Intelligent Search, NLP and Machine Learning Methods. CEUR Workshop Proceedings. 2021. Vol. 2917. P. 584-617.

67. Kholodna N., Vysotska V., Albota S. A Machine Learning Model for Automatic Emotion Detection from Speech. CEUR Workshop Proceedings. 2021. Vol. 2917. P. 699-713.
68. Kravets P., Lytvyn V., Dobrotvor I., Sachenko O., Vysotska V., Sachenko A. Matrix Stochastic Game with Q-learning for Multi-agent Systems. Lecture Notes on Data Engineering and Communications Technologies. 2021. Vol. 83. P. 304–314.
69. Kravets P., Burov Y., Lytvyn V., Vysotska V., Ryshkovets Y., Brodyak O., Vyshemyrska S. Markovian Learning Methods in Decision-Making Systems. Lecture Notes on Data Engineering and Communications Technologies. 2022. Vol. 77. P. 423-437.
70. Vysotska V., Berko A., Lytvyn V., Kravets P., Dzyubyk L., Bardachov Y., Vyshemyrska S. Information resource management technology based on fuzzy logic. Advances in Intelligent Systems and Computing (AISC). 2020. Vol. 1246. P. 164–182.

Монографії

71. Lytvyn V., Vysotska V., Chyrun L., Dosyn D. Methods based on ontologies for information resources processing. Saarbrücken: LAP Lambert Academic Publishing, 2016. 324 p.
72. Литвин В. В., Висоцька В. А., Досин Д. Г. Методи та засоби опрацювання інформаційних ресурсів на основі онтологій. Львів: Піраміда, 2016. 404 с.
73. Висоцька В. А. Технології електронної комерції та Інтернет-маркетингу. Saarbrücken: LAP Lambert Academic Publishing, 2018. 285 с.
74. Vysotska V., Lytvyn V. Web resources processing based on ontologies. Saarbrücken: LAP Lambert Academic Publishing, 2018. 232 p.
75. Vysotska V Internet systems design and development based on Web Mining and NLP. Saarbrücken: LAP Lambert Academic Publishing, 2018. 316 p.
76. Vysotska V. Computer linguistics for online marketing in information technology. Saarbrücken: LAP Lambert Academic Publishing, 2018. 396 p.
77. Висоцька В. А., Досин Д. Г., Микіч Х. І., Завущак І. І., Рибчак З. Л. Методи та засоби функціонування систем підтримки прийняття рішень на основі онтологій. Львів: Новий світ, 2019. 334 с.

Статті у матеріалах конференцій, індексованих у Scopus та Web of Science

78. Vysotska V. Chyrun L. Methods of information resources processing in electronic content commerce systemsю CADSM: матеріали XIII Міжнар. наук.-техн. конф., 24–27 лют., Львів, Поляна, 2015. С. 328–332.
79. Vysotska V., Chyrun L. Analysis features of information resources processing. CSIT : proc. of the Xth Intern. conf., 14–17 Sept., Lviv, Ukraine, 2015. P. 124–128.
80. Vysotska V. Linguistic analysis of textual commercial content for information resources processing. TCSET: proc. of the XIII Intern. conf, Feb. 23–26, Lviv, Slavske, Ukraine, 2016. P. 709–713.
81. Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. Content linguistic analysis methods for textual documents classification. CSIT: proc. of the XIth Intern. conf., 6–10 Sept., Lviv. P. 190–192.
82. Lytvyn V., Vysotska V., Chyrun L., Chyrun L. Distance learning method for modern youth promotion and involvement in independent scientific researches. The 1st IEEE International conference on data stream mining and processing, DSMP: proc. Aug. 23–27, Lviv, Ukraine. 2016. P. 269–274.
83. Lytvyn V., Vysotska V., Dosyn D., Holoschuk R., Rybchak Z. Application of sentence parsing for determining keywords in Ukrainian texts. CSIT : proc. of the XIIth Intern. conf., 5–8 Sept., Lviv, Ukraine. 2017. P. 326–331.
84. Su J., Vysotska V., Sachenko A., Lytvyn V., Burov Y. Information resources processing using linguistic analysis of textual content. IDAACS : proc. conf., Bucharest, Sept. 21–23. 2017. P. 573–578.
85. Lytvyn V., Vysotska V., Burov Y., Demchuk A. Defining author's style for plagiarism detection in academic environment. DSMP : proc. of Intern. conf., Aug. 21–25, Lviv, Ukraine. 2018. P. 128–133.

86. Lytvyn V., Vysotska V., Burov Y., Bobyk I., Ohirko O. The linguometric approach for co-authoring author's style definition. IDAACS-SWS : proc., Lviv, 20–21 September 2018. P. 29–34.
87. Vysotska V., Lytvyn V., Hrendus M., Brodyak O., Kubinska S. Method of textual information authorship analysis based on stylometry. CSIT: proc. of Intern. conf., 11–14 вер., Львів. 2018. С. 9–16.
88. Vysotska V., Kanishcheva O., Hlavcheva Y. Authorship identification of the scientific text in Ukrainian with using the lingvometry methods. CSIT : proc. of the Intern. conf., 11–14 вересня 2018 р., Львів. 2018. С. 34–38.
89. Gozhyj A., Kalinina I., Vysotska V., Gozhyj V. The method of web-resources management under conditions of uncertainty based on fuzzy logic. CSIT : proc. of the Intern. conf., Львів,11–14 вер. 2018. Т. 1. С. 343–346.
90. Lytvyn V., Peleshchak I., Peleshchak R., Vysotska V. Information encryption based on the synthesis of a neural network and AES algorithm. AICT : proc. of the 3rd Intern. conf., Lviv, Ukraine, July 2–6 2019. P. 447–450.
91. Lytvyn V., Vysotska V., Peleshchak I., Basyuk T., Kovalchuk V., Kubinska S., Chyrun L., Rusyn B., Pohreliuk L., Salo T. Identifying textual content based on thematic analysis of similar texts in big data. CSIT: proc. of Intern. conf., Львів, 17–20 вер. 2019. Т. 2. С. 84–91.
92. Shu C., Dosyn D., Lytvyn V., Vysotska V., Sachenko A., Jun S. Building of the predicate recognition system for the NLP ontology learning module. The 10th IDAACS : proc., September 18–21, Metz, France. 2019. P. 802–808.
93. Vysotska V., Berko A., Bublyk M., Chyrun L., Vysotsky A., Doroshkevych K. Methods and tools for web resources processing in e-commercial content systems. CSIT : proc. of the Intern. conf., Збараж, 23–26 вересня, 2020. P. 114–118.
94. Lytvyn V., Burov Y., Vysotska V., Pukach Y., Tereshchuk O., Shakleina I. Abstracting Text Content Based on Weighing the TF-IDF Measure by the Subject Area Ontology. SIST: proc. of the Intern. conf., 28-30 April 2021, Nur-Sultan, Kazakhstan. Art. 9465978.
95. Dmytriv A., Vysotska V., Bublyk M. The Speech Parts Identification for Ukrainian Words Based on VESUM and Horokh Using. CSIT : proc. of 16th Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. Vol. 2. P. 21–33.
96. Kubinska S., Vysotska V., Matseliukh Y. User Mood Recognition and Further Dialog Support. CSIT: proc. of Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. P. 34–39.
97. Ivanchyshyn D., Vysotska V., Albota S. The Film Script Generation Analysis Based on the Fiction Book Text Using Machine Learning. CSIT : proc. of the 16th Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. Vol. 2. P. 68–80.
98. Aksonov D., Gozhyj A., Kalinina I., Vysotska V. Question-Answering Systems Development Based on Big Data Analysis. CSIT: proc. of Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. P. 113–118.
99. Mykytiuk A., Vysotska V., Albota S. Spam Filtration System with the Use of Machine Learning Technology. CSIT: proc. of Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. Vol. 1. P. 124–130.
100. Zanchak M., Vysotska V., Albota S. The Sarcasm Detection in News Headlines Based on Machine Learning Technology. CSIT: proc. of Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. P. 131–137.

АНОТАЦІЇ

Висоцька В.А. Аналіз та синтез комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту. – На правах рукопису.

Дисертаційна робота на здобуття наукового ступеня доктора технічних наук за спеціальністю 10.02.21 – структурна, прикладна і математична лінгвістика. – Національний університет «Львівська політехніка», Міністерство освіти і науки України, Львів, 2023.

У дисертації вирішено важливу науково-прикладну проблему аналізу та синтезу комп'ютерних лінгвістичних систем (КЛС) для розв'язання різних задач

опрацювання україномовного текстового контенту на основі розроблення нових та удосконаленні відомих моделей, методів та засобів опрацювання природної мови (ОПМ).

Аналіз та синтез КЛС базується на застосуванні лінгвістичного аналізу україномовного текстового контенту, інтелектуальному опрацювання текстового потоку контенту, машинному навчанні системи на достовірних даних та статистичному аналізі для знаходження закономірностей появи лінгвістичних подій. Розроблена інформаційна технологія (ІТ) опрацювання україномовного текстового контенту на відміну від існуючих підтримує принцип модульності типової архітектури КЛС для розв'язку конкретної задачі ОПМ 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. Наведено приклади розроблених КЛС опрацювання україномовного текстового контенту для розв'язку відповідних задач ОПМ, функціонування яких ґрунтується на розроблених та вдосконалених моделях, методах та алгоритмах.

Удосконалена модель лінгвістичного опрацювання текстового контенту на основі графемного, морфологічного, лексичного, синтаксичного, семантичного, структурного, онтологічного та прагматичного аналізу для вирішення конкретної проблеми ОПМ. Це дало змогу сформулювати загальні вимоги до процесів опрацювання україномовного контенту. Удосконалення методів опрацювання інформаційних ресурсів, таких як інтеграція, управління та супровід україномовного контенту, дозволило адаптувати процес інтелектуального аналізу текстового потоку до розв'язку різних задач ОПМ та розробити КЛС, що ефективно функціонують, метрики для розв'язку різних задач ОПМ. Удосконалені методи ОПМ на основі регулярних виразів узгодження за шаблоном дозволили адаптувати алгоритми графемного та морфологічного аналізу для опрацювання україномовних текстів.

Удосконалено метод токенізації та нормалізації тексту каскадами простих підстановок регулярних виразів і кінцевих автоматів, що дало змогу адаптувати алгоритм лексичного та синтаксичного аналізів для опрацювання україномовних текстів. Удосконалено метод морфологічного аналізу, заснований на сегментації та нормалізації слів, сегментації речень і модифікованому алгоритмі стемінгу Портера як ефективного засобу ідентифікації афіксів лем для можливості розмічування аналізованого слова, що дозволило підвищити точність пошуку ключових слів на 9%.

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 на 6-9%, здійснити пошук стійких словосполучень та рубрикацію контенту. 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 6-7%.

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 на 6-12%. 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 до [9;34]% із загальної кількості учасників проекту.

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 ОПМ шляхом реалізації типового програмного забезпечення таких систем.

КЛС реалізовано на інформаційному ресурсі http://victana.lviv.ua засобами CMS Joomla! (для розроблення е-каркасу сайту), 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 – 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.

Ключові слова: NLP, комп'ютерна лінгвістика, текстовий контент, українська мова, графемний аналіз, морфологічний аналіз, лексичний аналіз, синтаксичний аналіз, семантичний аналіз, структурний аналіз, прагматичний аналіз, інформаційна технологія, машинне навчання, опрацювання природної мови, інформаційна система, онтологія, ключові слова, стійкі словосполучення, стиль автора, ідентифікація автора, психологічний аналіз тексту.

Vysotska V.A. Analysis and synthesis of computational linguistic systems for processing Ukrainian textual content. – Manuscript.

Thesis for a Doctoral degree in Technical Science, speciality 10.02.21 – structural, applied and mathematical linguistics. – Lviv Polytechnic National University, Ministry of Education and Science of Ukraine, Lviv, 2023.

The dissertation solves an important scientific and applied problem of analysis and synthesis of computer linguistic systems (CLS) for solving various problems of processing Ukrainian-language text content. It is based on the development and improvement of new and existing models, methods and tools for natural language processing (NLP).

The analysis and synthesis of CLS is based on the application of linguistic analysis of Ukrainian-language textual content, intelligent processing of textual flow of content, machine learning of the system based on reliable data, and statistical analysis to find patterns in the appearance of linguistic events. Developed information technology (IT) for processing of Ukrainian-language textual content, unlike the existing ones, supports the

modularity principle of the typical architecture of the CLS for solving a specific task of the NLP and analysing a set of parameters and metrics of effectiveness of the system in accordance with the behaviour of the target audience. The general structure of the CLS for the processing of text content in the Ukrainian language and the conceptual scheme/model of functioning of a typical CLS based on the modelling of the interaction of the main processes and components of the system were developed, which made possible to improve IT intellectual analysis of the text flow based on the processing of information resources. There are examples of developed CLS for processing Ukrainian-language textual content for solving relevant tasks of the NLP, functioning of which is based on developed and improved models, methods and algorithms.

An improved model of linguistic processing of textual content based on graphemic, morphological, lexical, syntactic, semantic, structural, ontological and pragmatic analysis to solve a specific problem of NLP is introduced. It has enabled the formulation of general requirements for Ukrainian content processing. Process improvement methodologies for information resources such as integration, management and content support of the Ukrainian language allow to adapt the intellectual analysis of the text stream processing to the solution of various tasks of NLP and develop effective CLS and metrics to solve various NLP problems. NLP methods based on regular pattern-matching expressions are improved and it has allowed the adaptation of grapheme and morphological analysis algorithms to Ukrainian text processing.

A method of tokenisation and normalisation of text by cascades of simple substitutions of regular expressions and finite state machines is upgraded and resulted in the adaptation of the lexical and syntactic analysis algorithm for Ukrainian text processing. The morphological analysis method based on word segmentation and normalisation, sentence segmentation, and a modified Porter stemming algorithm as an effective tool for identifying lemmas affixes to tag the analysed word is improved. It has resulted in a 9% increase in keyword search accuracy.

A method of identifying keywords in Ukrainian texts based on grapheme and morphological analysis of the word base using regular expressions and N-grams is elaborated. It has increased the accuracy of keyword searches by 6-9%, stable word combinations and categorise content search. A method for determining stable word combinations based on the identification of keywords in a Ukrainian text and the lexical coefficients analysis of the text author in the reference text is developed. The accuracy of the method for determining the author's style, based on statistical linguistics, has been improved by 6-7%.

A method for determining the author's style of thematic Ukrainian textual content based on the analysis of keywords, stable phrases, N-grams, linguometry and stylometry is developed. It has enabled the recognition of the stylistic contribution of each author and increased the accuracy of scientific and technical publications attribution by 6-12%. A method is developed to verify the authorship level of a Ukrainian text from the number of possible authors, based on a stylistic comparison analysis of the potential authors. It has improved the classification accuracy of style similarity to [9;34]% of the total number of project participants.

The analysis and synthesis methods of CLS are developed based on the creation of an organisational structure of the Ukrainian text processing system through the support of

modularity, and modelling the main processes and components interaction. It has improved the number of solutions to various typical NLP problems by implementing typical software systems.

CLS is realised on the platform http://victana.lviv.ua using CMS Joomla! (developing the site e-framework), PHP (implementation of text content processing methods), HTML (page mark-up), CSS (description of page styles), MySQL (storing data and dictionaries). An experimental study confirms the reliability of the method used to identify keywords and proves that for different source text processing algorithms, the average agreement of the identified keywords lists with the author's ones varies between 52.6 and 68.5%. The accuracy of keyword matching with the author's keywords ranges from 43.6% to 62.9%. The average match of meaningful keywords compared to all keywords found by the system varies between 38.9 and 75.8%, depending on the analysis stage of the article texts. The accuracy of matching keywords compared to all those found by the system varies between 34.3 and 71.9%, depending on the stage of analysis of the article texts.

Keywords: NLP, computational linguistics, text content, Ukrainian language, grapheme analysis, morphological analysis, lexical analysis, syntactic analysis, semantic analysis, structural analysis, pragmatic analysis, information technology, machine learning, natural language processing, information system, ontology, keywords, stable word phrases, author's style, author identification, psychological analysis of the text.