Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)
Кафедра комп'ютерних наук
(повна назва кафедри)

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня

бакалавр
(назва освітнього ступеня)

на тему: Дослідження засобів текстової аналітики для опрацювання даних про COVID-19

Виконав: студент IV курсу, групи СНс-42
спеціальності 122 Комп'ютерні науки
(шифр і назва спеціальності)

_____ Довгунь Д.О.
(підпис) (прізвище та ініціали)

Керівник _____ Пасічник В.В.
(підпис) (прізвище та ініціали)

Нормоконтроль _____ Шимчук Г.В.
(підпис) (прізвище та ініціали)

Завідувач кафедри _____ Боднарчук І.О.
(підпис) (прізвище та ініціали)

Рецензент _____ Гащин Н.Б.
(підпис) (прізвище та ініціали)

Тернопіль
2021

Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)

Кафедра комп'ютерних наук
(повна назва кафедри)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Боднарчук І.О.
(підпис) (прізвище та ініціали)
«_22_» _червня_ 2021 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

на здобуття освітнього ступеня Бакалавр
(назва освітнього ступеня)

за спеціальністю 122 Комп'ютерні науки
(шифр і назва спеціальності)

Студенту Довгунь Дмитро Олегович
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження засобів текстової аналітики для опрацювання даних про COVID-19

Керівник роботи Прізвище Ім'я По батькові, науковий ступінь, посада кафедри КН
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджені наказом ректора від «02» березня 2021 року № 4/7-171

2. Термін подання студентом завершеної роботи 23 червня 2021р.

3. Вихідні дані до роботи Наукові публікації про засоби аналітичного опрацювання текстових даних про COVID-19

4. Зміст роботи (перелік питань, які потрібно розробити)
Вступ. 1. Аналіз текстів у галузі COVID-19. Стан досліджень, корпуси та ресурси. 2. Засоби текстової аналітики для опрацювання даних про COVID-19. 3. Безпека життєдіяльності, основи хорони праці. Висновки. Перелік джерел. Додатки.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень, слайдів)
1. Титульний аркуш. 2. Тема, мета, завдання дослідження. 3. Актуальність дослідження. 4. Робочий процес для створення системи видобування тексту з літератури щодо COVID-19. 5. Корпус текстів. 6. Ресурси по вбудовуванню для дослідників та фахівців з майнінгу тексту в предметній області COVID-19. 7. Ресурси щодо анотацій, котрі призначені для дослідників та фахівців з майнінгу тексту COVID-19. 8. Ресурси мовних моделей для дослідників та фахівців з майнінгу тексту COVID-19. 9. Системи опрацювання текстів щодо COVID-19. 10. Системи опрацювання текстів щодо COVID-19 для вирішення пошукових задач. 11. Системи контролю якості текстів щодо COVID-19. 12. Процес формування системного огляду наукових публікацій щодо COVID-19. 13. Висновки. 14. Доповідь завершено.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Безпека життєдіяльності, основи хорони праці	Гурик О.Я., доцент кафедри МТ		

7. Дата видачі завдання 25 січня 2021 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1.	Ознайомлення з завданням до кваліфікаційної роботи	25.01.2021	*Виконано*
2.	Підбір джерел про засоби текстової аналітики для	26.01.2021-01.02.2021	*Виконано*
	опрацювання даних про COVID-19.		
3.	Переклад та опрацювання джерел про засоби	02.02.2021-08.02.2021	*Виконано*
	текстової аналітики для опрацювання даних про		
	COVID-19.		
4.	Виконання дослідження щодо аналізу засобів	09.02.2021-13.02.2021	*Виконано*
	текстової аналітики для опрацювання даних про		
	COVID-19.		
5.	Оформлення розділу «Аналіз текстів у галузі	14.02.2021-19.02.2021	*Виконано*
	COVID-19. Стан досліджень, корпуси та ресурси»		
6.	Оформлення розділу «Засоби текстової аналітики для	20.02.2021-24.02.2021	*Виконано*
	опрацювання даних про COVID-19»		
7.	Виконання завдання до підрозділу «Безпека	07.06.2021-08.06.2021	*Виконано*
	життєдіяльності»		
8.	Виконання завдання до підрозділу «Основи хорони	07.06.2021-08.06.2021	*Виконано*
	праці»		
9.	Оформлення кваліфікаційної роботи	07.06.2021-08.06.2021	*Виконано*
10.	Нормоконтроль	07.06.2021-08.06.2021	*Виконано*
11.	Перевірка на плагіат	08.06.2021	*Виконано*
12.	Попередній захист кваліфікаційної роботи	09.06.2021	*Виконано*
13.	Захист кваліфікаційної роботи	23.06.2021	

Студент _____ Довгунь Д.О.
(підпис) (прізвище та ініціали)

Керівник роботи _____ Пасічник В.В.
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Дослідження засобів текстової аналітики для опрацювання даних про COVID-19 // Кваліфікаційна робота освітнього рівня «Бакалавр» // Довгунь Дмитро Олегович // Тернопільський національний технічний університет імені Івана Пулюя, факультет комп'ютерно-інформаційних систем і програмної інженерії, кафедра комп'ютерних наук, група СНс-42 // Тернопіль, 2021 // С.51, рис.– 2, табл.– 7, кресл.– 14, додат. – 1, бібліогр. – 93.

Ключові слова: COVID-19, видобування тексту, відношення, знання, опрацювання природної мови, пошук інформації; узагальнення.

Кваліфікаційна робота присвячена дослідженню методів та засобів текстової аналітики для опрацювання даних про COVID-19. Метою роботи є підвищення рівня поінформованості наукової спільноти та дослідників щодо опіблікованих в предметній області COVID-19 статей.

В першому розділі кваліфікаційної роботи освітнього рівня «Бакалавр» подано розлогий аналіз предметної області опрацювання текстів в галузі COVID-19. Описано основи корпусів для видобування тексту щодо COVID-19. Виявлено та проаналізовано ресурси для моделювання текстового майнінгу для COVID-19.

В другому розділі кваліфікаційної роботи освітнього рівня «Бакалавр» описано системи опрацювання текстів щодо COVID-19. Досліджено пошукові текстові системи для публікацій щодо COVID-19. Проаналізовано системи опрацювання текстів щодо COVID-19 з функціями розвідки. Описано системи опрацювання текстів щодо COVID-19 спрямовані на дослідження та контроль якості джерел. Розглянуто системи узагальнення текстів щодо COVID-19. Досліджено системні огляди наукових джерел щодо COVID-19.

ANNOTATION

Study of text analytics capabilities for COVID-19 data processing // Qualification work of educational level "Bachelor" // Dovhun Dmytro Olehovych // Ternopil National Technical University named after Ivan Pulyuy, Faculty of Computer Information Systems and Software Engineering, Department of Computer Science Sciences, group CHc-42 // Ternopil, 2021 // P.51, fig. – 2, tables – 7, chair. – 14, annexes - 1, ref. – 93.

Key words: COVID-19, text extraction, attitudes, knowledge, natural language processing, information retrieval; generalization.

Qualification work is devoted to the study of methods and tools of text analytics for data processing on COVID-19. The aim of the work is to raise the level of awareness of the scientific community and researchers about the articles published in the subject area of COVID-19.

The first section of the qualification work of the educational level "Bachelor" presents an extensive analysis of the subject area of word processing in the field of COVID-19. The basics of text extraction enclosures for COVID-19 are described. Resources for text mining modeling for COVID-19 have been identified and analyzed.

The second section of the qualification work of the educational level "Bachelor" describes the word processing systems for COVID-19. Text search engines for publications on COVID-19 have been studied. COVID-19 word processing systems with intelligence functions are analyzed. The word processing systems for COVID-19 aimed at research and quality control of sources are described. Systems of generalization of texts concerning COVID-19 are considered. Systematic reviews of scientific sources on COVID-19 have been studied.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

CDC (англ. Centers for Disease Control and Prevention) – Центри з контролю та профілактики захворювань (США).

JSON (англ. JavaScript Object Notation) – запис об’єктів JavaScript – текстовий формат обміну даними між комп’ютерами.

NIH (англ. National Institutes of Health) – Національні інститути здоров’я (США).

QA (англ. Quality Assurance) – контроль якості.

PMC (англ. PubMed Central) – база даних COVID-19 ВООЗ

ВООЗ – Всесвітня організація охорони здоров’я.

БД – база даних.

БК – база ключів.

Корпус текстів – це вид корпусу даних, одиницями якого є тексти або їх достатньо значні фрагменти, що включають, наприклад, якісь повні фрагменти макроструктури текстів даної проблемної області [1].

ШІ – Штучний інтелект.

ЗМІСТ

ВСТУП

Актуальність теми. На даний час відбуваєть публікація значної кількості наукових публікацій присвячених COVID-19. Аналіз яких спричиняє значне навантаження на дослідників та зацікавлених в опрацюванні нових публікацій осіб. З часу відкриття нового коронавірусу SARS-CoV-2 [2] опубліковано понад сто тисяч статей та препринтів щодо викликаного COVID-19 захворювання. З початку 2021 року про COVID-19 було опубліковано понад п’ятдесять тисяч статей. На даний час щодня продовжує публікуватися декілька сотень нових статей за цею тематикою. COVID-19 прокотився по всьому світу і кардинально змінив усі аспекти нашого життя. Урядові установи, наукові та промислові дослідники об’єдналися навколо спільних цілей управління ресурсами в галузі охорони здоров’я, визначення соціальної політики, профілактики та лікування COVID-19, розробки вакцин, тощо. Наукове співтовариство швидко відреагувало на пандемію, а наукові публікації щодо COVID-19 з’являються з безпрецедентною швидкістю. Навіть відповідно до занижених оцінок, звичайні методи опрацювання публікацій шляхом перечитування виявляються не ефективними. Тому завдання формування автоматизованих підходів до аналітичного опрацювання тексту для ефективного опрацювання зростаючої хвилі результатів досліджень щодо COVID-19 є актуальним напрямком сучасних досліджень.

Мета і задачі дослідження. Метою даної кваліфікаційної роботи освітнього рівня «Бакалавр» є підвищення рівня поінформованості наукової спільноти та дослідників щодо опублікованих в предметній області COVID-19 статей. Для досягнення поставленої мети потребують вирішення ряд наступних завдань:

– Проаналізувати стан досліджень в галузі видобування тексту для предметної області COVID-19.

– Дослідити стан досліджень щодо корпусів видобування тексту щодо COVID-19.

– Описати ресурси для моделювання текстового майнінгу для COVID-19.

– Провести аналіз систем опрацювання текстів щодо COVID-19.

Практичне значення одержаних результатів, полягає в тому, що проаналізовано системні огляди наукових джерел щодо COVID-19 та виконано узагальнення результатів проведених наукових розвідок щодо COVID-19.

1 АНАЛІЗ ТЕКСТІВ У ГАЛУЗІ COVID-19. СТАН ДОСЛІДЖЕНЬ, КОРПУСИ ТА РЕСУРСИ

1.1 Аналіз предметної області

Однією з основних областей застосування біомедичного видобування тексту є управління перевантаженнями інформації [3]. Відповідно до [4], видобування тексту зосереджується на вирішенні конкретних проблем, таких як отримання відповідних документів або вилучення блоків інформації з цих документів. У процесі майнінгу системи видобування тексту можуть використовувати методи пошуку інформації, вилучення інформації, класифікації тексту тощо та використовувати методи із суміжних областей, зокрема, обробки природної мови, побудова бази знань тощо. Незважаючи на відсутність усталеного консенсусу щодо точних взаємозв'язків між цими різними за своєю природою завданнями або галузями дослідження доцільно зосередитись на підходах до опрацювання інформації [5]. При цьому слід застосовувати «видобування тексту» як загальний термін для позначення методів із вищеперелічених предметних областей.

У відповідь на великий обсяг літератури, опублікованої щодо COVID-19, дослідницька спільнота представила сформувала корпуси для видобування тексту, моделюючи при цьому ресурси, системи та загальні спільні завдання, характерні для предметної області COVID-19, для вирішення зростаючої проблеми. Корпуси – це колекції документів, попередньо опрацьовані для виділення машиночитаного тексту, які використовуються для видобування тексту. У даному випадку доцільно зосередитись на наукових публікаціях, що містять корпуси. Ресурси моделювання можуть включатись спеціалістами з розробки текстів у виробничі системи і складатися з засобів для вбудовування тексту, анотації даних, попередньо навчених мовних моделей, графіків знань тощо.

Лінгвістичні системи – це програми, що включають моделі для аналізу тексту та користувацькі інтерфейси для забезпечення функціональних можливостей пошуку, виявлення або візуалізації вмісту статтей. Спільні завдання – це змагання наукової громади, що сприяють концентрованій роботі над конкретними науковими завданнями.

На рисунку 1.1 подано ілюстрацію підходу фахівців із опрацювання текстів до розробки системи для вирішення задач, пов'язаних із перевантаженням інформації для дослідників в пов'язаних з COVID-19 галузях знань.

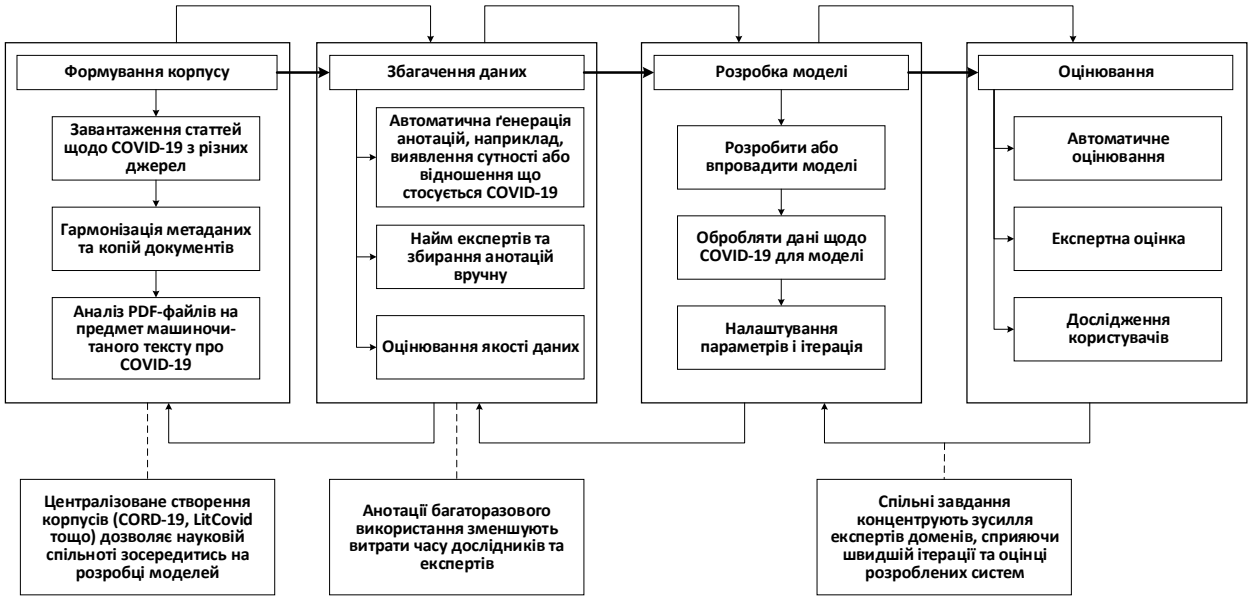

Рисунок 1.1 – Робочий процес для створення системи видобування тексту з літератури щодо COVID-19

При цьому типовий робочий процес складається з формування корпусу, збагачення даних, розробки моделі та оцінювання. На жаль, процес побудови корпусу в контексті COVID-19, збагачення даних, розробки моделі, оцінювання та розгортання в окремих випадках може зайняти місяці або роки, що неприпустимо під час кризи в галузі охорони здоров'я, пов'язаної з COVID-19. У ситуації, що склалася на даний час, державні інституції беруть на себе витрати щодо створення корпусів. Водночас спільні анотації

сформовані науковими та дослідницькими спільнотами сприяють вирішенню проблем збагачення та анотації даних. Спільні завдання допомагають сприяти швидшій ітерації описаного процесу, централізуючи при цьому оцінювання та слугуючи джерелом анотованих даних.

Тому доцільно провести дослідження:

– сформованих на даний час корпусів;
– ресурсів для моделювання та видобування тексту;
– систем видобування тексту;
– спільних завдань.

Які були створені або реалізовані для підтримки процесів видобування тексту з наукових літературних джерел, опублікованих щодо COVID-19. При цьому слід відзначити унікальні та іноваційні системи, які забезпечують високу ефективність виконання фундаментальних завдань щодо пошуку, відповіді на запитання або забезпечують нові функціональні набори, зокрема узагальнення декількох документів або зв’язок між статтями щодо COVID-19 та клінічними випробуваннями. Також слід провести аналіз стратегій побудови ефективних та корисних систем тарґетованих на COVID-19, зокрема таких, що сприяють виробленню системних оглядів, або таких, що безпосередньо відповідають потребам медиків, дослідників та службовців в галузі охорони здоров’я.

1.2 Корпуси видобування тексту щодо COVID-19

Одним з перших та найбільших літературних корпусів, створених для підтримки видобування тексту щодо COVID-19, є відкритий набір даних щодо COVID-19 – CORD-19 [6], який містить сформований та оновлюваний корпус метаданих та повні тексти публікацій та препринтів, що стосуються COVID-19. Зазначені статті регулярно публікуються «Semantic Scholar» у співпраці з «Microsoft Research», «IBM Research», «Kaggle», «ініціативою

Чан-Цукерберг», Національною медичною бібліотекою Національних інститутів здоров'я (NIH) та Джорджтаунським центром безпеки та нових технологій. Корпус був вперше опублікований 16 березня 2020 року на запит Управління науки і технологій Білого дому, щоб підтримати зусилля наукової та дослідницької спільноти щодо застосування методів видобування тексту до літератури про COVID-19. Корпус об'єднує документи з PMC та сервери препринту «bioRxiv», «medRxiv» та «arXiv». Метадані публікацій з цих джерел проходять процедури узгодження, PDF-файли перетворюються на машиночитаний JSON за допомогою конвеєра «S2ORC», описаного в [7]. А HTML-представлення таблиць у документах додаються за допомогою «IBM Watson Discovery's Global Table Extractor» [8]. Корпус містить понад двісті шісдесят тисяч статей на паперових носіях. З них сто п'ять тисяч – це повнотекстові записи. Більшість систем видобування тексту, спрямованих на COVID-19, певною мірою використовують цей корпус.

«LitCovid» – це підготовлений набір пов'язаних з COVID-19 статей, опублікованих у відкритому доступі «PubMed» [9]. «LitCovid» містить понад п'ятдесят дві тисячі статей. Ряд систем опрацювання текстів використовують «LitCovid» як джерело даних. На початку тільки «LitCovid» надав необхідний додатковий набір статей щодо CORD-19. Оскільки перші випуски CORD-19 були зосереджені на «PMC», «bioRxiv» та «medRxiv» як джерелах і не включали документи з «PubMed». Однак сучасні випуски CORD-19 включають «PubMed» як джерело статей.

На даний час доступні інші набори публікацій щодо COVID-19, зокрема наприклад БД ВООЗ COVID-19 або БД CDC дослідницьких статей щодо COVID-19. Ці БД перекриваються з іншими корпусами, наприклад, БД ВООЗ включена до CORD-19, а значна частина бази даних CDC включена до «PubMed», PMC, CORD-19 та «LitCovid». БД CDC також забезпечує додаткову колекцію документів та технічних звітів щодо COVID-19.

Декілька видавців зібрали та випустили збірки своєї літератури про COVID-19. Зокрема, Інформаційний центр про коронавірус Elsevier [10], Основні моменти досліджень щодо коронавірусу Springer Nature [11], колекція COVID-19 мережі JAMA [12] або колекція «Science COVID-19» [13]. Велика кількість видавців надали літературу про COVID-19 під тимчасові ліцензії на відкритий доступ через ініціативу «PMC's Public Health Emergency COVID-19 Initiative» [14], роблячи таким чином тексти, доступні для громадськості через PMC та агреговані корпорації, наприклад, CORD-19. Це не включає всіх видавців, наприклад, JAMA не включено до цієї ініціативи. А це обмежує можливості наукової та дослідницької спільноти щодо формування справді всеосяжного корпусу. Повний текст також може бути недоступний у деяких випадках або може бути доступний лише у формі PDF-файлів, які повинні пройти складніше попереднє опрацювання для отримання повного тексту з метою використання процедур видобування тексту. Оскільки статус відкритого доступу для багатьох статей недостатньо визначений, то це може призвести до відкликання ліцензій та скасування наборів даних та систем у перспективі.

Практикуючий фахівець в галузі видобування тексту, наприклад, інженер, дослідник, ентузіаст тощо, може нести відповідальність за кожен із виконаних кроків виконаних у зоні з недостатньо чітко визначеними авторськими правами та ліцензіями, шляхом ідентифікації та адаптації існуючих наборів даних та моделей, або шляхом створення власних. Для COVID-19 централізація частин зазначеного робочого процесу допомогла зменшити навантаження на окремі з етапів ліцензування та авторства.

1.3 Ресурси для моделювання текстового майнінгу для COVID-19

Розглянемо детальніше ресурси для моделювання, які в основному використовуються для підтримки додаткових програм з метою видобування

текстів щодо COVID-19. Зазначені ресурси включають вкладені папери та концепції, текстові анотації багаторазового використання, графіки знань або контекстні мовні моделі, адаптовані до домену COVID-19.

Проаналізуємо відомості та моделі, що використовуються при створенні кожного ресурсу, а також подамо короткий його опис, який може допомогти в його використанні у системах видобування тексту.

Вбудовування (див. таблицю 1.1) – це обчислювані векторні подання інтервалів тексту, які фіксують семантичну та синтаксичну подібність між ними. Вбудовування можуть бути обчислені на різних рівнях деталізації, для лексем слова, іменованих сутностей, речень, абзаців, документів тощо. Існують десятки різних методологій вбудовування; для отримання додаткової інформації щодо їх використання [15].

Таблиця 1.1 – Ресурси по вбудовуванню для дослідників та фахівців з майнінгу тексту в предметній області COVID-19

Назва ресурсу	**Використані дані, модель**	**Приналеж-ність**	**Опис**
SPECTER CORD-19 вкладання	CORD-19	Інститут Аллена для ШІ	Вбудовування SPECTER [16] для статей CORD-19
COVID-19 вкладення концепції	CORD-19, SNOMED-CT	Університет штату Огайо	Вбудовування JET [17] для клінічних установ (SNOMED-CT) у корпусі CORD-19
CORD-19 SeVeN вкладання	CORD-19	Кардіфський університет	SeVeN [18] вбудовування слів, навчених на CORD-19
Вбудовані мережі спіль-ного викорис-тання [17]	CORD-19-on-FHIR	Клініка Майо	Мережеві вбудовування спільних випадків, навчені семантично анотованій версії CORD-19 (CORD-19-on-FHIR)

Декілька систем використовують вкладені документи та концепції для підтримки пошуку в літературі щодо COVID-19. Метод вбудовування «SPECTER» обчислює вбудовування документів за допомогою моделі SciBERT [19], яка попередньо підготовлена до сигналів спорідненості, отриманих з графіка цитування [20]. Вбудовування документів «SPECTER» успішно фіксує схожість документів [21] доступні для всіх публікацій у CORD-19. Також для статей на CORD-19 доступні вбудовування клінічної концепції, навчені за допомогою алгоритму «JET» [22], вбудовування відношень, навчені за допомогою «SeVeN» [23], та вбудовування спільних мереж [24] для біомедичних осіб, обчислених за допомогою «CORD-19-on-FHIR». Вбудовування фіксують схожість тексту і можуть використовуватися для отримання подібних текстів, наприклад вбудовування тексту запиту може бути використано для отримання відповідних документів з того самого простору вбудовування.

Анотації (див. таблицю 1.2) містять інформацію, що додається до метаданих та тексту публікації щодо COVID-19.

Таблиця 1.2 – Ресурси щодо анотацій, котрі призначені для дослідників та фахівців з майнінгу тексту COVID-19

Назва ресурсу	**Використані дані, модель**	**Приналеж-ність**	**Опис**
1	**2**	**3**	**4**
CODA-19 [29]	CORD-19	Пенський державний університет, UCSF, Університет Карнегі-Меллона	Набір даних про анотації аспектів дослідження публікацій у CORD-19

Продовження таблиці 1.2

1	2	3	4
CORD-19-on-FHIR	CORD-19, FHIR	Клініка Майо	Версія CORD-19 FHIR RDF з анотаціями стану, ліків та процедур клінічних осіб
COVID-19 DistillerSR	CORD-19, ClinicalTrials.gov	Evidence Partners	Зв’язки між ідентифікаторами клінічних випробувань та документами в CORD-19

В окремих дослідників можуть бути сформовані потреби ідентифікації будь-яких, пов’язаних з COVID-19, сутностей, біомедичних чи клінічних властивостей, відношень чи атрибутів у тексті статті [25]. Анотації можуть створюватися автоматично, наприклад, за допомогою попередньо навчених моделей для розпізнавання іменованих сутностей та зв’язків сутності «КВ», за допомогою інструментів «MetaMap Lite» або «ScispaCy» [26]. Або вручну за допомогою експертних анотацій, наприклад, запрошення експерта для позначення блоків, що описують популяцію, впливи, порівняння та результати у документах щодо клінічних випробувань та COVID-19.

Декілька груп опублікували анотації, що багаторазово використовуються незалежно або через платформи для обміну анотаціями, зокрема «PubTator» [27] або «PubAnnotation». Зокрема, у «PubAnnotation» для корпоративних CORD-19 та «LitCovid» доступні автоматично зґенеровані анотації термінів з множини онтологій та PICO-елементи.

Доступна також «CORD-19-on-FHIR» [28] – це версія CORD-19 із семантичними анотаціями для клінічного спостереження осіб у категоріях стану, ліків та процедур. Ця версія адаптована з метою полегшення процесів інтеграції в клінічні робочі процеси або використана для формування відомостей для клінічної підтримки прийняття рішень. Для проекту анотацій CODA-19 [29] автори демонструють функціональні можливості для

створення анотацій з використанням множини робіт у CORD-19. Спільні завдання також є джерелом анотацій створених експертами, зокрема, EPIC-QA формує позначені інтервали відповідей, а TREC-COVID – рейтинги документів. Це може бути використано практикуючими дослідниками в галузі текстового майнінгу для створення ефективніших систем.

Графи знань забезпечують модель сутностей та взаємозв'язків у певному домені (див. додаток А) можуть бути використані для подання фонових знань, виведення або виявлення нових відношень на основі логічних виведень та міркувань. Ряд графів знань COVID19 були побудовані шляхом об'єднання виявлених у літературі зв'язків з іншими онтологіями та базами даних структурованих зв'язків. Зокрема «CovidGraph» [30] на даний час є, практично найбільшим. Він поєднує літературу, статистику випадків, геномні та молекулярні дані. Проєкт інструмента графа знань [31], інтегрує корпус CORD-19 з генами, хімічними речовинами, хворобами та таксономічною інформацією з Вікіданих [32] та Порівняльної БД токсиксиноміки [33]. Blender Lab COVID-KG [34] презентує інший граф знань щодо COVID-19, орієнтований на переназначення медичних препаратів. Зазначені графи знань використовуються декількома системами видобування тексту з метою забезпечення перебігу процесів вивчення літератури на основі сутностей або відношень, або як спосіб візуалізації даних. Графи знань можуть підтримувати автоматизовані логічні міркування та висновки, потенційне відкриття нових відношень тощо.

Спеціально попередньо навчені контекстні мовні моделі (див. таблицю 1.3), є повсюдними в сучасних системах аналізу текстів. Ці моделі є найсучаснішими при опрацюванні природних мов і суттєво перевершили попередні базові показники у всьому спектрі мовних задач [35]. Багато проєктів в галузі текстового майнінгу використовують адаптовані до домену моделі BERT [36]. Зокрема SciBERT [37] та BioBERT [38] були

доопрацьовані відповідно до наукового та біомедичного тексту в області COVID-19.

Таблиця 1.3 – Ресурси мовних моделей для дослідників та фахівців з майнінгу тексту COVID-19

Назва ресурсу	**Використані дані, модель**	**Приналеж-ність**	**Опис**
COVID-KOP [39]	ROBOKOP, GO-анотації, CORD-19 анотації	UNC Chapel Hill SciBite	Поєднує графік біомедичних знань ROBOKOP з інформацією, вилученою з анотацій SciBite CORD-19
CovidBERT	CORD-19, BioBERT	ClinicalBERT	BioBERT та ClinicalBERT точно налаштовані на CORD-19
Green CovidSQu AD BERT [41]	CORD-19, Word2vec,	LMU Munich, Siemens SQuADBERT	Дешевий та ефективний спосіб досягнення домену. Адаптація для моделей BERT; досягається шляхом навчання. Word2vec та вирівнювання вбудовувань Word2vec до BERT-текстури.

Варіанти моделей BERT [40], доопрацьовані за літературою COVID-19, доступні у формі BioCovidBERT та ClinicalCovidBERT [41]. Авторами [42] обговорюється техніка адаптації домену, коли вектори word2vec [43], навчені в цільовому домені, використовуються для оновлення вбудовувань текстових заголовків у загальну модель мови домену на зразок BERT [44]. Що призводить до зменшення вартості та ресурсоємності. Але на даний час така альтернатива ще не є достатньо ефективною альтернатива. Попередньо навчені моделі забезпечують альтернативний спосіб опрацювання та пошуку вкладених текстів і можуть бути використані для пошуку або класифікації подібним чином до описаних вище інших типів векторних вбудовувань.

1.4 Висновок до першого розділу

В першому розділі кваліфікаційної роботи освітнього рівня «Бакалавр» подано розлогий аналіз предметної області опрацювання текстів в галузі COVID-19. Описано основи корпусів для видобування тексту щодо COVID-19. Виявлено та проаналізовано ресурси для моделювання текстового майнінгу для COVID-19. Зокрема подано аналіз ресурсів щодо вбудовування, анотацій, графів знань та спеціально попередньо навчених контекстних мовних моделей.

2 ЗАСОБИ ТЕКСТОВОЇ АНАЛІТИКИ ДЛЯ ОПРАЦЮВАННЯ ДАНИХ ПРО COVID-19

2.1 Системи опрацювання текстів щодо COVID-19

На даний час випущено обширне число систем для аналізу текстів для літератури щодо COVID-19. За відомостями [45] список присутніх на ринку налічує більш як тридцять дев’ять систем. Актуальний список систем можна переглянути на сторінці CORD-19 GitHub [46]. Системи видобування тексту включені до зазначеного списку збираються у відкритій формі на веб-сайті CORD-19, шляхом пошуку статей та препринтів щодо COVID-19 у корпусі CORD-19 та у соціальних мережах. Автори [45] пропустили системи, які з’являються у публікаціях на основі готового програмного забезпечення без додаткових даних чи методологічних розширень, або без подання характерних особливостей систем.

Всі включені до згаданого вище списку системи полегшують пошук та наукові розвідки літератури щодо COVID-19. Хоча деякі системи містять конкретніші завдання щодо розуміння тексту, зокрема, узагальнення, перевірка якості та перевірка претензій тощо. Для полегшення процедури порівняння між системами, доцільно використати ряд критеріїв, зокрема:

– дані, що використовуються;

– моделі та методи, що використовуються або впроваджуються кожною системою;

– підтримуваний функціонал користувацького інтерфейсу.

У деяких випадках відомості щодо систем не надаються або є неповними, зокрема щодо даних або моделей та методів, що використовуються.

Більшість задокументованих систем, використовують публічні корпуси та ресурси даних, до яких можна отримати доступ за відповідними

посиланнями на джерела. Корпорації, зокрема, CORD-19 та «LitCovid» тощо, задіюють широко використовувані ресурси даних, зокрема «ClinicalTrials.gov», UMLS та біомедичні онтології. При цьому вони дотримуються справедливих принципів доступності, сумісності та багаторазового використання [47]. Хоча деякі системи, зокрема «CovidScholar», «DOCSearch», «COVID-19 Intelligent Insight» використовують власні корпуси або приватні анотації разом з публічними наборами даних. Досить багато систем мають прозорі методи або відкритий вихідний код для відтворюваності. Проте деякі системи цього не роблять, тому щодо них немає можливості сформувати адекватні описи моделей.

Тому в процесі порівняння систем для опрацювання текстів щодо COVID-19 слід сформувати завдання на опрацювати тексту, які використовуються для категоризації та оцінки систем. Для кожного сформованого завдання потрібно:

– узагальнити особливості та методологію, що використовуються відповідними системами;

– виділити конкретні системи, які зробили додаткові кроки для адаптації свого інтерфейсу для реального використання біомедичними та клінічними дослідниками та практиками в предметній області COVID-19.

Такі додаткові кроки включають об’єднання даних літературних джерел з біомедичними БД, що використовуються в клінічних умовах. Або додавання анотацій, створених медичними експертами спеціально для завдань, пов’язаних з COVID-19.

2.2 Пошукові текстові системи для публікацій щодо COVID-19

Пошукові системи забезпечать наукові розвідки, в яких користувачам буде видано запити, що відображіють інформаційні потреби щодо COVID-19,

які кожна система задовольняє в повернутій множині відповідних документів.

Індексація та пошук можуть бути реалізовані за допомогою інструментів з відкритим кодом, зокрема Anserini [48], або комерційного програмного забезпечення, зокрема Amazon Kendra або Azure Cognitive.

Розглянемо детальніше деякі системи опрацювання текстів щодо COVID-19 для вирішення пошукових задач (див. таблицю 2.1).

Таблиця 2.1 – Системи опрацювання текстів щодо COVID-19 для вирішення пошукових задач

Система	**Дані**	**Методи/Моделі**	**Інтерфейс користувача**
Covidex [49]	CORD-19, ClinicalTrials.gov through TrialStreamer	Отримує фрагменти за допомогою Anserini [50]. Повторне оцінювання з використанням базової моделі T5 [51] доопрацьовано на біомедичному тексті та підготовлено для ранжування на MS MARCO [52].	Підтримує запити ключових фраз, написані користувачем. Виділяє відповідні терміни абстрактно. Перемикач для пошуку різних корпусів.
COVID papers browser	CORD-19, SNLI, MultiNLI	Відповідає запитам до статей за допомогою попередньо навчених вкладених речень: процедура навчання Sent BERT [53] із SciBERT [19], BioBERT [38], CovidBERT та ClinicalCovidBERT. Навчання на наборах даних SNLI [54] та MultiNLI [55].	Підтримує інтерактивні запити через командний рядок.

Запити можуть бути наборами ключових фраз, які подібні до підтримуваних традиційними пошуковими системами, зокрема Google або «PubMed». Браузер публікацій COVID, CoronaSearch та CovidScholar, обчислюють вбудовування для запитів та текстових інтервалів документу, речень або сутностей і повертають множину документів, що містять найближчих сусідів. Деякі системи обмежують запитуваний словниковий запас до сутностей у відомій БД. Наприклад, навігатор COVID-19 дозволяє терміни запитів у формі концепцій UMLS. A SPIKE-CORD [38] підтримує специфікацію регулярних виразів для забезпечення користувачам більшого рівня контролю результатів пошуку.

2.3 Системи опрацювання текстів щодо COVID-19 з функціями розвідки

Серед цих пошукових систем Covidex [56], fatcat, пошук по DOC, інтелектуальний аналіз COVID-19, пошук на основі технології Covid AI, навігатор COVID19 та CovidScholar інтегрують дані з багатьох джерел, виходячи за межі документів CORD-19 та «LitCovid», до інших баз даних, зокрема «ClinicalTrials.gov», «Lens», «Dimensions», документи з веб-сайтів ВООЗ або CDC тощо. Деякі системи також використовують зовнішні БД для зв’язування сутностей, зокрема «Vapur», який посилається на «ChemProt» [57], навігатор COVID-19 та «EVIDENCEMINER» (див. таблицю 2.2), які посилаються на UMLS. Або «AWS CORD- 19 Search» [58], який використовує зовнішні знання з «Comprehend Medical KB» [59].

Пошук DOC та «COVID-SEE» [60] – це цікаві системи, що включають видобуті елементи «PICO» та взаємозв’язки у візуалізації та дослідженні. Це може бути корисним при перегляді результатів щодо пов’язаних з COVID-19 клінічних досліджень. «COVIDExplorer» кластеризує видобуті ключові фрази без нагляду. «TopicForest» використовує вивчену ієрархію тем та організовує

видобуті ключові фрази для користувачів, хоча на даний час користувальницький інтерфейс недостатньо розроблений.

Таблиця 2.2 – Системи опрацювання текстів щодо COVID-19 для вирішення пошукових задач з функціями розвідки

Система	**Дані**	**Методи/Моделі**	**Інтерфейс користувача**
EVIDEN-CEMINER [61]	CORD-19, PubMed, UMLS	Отримує речення із подібними до запиту біомедичними об’єктами за допомогою контрольованих NER та OpenIE, детально в [62].	Підтримує написані користувачами запити ключових фраз, які можуть мати різну форму, результати класифікуються за рівнем доказовості, яку вони надають щодо запиту. Суб’єкти виділяються в результатах.
SPIKE-CORD [50]	CORD-19	Сутності та синтаксис витягнуті за допомогою ScispaCy [26]. Дані індексуються за допомогою Одинсона [63]. Підтримка синтаксису запитів.	Спеціалізована мова запитів підтримує оператори регулярних виразів (наприклад, символи підстановки, кількість збігів), відповідність за типами сутностей та синтаксичні шаблони.

2.4 Системи опрацювання текстів щодо COVID-19 спрямовані на дослідження та контроль якості джерел

Системи, зорієнтовані на дослідження COVID-19, допомагають користувачам знаходити та розуміти документи в корпусі. Такі системи можуть не бути спрямовані на задоволення конкретної інформаційної потреби, а скоріше використовуються щоб допомогти користувачам

зрозуміти основне джерело даних. Їх інтерфейси сприяють розфокусованому дослідженню даних та повторній взаємодії. Замість підтримки довільних запитів, написаних користувачем, ці системи можуть надавати заздалегідь визначений набір тем або ключових фраз, за допомогою яких можна фільтрувати документи. Теми публікації або документу можна аналогічним чином призначати за допомогою класифікації документів, як у «AWS CORD-19 Search» [67] (див. таблицю 2.3) яка класифікує документи, використовуючи сутності в БК «Comprehend Medical».

Таблиця 2.3 – Системи контролю якості текстів щодо COVID-19

Система	**Дані**	**Методи/Моделі**	**Інтерфейс користувача**
AWS CORD-19 Search [58]	CORD-19, медичні поняття Amazon	Багатозначна класифікація тем за статтями щодо медичних уявлень Amazon [59]. Можливий пошук за допомогою Amazon Kendra. Теми досліджень, вивчені за допомогою LDA.	Підтримує написані користувачем природні запитання, повертає рейтинговий список відповідних статей із виділеними інтервалами відповідей. Фільтрує результати за темами. Рекомендує подібні статті.
covidAsk [77]	CORD-19, Натуральні запитання, SQUAD	DenSPI для довших питань. BERN [65] для вилучення названої сутності. BioSyn [67] для організації, що посилається на CTD або NCBI. Навчено на природничих питаннях та наборах даних SQuAD.	Підтримує сформовані користувачем природні запитання, повертає рейтинговий список відповідних статей із виділеними інтервалами відповідей. Сутності в тексті документа також пов’язані із зовнішніми базами даних.

Ключові слова або фрази можуть бути видобуті з документів за допомогою контрольованого видобування біомедичних об’єктів, зокрема, «ScispaCy» [64] та «BERN» [65], або видобування ключових фраз без нагляду, зокрема «SGRank» [66]. Серед систем, що використовують БК є такі, що використовують куровані доменні БК. Вони забезпечують кращі результати роботи, оскільки сутності та відносини в цих БК перевіряються експертами відповідних доменів, зокрема COVID-19. Навігатор «COVID-19 IBM Watson» дозволяє користувачам виконувати логічні запити з використанням концепцій UMLS та семантичних типів [68].

Системи контролю якості (див. табл. 2.3) приймають запити у формі запитань та надають витягнуті фрагменти відповідей з документів. Більшість систем контролю якості літературних джерел щодо COVID-19 надають функції пошуку та контролю якості, отримуючи відповідні документи та виводячи інтервали відповідей. Через відсутність обширної множини верифікованих навчальних даних, специфічних для COVID-19, більшості існуючих систем контролю якості потрібно було завантажити власні навчальні дані щодо контролю якості або навчатись на ненаукових наборах даних, наприклад «SquAD» [70], або менших наборах QA з біомедичного домену, зокрема «BioASQ» [71] що може призвести до зменшення показників їх ефективності. «EPIC-QA» [72] має за мету створення загальнодоступних наборів даних COVID-19 для доопрацювання систем контролю якості.

2.5 Системи узагальнення текстів щодо COVID-19

Системи узагальнення мають на меті надати скорочену версію довшого фрагмента тексту. Щоб дозволити читачам проаналізувати ключові тези документа не витрачаючи зусиль на читання. Або сформувати швидкий анотований зміст документа, щоб читач прийняв рішення, чи варто витрачати більше часу на читання. Деякі надають додаткові функції, зокрема, «CAiRE-

Covid» [69] має функціонал для ґенерації підсумків для відповідей, а «Google COVID-19 Research Explorer» надає функціональні можливості для формування подальших запитань.

Дві системи в таблиці 2.4 мають інтегровані компоненти підсумовування. «Vespa CORD-19 Search» ґенерує зведення на рівні документа, а «CAiRE-Covid» [62] – система контролю якості, яка ґенерує зведення декількох документів по інтервалах відповідей.

Таблиця 2.4 – Системи узагальнення текстів щодо COVID-19

Система	**Дані**	**Методи/Моделі**	**Інтерфейс користувача**
CAiRE-Covid [17]	CORD-19, Біомедичні анотації	Отримання абзаців на основі ключових слів за допомогою Anserini. Повторне оцінювання та вибір відповідей із використанням ансамблю моделі контролю якості BioBERT [38] та узагальненої моделі MRQA. Абстрактно узагальнює відповіді.	Підтримує написані користувачем природні запитання, повертає рейтинговий список відповідних статей із виділеними інтервалами відповідей. Надає екстрактивні та абстрактні резюме за всі інтервали відповідей.
CORD-19 Search	CORD-19	Формує зведення доповідей за допомогою T5. Рекомендує подібні статті з використанням вкладених матеріалів SPECTER [20].	Підтримує запити ключових фраз, написані користувачем. Рекомендує подібні статті.

Система «CAiRE-Covid» ґенерує як екстракційні, так і абстрактні резюме, агрегуючи інформацію по інтервалах відповідей для вхідного запиту та забезпечуючи швидкий огляд поточного дослідження на високому якісному рівні. «Kbconstruction» описує системи, які створюють БК, видобуваючи сутності та відношення з тексту. БК може використовуватися для підтримки пошуку чи дослідження. Або може бути основною метою, як у базі даних «AIM COVID-19», яка надсилає документи на відповідні клінічні випробування та результати випробувань. БД AIM дозволяє користувачам відстежувати стан лікування та розроблення вакцин проти COVID-19.

Візуалізація забезпечує візуальний спосіб взаємодії та розуміння даних. Візуалізації, як правило, поєднуються з вилученими БК або мережами цитування, що забезпечують альтернативний спосіб дослідження корпусу наукових робіт щодо COVID-19. Зокрема «SemViz» [73] зосереджений на дослідженні корпусу CORD-19, лабораторії «Blender COVID-KG» та наборів даних про взаємодію білків та нуклеотидів. «SciSight» [74] дозволяє користувачам переглядати документи в CORD-19 вибираючи та сортуючи їх за автором, організаційною приналежністю, видобутими організаціями та мережевими відношеннями. Доповнені системи читання намагаються покращити стандартний досвід читання статей, надаючи функції підсвічування сутностей в документі та між посиланнями на документи. Зокрема, «COOVID-19 Intelligent Insight» виділяє видобуті сутності безпосередньо в PDF-документі.

2.6 Системні огляди наукових джерел щодо COVID-19

Мета системних оглядів – це синтез результатів за всіма відповідними опублікованими дослідженнями по темі COVID-19. При цьому потрібно забезпечити найвищу якість доведень та рекомендацій в процесі прийняття рішень в галузі охорони здоров’я. Системні огляди наукових публікацій

стали важливим елементом біомедичної літератури. Для неї на даний час існує багато встановлених протоколів щодо реєстрації, виробництва, публікації та оновлення [75]. На рисунку 3.1 подано ключові етапи формування системного огляду сформовані на основі відомостей поданих в [76].

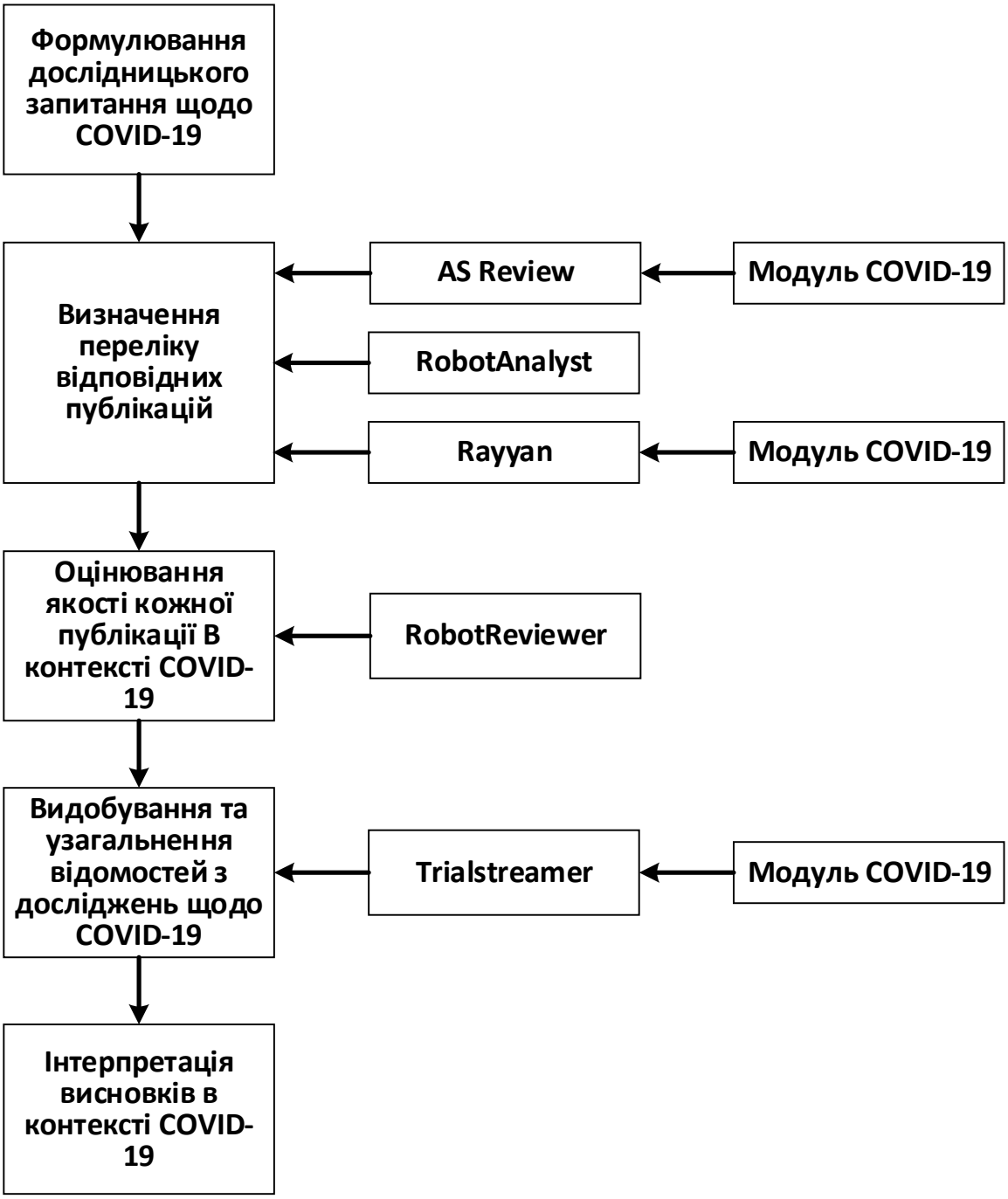

Рисунок 3.1 – Процес формування системного огляду наукових публікацій щодо COVID-19

Проведення системного огляду буде корисним при формуванні резюме видобутих тверджень та відношень, у процесах перезавантаження відомостей та публікацій. Багато завдань що стосуються опрацювання текстів в галузі COVID-19 можна сформулювати в контексті системного формування огляду наукових джерел. Пошук та перевірка якості можуть допомогти визначити відповідні документи та текстові фрагменти, заповнення таблиць допоможе видобути структуровані відомості з різних досліджень, а узагальнення множини документів – це спосіб агрегування відомостей між дослідженнями.

Системні огляди відіграють важливу роль у наукових дослідженнях в галузі COVID-19. Поширеними є швидкі огляди, які ущільнюють та скорочують багатомісячний або багаторічний системний огляд [77]. На даний час опубліковані експрес-огляди досліджень щодо рівня COVID-19 та смертності [78], клінічних характеристик субпопуляцій [79], симптомів захворювання [80], переназначення ліків [81], COVID- 19 управлінських стратегій [82], взаємодії між COVID-19 та іншими захворюваннями [83]. Через велику кількість оглядів в галузі COVID-19, що налічують тисячі, ми вирішили навести тут такі, що використовують корпуси COVID-19, такі як CORD-19 або LitCovid, як джерело досліджень на додаток до традиційних баз даних, таких як PubMed .

У міру зростання кількості публікацій про COVID-19 стає все важче та дорожче формувати та оновлювати огляди. Потрібні системи, які допомагають або автоматизують частини процесу перевірки. Декілька існуючих систем зосереджуються на автоматизованих частинах системного процесу опрацювання публікацій досить широко [84]. Ці системи зосереджені на підтримці виявлення відповідних досліджень в галузі COVID-19 [85] або вилученні елементів PICO [86]. Недавно випущена система «Trialstreamer», яка дозволяє користувачам відкривати нові клінічні випробування за допомогою пошуку на основі PICO [87]. «ASReview» [88],

«Rayyan» [89] та «Trialstreamer» [90] мають модулі COVID-19. Це дозволяє користувачам зосередитися виключно на роботах щодо COVID-19.

Процеси створення системних оглядів суттєво розвинулись впродовж декількох останніх десятиліть. Огляди надають надійні відомості дослідникам, лікарям та менеджерам. Ці відомості корисні для вирішення завдань, пов’язаних із перевантаженням інформації, оскільки досліджують та узагальнюють інформацію в рамках множини численних досліджень. Тарґетовані методи та системи, які допомагають або автоматизують системні огляди щодо COVID-19, ставатимуть значимішшими в майбутньому.

2.7 Узагальнення результатів проведених наукових розвідок щодо COVID-19

З початку пандемії COVID-19 наприкінці 2019 року та до теперішнього часу дослідницька спільнота представила обширну множину ресурсів та систем для видобування тексту. Які спрямовані на боротьбу із наростаючою хвилею нових наукових публікацій в предметній області COVID-19. На даний час сформована значна кількість версій корпусів, моделей, систем та спільних завдань. Незважаючи на значний прогрес, залишається багато відкритих питань. Підсумовуємо висновки та проблеми:

– Корисно та зручно мати централізований корпус документів. Наприклад CORD-19 або «LitCovid», які регулярно підтримуються та оновлюються. Існування цих корпусів дозволяє сучасній громаді дослідників зосередитись на формуванні моделей та розробці систем, пришвидшуючи ітерацію та створення нових методологій.

– Проміжна інфраструктура для спільного використання анотацій даних створених в автоматичному та ручному режимах збільшує сферу охоплення анотацій, зокрема «PubTator» та «PubAnnotation». Анотації, які

спільно використовуються цими платформами, можуть бути використані обширним переліком застосунків та програмно-алгоритмічних комплексів.

– Спільні завдання наукової спільноти можуть бути використані для об’єднання ресурсів для оцінювання та формування експертних оцінок щодо ефективності різних систем. Для COVID-19 цикли швидкого подання та оцінювання, що використовуються завданнями «Kaggle» та «TREC-COVID», імітують реалістичні задачі швидкого розвитку та розгортання систем. Ця реалістична чутливість, хоч і складна для реалізації для розробників, може привести до створення надійніших систем, які можуть швиде адаптуватися до швидкозмінних даних та потреб користувачів і дослідників.

– Важливо залучати експертні спільноти в предметній області COVID-19 на ранніх етапах якомога частіше, щоб зосередитись на реальних завданнях та потребах користувачів і дослідників. Завдання слід вибирати, відповідно до максимізації схожості з відповідними робочими процесами в предметній області COVD-19, зокрема пошук публікацій або системне формування огляду. Оскільки існуючі на даний час робочі процеси перевірені та відомі як корисні, прикріплення спільних завдань до цих робочих процесів, призведе до формування ефективніших систем.

Значна частина інформаційно-технологічної та обчислювальної інфраструктури існує десятиліттями. Реалії COVID-19 змусили наукову спільноту пришвидшити процеси проведення досліджень, включаючи етапи розробки наборів даних, розробки та впровадження моделей, оцінювання та публікації результатів досліджень. Адаптація до зазначених змін спричинила рад труднощів. Закрема:

– Попередні випуски корпусу CORD-19 були нестабільними, регулярно змінювались формати відповідно до адаптації для інженерних завдань та запитів користувачів.

– Спільні завдання повинні відповідно коригуватися. Наприклад, «TREC-COVID» був організований у п’яти етапів. Це вимагало дуже

швидкого реагування розробників, що подають систему на розгляд та експертів-оцінювачів, які зазвичай працюють спокійніших часових обмеженнях.

– Потрібен час для визначення найкращих способів залучення медичних експертів до оцінювання. Для аналітичних систем такого рівня складності завдання спеціального пошуку є чітко визначеними та історично визнано корисним та важливим завданням з видобування тексту щодо COVID-19. Е звязку з пандемією COVID-19, наукові дослідження в галузі видобування текстів виконануються у вужчих часових проміжках, ніж зазвичай. Проте, завдяки напряцюванням з видобування текстів у інших галузях, відбулося відносно легка адаптація експертів-оцінювачів в предметній області COVSD-19. Однак у випадку з «Kaggle» початковы завдання, сформованы на першому етапы, були занадто відкритими, і, як наслыдок, подання було відповідно різноманітним і важким для порівняння. Для експертів-медиків формувалась множина з понад п’ятсот подань для ручного оцінювання.

Методи видобування тексту значно вдосконалились впродовж останніх десятиліть. Внаслідок COVID-19 постала задача тестування існуючих методів та розроблення нових в умовах з обмеженим часом та ресурсами. В таких випадках автоматизація або обчислювальна допомога можуть бути найбільш корисними. Попередні результати наукових розвідок є багатообіцяючими. В продовж останнього періоду часу спроєктовано, розроблено та випущено декілька десятків виробничих систем, пристосованих до різних аспектів пошуку в предметній області COVID-19. До проведення досліджень залучено обширну наукову спільноту, котра включає багато біологічних та медичних експертів, котрі залучені для оцінювання багатьох аналітичних систем та інструментів, які на даний час є загальнодоступними. Прагнучи допомогти дослідникам управляти інформаційним перевантаженням, системи використовують методи

видобування тексту, щоб допомогти у формуванні швидких оглядів наукової літератури в предметній області COVID-19. Хоча було досягнуто значного прогресу в наукових дослідженнях щодо COVID-19, необхідні значні вдосконалення для забезпечення значущих та ефективних результатів у боротьбі з глобальною пандемією COVID-19.

2.8 Висновок до другого розділу

В другому розділі кваліфікаційної роботи освітнього рівня «Бакалавр» описано системи опрацювання текстів щодо COVID-19. Досліджено пошукові текстові системи для публікацій щодо COVID-19. Проаналізовано системи опрацювання текстів щодо COVID-19 з функціями розвідки. Описано системи опрацювання текстів щодо COVID-19 спрямовані на дослідження та контроль якості джерел. Розглянуто системи узагальнення текстів щодо COVID-19. Досліджено системні огляди наукових джерел щодо COVID-19. Проведено узагальнення результатів проведених наукових розвідок щодо COVID-19.

3 БЕЗПЕКА ЖИТТЄДІЯЛЬНОСТІ, ОСНОВИ ХОРОНИ ПРАЦІ

3.1 Долікарська допомога при вивихах

Вивихом суглоба називається зміщення суглобової поверхні кісток, що супроводжується пошкодженням капсульно-зв'язкового апарату. Вивих супроводжується, як правило, пошкодженням м'яких тканин суглоба. При цьому можуть рватися зв'язки з судинами, капсула, а також сухожилля прилеглих м'язів [92]. Це пошкодження веде до серйозних порушень функціональності як окремого суглоба, так і всієї кінцівки. Необхідно зазначити, що велику роль для повноцінного відновлення функції відіграє правильне та своєчасне надання першої допомоги при вивихах суглобів.

Під час вивиху відбувається заміщення кісток у суглобі. Залежно від місця вивиху зовнішні прояви можуть відрізнятися, але симптоми у них схожі. Вивих плечового суглоба. Залежно від того, в який бік зміщена голівка плечової кістки, пошкодження може бути нижнім, заднім або переднім. Останній зустрічається в 98% випадках. Біль при будь-якій спробі руху в суглобі, набряк, зниження рухливості суглоба, порушення чутливості та кровопостачання – основні симптоми цього стану. Лікуванням вивиху плечового суглоба повинен займатися виключно фахівець.

Причинами вивиху колінного суглоба може бути сильний удар, падіння з висоти, вроджені аномалії і різке скорочення м'яза стегна при швидкій ходьбі. Симптомами такого ушкодження будуть біль, набряклість, оніміння кінцівки, зміна форми коліна, специфічне положення ноги і неможливість рухатися. Оскільки травма вважається серйозною, перша допомога, лікування та реабілітація після вивиху колінного суглоба має проходити під контролем лікаря.

Вивих кульшового суглоба буває переднім і заднім. Передній може з'явитися в результаті падіння з висоти. Причиною заднього найчастіше є

автодорожня травма. До симптомів ушкодження відносяться сильний біль, вимушене положення кінцівки, зміна форми суглоба і невелике вкорочення постраждалої кінцівки.

Після проведення огляду і пальпації пошкодженої ділянки травматолог (а саме він займається такими травмами, включаючи вроджені вивихи суглобів) направить пацієнта на рентгенографію. Після підтвердження діагнозу лікар може надати повну профільну допомогу, починаючи з усунення вивиху і закінчуючи необхідною іммобілізацією або фіксацією з доповненням при необхідності медикаментозним і загальноукріплюючим лікуванням. По закінченні терміну іммобілізації назначається програма по реабілітації з відновленням рухів та прогнозуванням результатів лікування.

Основним завданням першої допомоги є повне знерухомлення пошкодженого суглоба в положенні мінімальної болісності (шинування, бинтування, до тіла/кінцівки, туга пов’язка та ін.). Якщо в області травми є ушкодження шкіри, його необхідно обробити перекисом водню та накласти чисту (в ідеалі стерильну) пов’язку або забинтувати рану бинтом. Для зменшення набряку до суглоба прикладають холодний компрес. Не рекомендується давати постраждалому воду або знеболювальне у вигляді таблеток. Вищеперерахованих заходів буде в більшості випадків достатньо до приїзду швидкої, для надання допомоги з усуненням вивиху в медичному закладі в найкоротший термін, основною умовою якого є порожній шлунок травмованого.

Лікування будь-якого вивиху в першу чергу передбачає вправляння суглоба і повернення його у фізіологічне положення. Цю процедуру рекомендовано проводити в найкоротші терміни під загальним наркозом (особливо у дітей) і в деяких випадках під місцевою анастезією (вивихи області кисті або стопи). Рішення в такому випадку приймає лікар, враховуючи стан хворого та ступень ураження. Операцію при звичному

вивиху плечового суглоба проводить травматолог-ортопед або хірург, що спеціалізується на спортивних травмах.

У разі несвоєчасного лікування існує ймовірність розвитку контрактури. При застарілих вивихах (більше 3 тижнів) функція суглоба може бути незворотно втрачена.

До основних методів реабілітації, в тому числі і відновлення після вивиху променево-зап’ясткового та локтьового суглоба в дорослих, відносяться ЛФК, масаж і фізіотерапія. У якості останньої широко застосовують лазеротерапію, ультразвук, міоелектростимуляцію, магнітотерапію та інтерференцтерапію.

Виконуючи лікувальні вправи після вивиху плечового або колінного суглоба, ви в найкоротші терміни відновите рухливість суглоба, поліпшите кровообіг і збільшите його гнучкість. Із цією ж метою під час реабілітації призначаються різні види масажу. Зауважте, що після вивиху суглоба період відновлення в кожного пацієнта індивідуальний і потребує регулярного нагляду ортопеда та реабілітолога, в деяких випадках невролога, масажиста та фізіотерапевта.

Профілактикою вивихів може бути тільки профілактика травматизму або рання діагностика (проф. огляд при вродженій патології) у новонароджених.

3.2 Правила техніки безпеки при експлуатації обладнання

Аналітичне опрацювання даних про COVID-19 здебільшого відбувається з використанням онлайн засобів, котрі реалізовано з використанням серверного обладнання. Тому розглянемо детальніше правила техніки безпеки при його експлуатації. Побудова і облаштування серверного приміщення (серверної кімнати або апаратної) є першим і основним кроком в організації телекомунікаційної інфраструктури будь-якого підприємства.

Серверне приміщення слід розміщувати якомога ближче до магістральних кабельних каналів. Або ж проектувати майбутню кабельну інфраструктуру відповідним чином [93].

В ідеалі, серверне приміщення повинно бути поруч з головним розподільчим пунктом (Main Cross, MC), а якщо є можливість, то слід організувати головний розподільний пункт безпосередньо в самому серверному приміщенні.

Переважно розміщувати апаратну недалеко від вантажних або вантажопасажирських ліфтів, використовуваних для транспортування важкого обладнання, наприклад ДБЖ. Водночас, слід уникати близького розміщення потужних джерел електричних і магнітних полів, а також обладнання, яке може викликати підвищену вібрацію.

Вібрація негативно впливає на роботу активного обладнання, контакти і з'єднання. У діапазоні частот до 25 Гц амплітуда коливань не повинна перевищувати 0.1 мм.

Багато джерел рекомендують розташовувати апаратну в геометричному центрі будівлі хоча б тому, що це дозволяє істотно заощадити на прокладці кабелю.

Через апаратну не повинні прокладатися транзитом трубопроводи інженерних систем будівлі Забороняється розташовувати апаратну поряд з приміщеннями для зберігання пожежонебезпечних або агресивних хімічних матеріалів. Забороняється розташовувати апаратну в приміщенні, суміжному з приміщенням виробництв з мокрими технологічними процесами.

Не рекомендується виділяти приміщення для апаратної на верхніх поверхах будівлі, оскільки вони найбільш схильні до пошкоджень у разі пожежі і можуть заливатися при протіканнях даху.

Не рекомендується розміщувати серверне приміщення поруч з сходовими прольотами, ліфтовими шахтами, великими вентиляційними

каналами та іншими елементами будівлі, які, згідно існуючого планування, можуть обмежити розширення серверного приміщення в майбутньому.

Не допускається розміщення апаратної під приміщеннями, пов'язаними зі споживанням води: туалети, душові, їдальні, буфети і т. д.) Серверне приміщення потрібно розмістити осторонь від джерел електромагнітних завад на такій відстані, щоб напруженість електричного поля не перевищувала 3 В/м у всьому спектрі частот.

Система контролю і управління мікрокліматом повинна забезпечити заданий рівень вологості і температури необхідний для нормального функціонування активного обладнання. Система мікроклімату повинна забезпечити підтримку температурного режиму не тільки влітку, а й взимку і розрахована на цілодобову безперервну роботу. Якщо централізована система мікроклімату в будівлі не може забезпечити безперервну роботу і заданий рівень температури і вологості, то необхідно встановити автономну систему.

Після прокладки кабелів необхідно закрити вогнетривким матеріалом всі кабельні вводи в серверне приміщення. Для цих цілей можна використовувати спеціальні заглушки, які встановлюються в кабельному вводі, які у разі виникнення пожежі розширюються, перекривають простір і не дозволяють поширитися вогню і диму. Стельові перекриття, стіни і перегородки серверного приміщення повинні бути негорючими та забезпечувати вогнестійкість не менше 45 хвилин.

3.3 Висновок до третього розділу

В третьому розділі кваліфікаційної роботи описано долікарську допомогу при вивихах. Подано правила техніки безпеки при експлуатації серверного обладнання.

ВИСНОВКИ

В першому розділі кваліфікаційної роботи освітнього рівня «Бакалавр»:

– Проведено аналіз предметної області.

– Подано опис корпусів для видобування тексту щодо COVID-19.

– Описано ресурси для моделювання текстового майнінгу для COVID-19.

В другому розділі кваліфікаційної роботи:

– Описано системи опрацювання текстів щодо COVID-19.

– Досліджено пошукові текстові системи для публікацій щодо COVID-19.

– Проаналізовано системи опрацювання текстів щодо COVID-19 з функціями розвідки.

– Описано системи опрацювання текстів щодо COVID-19 спрямовані на дослідження та контроль якості джерел.

– Розглянуто системи узагальнення текстів щодо COVID-19.

– Досліджено системні огляди наукових джерел щодо COVID-19.

– Проведено узагальнення результатів проведених наукових розвідок щодо COVID-19.

У розділі «Безпека життєдіяльності, основи хорони праці» описано долікарську допомогу при вивихах. Подано правила техніки безпеки при експлуатації серверного обладнання.

ПЕРЕЛІК ДЖЕРЕЛ

1 Oleksii Duda, Liliana Dzhydzhora, Oleksandr Matsiuk, Andrii Stanko, Nataliia Kunanets, Volodymyr Pasichnyk, Oksana Kunanets. Mobile Information System for Monitoring the Spread of Viruses in Smart Cities. SISN. 2020; Volume 8: pp. 65 - 70.

2 Oleksii Duda, Oleksandr Matsiuk, Nataliia Kunanets, Volodymyr Pasichnyk, Antonii Rzheuskyi and Yuriy Bilak, “Formation of Hypercubes Based on Data Obtained from Systems of IoT Devices of Urban Resource Networks”, International Journal of Sensors, Wireless Communications and Control (2020) 10: 1. ISSN 2210-3287.

3 Kilicoglu H. Biomedical text mining for research rigorand integrity: tasks, challenges, directions. *Brief Bioinform* 2018;19:1400–14.

4 Zweigenbaum P, Demner-Fushman D, Yu H, *et al.* Frontiers of biomedical text mining: current progress. *Brief Bioinform* 2007;8:358–75.

5 Duda, O., Kunanets, N., Martsenko, S., Matsiuk, O., Pasichnyk, V., Building secure Urban information systems based on IoT technologies. CEUR Workshop Proceedings 2623, pp. 317-328. 2020.

6 CORD-19. COVID-19 Open Research Dataset. https://www.semanticscholar.org/cord19.

7 Lo K, Wang LL, Neumann M, *et al.* S2ORC: the semantic scholar open research corpus. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online.* 2020.

8 Zheng X, Burdick D, Popa L, *et al.* Global table extractor (GTE): a framework for joint table identification and cell structure recognition using visual context. 2020. Preprint. Archive: arXiv; Identifier: 2005.00589.

9 Chen Q, Allot A, Zhiyong L. Keep up with the latest coronavirus research. *Nature* 2020;579:193.

10 Novel Coronavirus Information Center. Elsevier’s free health and medical research on the novel coronavirus (SARS-CoV-2) and COVID-19. https://www.elsevier.com/connect/coronavirus-information-center.

11 Coronavirus (COVID-19) Research Highlights. https://www.springernature.com/gp/researchers/campaigns/coronavirus.

12 Coronavirus Disease 2019 (COVID-19). https://jamanetwork.com/ journals/jama/pages/coronavirus-alert.

13 Coronavirus: Research, Commentary, and News. https://www.sciencemag.org/collections/coronavirus.

14 Public Health Emergency COVID-19 Initiative. https://www.ncbi.nlm.nih.gov/pmc/about/covid-19/.

15 Duda, O., et al, Selection of Effective Methods of Big Data Analytical Processing in Information Systems of Smart Cities. CEUR Workshop Proceedings 2631, pp. 68-78. 2020.

16 Dong L, Yang N, Wang W, *et al.* Unified language model pre-training for natural language understanding and generation.In: *Advances in Neural Information Processing Systems*, Vol. 29, . Vancouver, Canada, 2019.

17 Benjamin E, Nye JJL, Patel R, *et al.* A corpus with multilevel annotations of patients, interventions and outcomes to support language processing for medical literature. *Proc Conf Assoc Comput Linguist Meet* 2018;2018:197–207.

18 Köksal A,Dönmez H,Ozçelik R,*et al.*Vapur: a search engine to find related protein—compound pairs in COVID-19 literature. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: NLP-COVID Workshop, Online*. 2020.

19 Beltagy I, Lo K, Cohan A. SciBERT: a pretrained languagemodel for scientific text. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on*

Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019, 3615–20.

20 Cohan A, Feldman S, Beltagy I, *et al.* Specter: documentlevel representation learning using citation-informed transformers. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. 2020.

21 Lewis M, Liu Y, Goyal N, *et al.* BART: denoising sequenceto-sequence pre-training for natural language generation, translation, and comprehension. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, 7871–80.

22 Newman-Griffis D, Lai AM, Fosler-Lussier E. Jointly embedding entities and text with distant supervision. In: *Proceedings of the Third Workshop on Representation Learning for NLP*. Melbourne, Australia: Association for Computational Linguistics, 2018, 195–206.

23 Espinosa-Anke L, Schockaert S. SeVeN: augmenting wordembeddings with unsupervised relation vectors. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018, 2653–65.

24 Oniani D, Jiang G, Liu H, *et al.* Constructing co-occurrence network embeddings to assist association extraction for COVID-19 and other coronavirus infectious diseases. *J Am Med Inform Assoc* 2020;27(8):1259–67.

25 Duda O., Kunanets N., Matsiuk O., Pasichnyk V., Rzheuskyi A. (2021) Aggregation, Storing, Multidimensional Representation and Processing of COVID-19 Data. In: Shakhovska N., Medykovskyy M.O. (eds) Advances in Intelligent Systems and Computing V. CSIT 2020. Advances in Intelligent Systems and Computing, vol 1293, pp 875-889. Springer, Cham. ISBN978-3-030-63270-0

26 Neumann M, King D, Beltagy I, *et al.* ScispaCy: fast and robust models for biomedical natural language processing. In: *Proceedings of the 18th BioNLP*

Workshop and Shared Task. Florence, Italy: Association for Computational Linguistics, 2019, 319–27.

27 U.S. National Library of Medicine. National Center for Biotechnology Information. PubTator Central. https://www.ncbi.nlm.nih.gov/research/pubtator/.

28 CORD-19-on-FHIR -- Semantics for COVID-19 Discovery. https://github.com/fhircat/CORD-19-on-FHIR.

29 Huang T-H, Huang C-Y, Ding C-KC, *et al.* CODA-19: reliably annotating research aspects on 10,000+ CORD-19 abstracts using a non-expert crowd. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: NLPCOVID Workshop, Online*. 2020.

30 COVID-19 Knowledge Graph. https://covidgraph.org/.

31 Ilievski F, Garijo D, Chalupsky H, *et al.* KGTK: a toolkit for large knowledge graph manipulation and analysis. In: *Proceedings of the 19th International Semantic Web Conference, Online.* 2020.

32 WIKIDATA. https://www.wikidata.org/wiki/Wikidata:Main_Page.

33 Comparative Toxicogenomics Database. Illuminating how chemicals affect human health. http://ctdbase.org/.

34 Wang Q, Li M, Wang X, *et al.* COVID-19 literature knowledge graph construction and drug repurposing report generation. 2020. Preprint. Archive: arXiv; Identifier: 2007.00576.

35 Liu Y, Ott M, Goyal N, *et al.* RoBERTa: a robustly optimized BERT pretraining approach. 2019.

36 Devlin J, Chang M-W, Lee K, *et al.* BERT: pre-training of deep bidirectional transformers for language understanding.In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Vol. 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019, 4171–86.

37 Peters M, Neumann M, Iyyer M, *et al.* Deep contextualized word representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics, 2018, 2227–37.

38 LeeJ,YoonW,KimS,*etal*.BioBERT:apre-trainedbiomedical language representation model for biomedical text mining. *Bioinformatics* 2019;36(4): 1234–40.

39 Korn D, Bobrowski T, Li M, *et al.* COVID-KOP: integrating emerging COVID-19 data with the ROBOKOP database. *ChemRxiv* 2020. Preprint. Archive: ChemRxiv, Identifier:10.26434/chemrxiv.12462623.

40 Panahi L, Amiri M, Pouy S. Clinical characteristics ofCOVID-19 infection in newborns and pediatrics: a systematic review. *Arch Acad Emerg Med* 2020;8(1):e50. https://doi. org/10.22037/aaem.v8i1.634.

41 Covid-BERTs. https://github.com/manueltonneau/covid-berts.

42 Poerner N, Waltinger U, Schutze H. Inexpensive domainadaptation of pretrained language models: case studies on biomedical NER and COVID-19 QA. 2020. Preprint. Archive: arXiv, Identifier: 2004.03354.

43 Mikolov T, Sutskever I, Chen K, *et al.* Distributed representations of words and phrases and their compositionality. In: *Proceedings of Advances in Neural Information Processing Systems 26 (NIPS)*. Lake Tahoe, USA, 2013.

44 Popa IV, Diculescu M, Mihai C, *et al.* COVID-19 and inflammatory bowel diseases: risk assessment, shared molecular pathways and therapeutic challenges. *Gastroenterol Res Pract* 2020;2020:1918035. doi: 10.1155/2020/1918035.

45 Wang, Lucy Lu, and Kyle Lo. "Text mining approaches for dealing with the rapidly expanding literature on COVID-19." *Briefings in Bioinformatics* 22.2 (2021): 781-799.

46 Allenai cord19. https://github.com/allenai/cord19.

47 Wilkinson M, Dumontier M, Aalbersberg IJ, *et al*. The fair guiding principles for scientific data management and stewardship. *Sci Data* 2016;3:160018.

48 Yang P, Fang H, Lin J. Anserini: enabling the use of Lucenefor information retrieval research. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17*. New York, NY, USA: Association for Computing Machinery, 2017, 1253–6.

49 Taboureau O, Nielsen SK, Audouze K, *et al*. Chemprot: a disease chemical biology database. *Nucleic Acids Res* 2011;39:D367–72.

50 Tabib HT, Shlain M, Sadde S, *et al.* Interactive extractive search over biomedical corpora. In: *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*.Association for Computational Linguistics, 2020, 28–37.

51 Raffel C, Shazeer N, Roberts A, *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*. 2019;21(140):1–67.

52 Campos DF, Nguyen T, Rosenberg M, *et al.* MS MARCO: a human generated machine reading comprehension dataset. In: *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS)*. Barcelona, Spain, 2016.

53 Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, 3982–92.

54 Bowman SR, Angeli G, Potts C, *et al.* A large annotated corpus for learning natural language inference. In: *Proceedings of the 2015 Conference on*

Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015, 632–42.

55 Williams A, Nangia N, Bowman S. A broad-coverage challenge corpus for sentence understanding through inference. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long Papers). New Orleans,Louisiana: Association for Computational Linguistics, 2018, 1112–22.

56 Zhang E, Gupta N, Tang R, *et al.* Rapidly Deploying a Neural Search Engine for the COVID-19 Open Research Dataset: Preliminary Thoughts and Lessons Learned. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: NLP-COVID Workshop, Online*. 2020.

57 Taboureau O, Nielsen SK, Audouze K, *et al.* Chemprot: a disease chemical biology database. *Nucleic Acids Res* 2011;39:D367–72.

58 Bhatia P, Arumae K, Pourdamghani N, *et al.* AWS CORD19search: a scientific literature search engine for COVID-19. 2020. Preprint. Archive: arXiv; Identifier: 2007.09186.

59 Bhatia P, Celikkaya B, Khalilia M, *et al.* Comprehend medical: a named entity recognition and relationship extraction web service. In: *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. Baton Rouge, USA, 2019, 1844–51.

60 Verspoor K, Suster S, Otmakhova Y,ˆ *et al.* COVID-see: scientific evidence explorer for COVID-19 related research. 2020. Preprint. Archive: arXiv; Identifier: 2008.07880.

61 Wang X, Liu W, Chauhan A, *et al.* Automatic textual evidence mining in COVID-19 literature. 2020. Preprint. Archive: arXiv; Identifier: 2004.12563.

62 Wang X, Guan Y, Liu W, *et al.* EVIDENCEMINER: textual evidence discovery for life sciences. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 2020, 56–62.

63 Valenzuela-Escárcega MA, Hahn-Powell G, Bell D. Odinson: a fast rule-based information extraction framework. In: *Proceedings of the 12th Language Resources and Evaluation Conference*.Marseille,France:EuropeanLanguageResources Association, 2020, 2183–91.

64 Portenoy J, West JD. Constructing and evaluating automated literature review systems. *Scientometrics* 2020;1–19.

65 Kim D, Lee J, So CH, *et al.* A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE*;7:73729–40.

66 Danesh S, Sumner T, Martin JH. SGRank: combining statistical and graphical methods to improve the state of the art in unsupervised keyphrase extraction. In: *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*.Denver,Colorado:AssociationforComputational Linguistics, 2015, 117–26.

67 Sung M, Jeon H, Lee J, *et al.* Biomedical entity representations with synonym marginalization. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online.* 2020.

68 Bodenreider O. The unified medical language system(UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;32(Database issue):D267–70.

69 Su D, Xu Y, Yu T, *et al.* CAiRE-COVID: a question answering and multi-document summarization system for COVID-19 research. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: NLP-COVID Workshop, Online*. 2020.

70 Rajpurkar P, Zhang J, Lopyrev K, *et al.* SQuAD: 100,000+ questions for machine comprehension of text. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, 2016, 2383–92.

71 Tsatsaronis G, Balikas G, Malakasiotis P, *et al.* An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform* 2015;16: article number 138.

72 Tang R, Nogueira R, Zhang EM, *et al.* Rapidly bootstrapping a question answering dataset for COVID-19. 2020. Preprint. Archive: arXiv; Identifier: 2004.11339.

73 Jingxuan T, Verhagen M, Cochran BH, *et al.* Exploration and discovery of the COVID-19 literature through semantic visualization. 2020. arXiv abs/2007.01800.

74 Hope T, Portenoy J, Vasan K, *et al.* SciSight: combining faceted navigation and research group detection for COVID-19 exploratory scientific search. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online*. 2020.

75 Booth A, Clarke M, Dooley G, *et al.* The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. *Syst Rev* 2012;1:2.

76 Khan KS, Kunz R, Kleijnen J, *et al.* Five steps to conducting a systematic review. *J R Soc Med* 2003;96:118–21.

77 Khangura SD, Konnyu KJ, Cushman R, *et al.* Evidence summaries: the evolution of a rapid review approach. *Syst Rev* 2012;1:10.

78 Yang H, Li VOK, Lam JCK, *et al.* Who is more susceptible to COVID-19 infection and mortality in the states? *medRxiv* 2020.Preprint.Archive: medRxiv,Identifier: 10.1101/2020.05.01.20087403.

79 El-shafeey F, Magdi R, Hindi N, *et al.* A systematic scoping review of COVID-19 during pregnancy and childbirth. *Int J Gynaecol Obstet* 2020;150(1):47–52.

80 Parasa S, Desai M, Chandrasekar VT, *et al.* Prevalence of gastrointestinal symptoms and fecal viral shedding in patients with coronavirus disease 2019. *JAMA Netw Open* 2020;3(6):e2011335.

81 Sadegh S,Matschinske J,Blumenthal DB,*et al.*Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. *Nature Communications*. 2020;11:Article 3518.

82 Yaacoub S, Schünemann HJ, Khabsa J, *et al.* Safe management of bodies of deceased persons with suspected or confirmed COVID-19: a rapid systematic review. *BMJ Glob Health* 2020;5(5):e002650.

83 Crisan-Dabija R, Grigorescu C, Pavel CA, *et al.* Tuberculosis and COVID-19in2020:lessons from the past viral out breaks and possible future outcomes. *Canadian Respiratory Journal* 2020;2020:1401053. https://doi.org/10.1155/2020/1401053.

84 Tsafnat G, Glasziou PP, Choong MK, *et al.* Systematic review automation technologies. *Syst Rev* 2014;3:74–4.

85 ASReview Core Development Team. *ASReview: Active Learning for Systematic Reviews*.Utrecht,The Netherlands:Utrecht University, 2019.

86 de Bruijn B, Carini S, Kiritchenko S, *et al.* Automated information extraction of key trial design elements from clinical trial publications. *AMIA Annu Symp Proc* 2008;141–5.

87 Nye B, Nenkova A, Marshall I, *et al.* Trialstreamer: mapping and browsing medical evidence in real-time. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 2020, 63–9.

88 van de Schoot R, de Bruin J, Schram RD, *et al.* ASReview: open source software for efficient and transparent active learning for systematic reviews. 2020. Preprint. Archive: arXiv; Identifier: 2006.12166.

89 Ouzzani M, Hammady HM, Fedorowicz Z, *et al.* Rayyan— a web and mobile app for systematic reviews. *Syst Rev* 2016;5:210.

90 Wang LL,Lo K,Chandrasekhar Y,*et al.*CORD-19: the COVID19 open research dataset. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: NLPCOVID Workshop, Online*. 2020.

91 Wise C, Ioannidis VN, Calvo MR, *et al*. COVID-19 knowledge graph: accelerating information retrieval and discovery for scientific literature. 2020. Preprint. Archive: arXiv; Identifier: 2007.12731.

92 ДОБРОБУТ. ЗДОРОВ'Я КРАЇНИ. Що включає в себе надання першої допомоги при вивихах суглобів. https://www.dobrobut.com/ua/services/library/c-cto-vklucaet-v-seba-okazanie-pervoj-pomosi-pri-vyvihah-sustavov.

93 Вимоги до серверної (серверному приміщенню, апаратної). https://shop.hypernet.com.ua/ua/trebovaniya-k-servernoy-komnate/.

ДОДАТКИ

Додаток А

Ресурси щодо графів знань, котрі призначені для дослідників та фахівців з майнінгу тексту COVID-19

Назва ресурсу	Використані дані, модель	Приналеж-ність	Опис
SciBite COVID-19 анотації	CORD-19	SciBite	Анотації спільного вживання речення та сутності; анотація об’єктів з MeSH, GO, HPO, HGNC, ChEMBL та інших
CovidGraph	CORD-19, Lens, Ensembl, NCBI гени, онтології генів, експериментальні дані, набір даних Johns Hopkins 2019-nCoV	Багато академічних та галузевих організацій	Графік знань статей COVID-19, статистики випадків, генів та функцій, а також молекулярних даних
KGTK COVID-19 KnowledgeGraph [36]	CORD-19, WikiData, CTD, Blender Lab COVID-KG	USC, Папський Католицький університет Ріо-де-Жанейро	Графік знань, що інтегрує корпус CORD-19 з генами, хімічними речовинами, таксономічною інформацією з баз даних Вікіданих та CTD та лабораторії Blender COVID-KG
Blender Lab COVID-KG [89]	CORD-19	UIUC	Графік знань з генами типів сутностей, хворобами, хімічними речовинами та організмами та підтипами, отриманими з тексту та співвідношеннями рисунок/ підпис у літературі
COVID-19 KnowledgeGraph [91]	CORD-19, зрозумійте медицину [47]	Amazon Web Services (AWS)	Графік знань COVID-19; вбудовані графіки використовуються для забезпечення пошуку AWS CORD-19