https://doi.org/10.31649/1997-9266-2024-177-6-135-144

УДК 004.9+578.834.1

А. В. Лосенко[1]
Є. М. Крижановський[1]
І. М. Штельмах[1]
І. В. Варчук[1]

ТЕХНОЛОГІЯ LLM-ВИДОБУВАННЯ ОЗНАК ТЕСТУВАННЯ ПАЦІЄНТІВ З ТЕКСТОВИХ ЗВІТІВ ДЛЯ УДОСКОНАЛЕННЯ ПРОГНОЗУВАННЯ КІЛЬКОСТІ ХВОРИХ НА КОРОНАВІРУС

[1]Вінницький національний технічний університет

Розглянуто застосування сучасних великих мовних моделей (LLM) для автоматизованого видобування важливих ознак з аналітичних текстових звітів про пандемію COVID-19 в Україні в період 2020—2022 років. Ці звіти охоплюють широкий спектр даних, включно з регіональними показниками захворюваності, динаміку тестування, результати вакцинації та демографічні характеристики пацієнтів. У дослідженні розглянуто інтеграцію цих видобутих ознак у моделі часових рядів для підвищення точності епідеміологічних прогнозів.

Центральним елементом дослідження є використання моделі Prophet, яку вдосконалено для врахування сезонних змін і аномалій у даних. У дослідженні вирішувалися такі виклики, як багатохвильовий характер часового ряду COVID-19, включно з різкими підйомами і спадами захворюваності. Здійснено коригування аномалій, спричинених змінами в карантинних заходах, політиці тестування та вакцинаційних кампаніях, особливо в періоди зимових сплесків.

Оптимізація моделі Prophet включала вдосконалене налаштування параметрів за допомогою таких методів, як Grid Search і стохастична оптимізація, адаптованих до специфічного епідеміологічного контексту України. Додатково у дослідженні оцінено потенціал нейромережевих моделей, зокрема LSTM (Long Short-Term Memory), для аналізу часових рядів. Здатність LSTM виявляти нелінійні залежності та обробляти велику кількість вхідних параметрів доповнює традиційні методи, забезпечуючи глибше розуміння довгострокових трендів і взаємозв'язків у даних.

Мета цієї статті полягає у створенні ефективного інструменту для прогнозування епідеміологічної динаміки, здатного враховувати багатофакторний характер даних, що описують пандемію COVID-19, шляхом інтеграції нових ознак, отриманих із текстових аналітичних звітів за допомогою великих мовних моделей (LLM), у часовий ряд.

Ключові слова: інформаційна технологія, інженерія ознак, прогнозування часових рядів, Prophet, штучний інтелект, великі мовні моделі, COVID-19

Вступ

Прогнозування кількості захворювань на COVID-19 стало критично важливим інструментом для прийняття обґрунтованих рішень у сфері громадського здоров'я. Це включає планування ресурсів, визначення пріоритетів у тестуванні та вакцинації, а також оцінку ефективності впроваджених заходів. Ефективне прогнозування дозволяє зменшити ризики перевантаження медичних закладів, забезпечити своєчасне постачання необхідних матеріалів та сприяти раціональному використанню ресурсів. У цьому контексті важливим є впровадження сучасних моделей аналізу даних, таких як Prophet.

Модель Prophet розроблено для обробки часових рядів з урахуванням сезонних трендів, нестабільностей та аномалій у даних. Її гнучкість дозволяє адаптувати модель до різних сценаріїв, через що вона є цінним інструментом для прогнозування епідемічних хвиль. Наприклад, у випадку

COVID-19, модель здатна враховувати фактори, такі як зростання або зниження кількості тестувань, регіональні особливості та вплив масових вакцинацій. Prophet довела свою ефективність у багатьох попередніх дослідженнях, спрямованих на аналіз складних динамічних процесів.

Попередні дослідження, проведені у 2020—2023 роках [1], [2], підкреслюють переваги використання Prophet у прогнозуванні динаміки COVID-19. Одним з ключових аспектів було налаштування параметрів моделі для врахування багатохвильового характеру пандемії, коли зростання та спад кількості захворювань відбувалися в декілька етапів. До того ж, приділялась увага обробці аномальних даних, спричинених різкими змінами у політиці тестування або вакцинації. Такі підходи дозволили значно підвищити точність прогнозів [3].

Водночас, потенціал для вдосконалення прогнозування все ще залишається. Зокрема, сучасні дослідження вказують на можливість збагачення моделей додатковими ознаками, отриманими з текстових джерел за допомогою великих мовних моделей (LLM). Ці моделі здатні автоматично видобувати релевантну інформацію з текстових звітів, новин та інших джерел, що дозволяє ідентифікувати приховані патерни в даних. Такий підхід відкриває нові горизонти для аналізу часових рядів, надаючи можливість точніше прогнозувати динаміку захворювань.

Інтеграція результатів, отриманих за допомогою LLM, з моделями часових рядів, такими як Prophet, створює синергію, яка може значно підвищити якість прогнозів. Наприклад, додаткові ознаки, такі як рівень громадської активності, соціальні настрої або політичні рішення, можуть бути враховані в моделях. Це дозволяє створювати точніші та адаптивніші прогнози, які враховують складну природу пандемій та їхній вплив на суспільство.

Мета статті полягає у створенні ефективного інструменту для прогнозування епідеміологічної динаміки, здатного враховувати багатофакторний характер даних, які описують пандемію COVID-19, шляхом інтеграції нових ознак, отриманих із текстових аналітичних звітів за допомогою великих мовних моделей (LLM), у часовий ряд.

Огляд методів інженерії ознак та способи застосування великих мовних моделей

Інженерія ознак є одним з ключових етапів у моделюванні часових рядів, оскільки якісно підготовлені ознаки напряму впливають на точність прогнозів. Наприклад, у завданнях прогнозування динаміки COVID-19 модель Prophet вже продемонструвала свою ефективність. Використовуючи такі ознаки, як сезонність, аномалії та тренди, модель Prophet дозволяє визначити хвилі поширення вірусу, враховуючи змінні фактори, такі як політика тестування чи вакцинація. Розглянемо типи ознак, які ця модель підтримує за замовчуванням.

Сезонна декомпозиція є важливим процесом, що включає виокремлення сезонних компонентів для розуміння базових трендів і циклів. Ідентифікація аномалій спрямована на визначення і корекцію аномальних значень, які можуть спотворювати результати прогнозу. Створення змінних, які базуються на попередніх значеннях часового ряду, дозволяє врахувати вплив минулих подій на поточну динаміку. Для фінансових даних це може бути критично важливим, дозволяючи врахувати циклічні закономірності. Календарні ознаки додають контекстуальні індикатори, які відображають святкові дні, вихідні або специфічні події, що можуть впливати на динаміку досліджуваних явищ. Для епідемічних даних ці ознаки можуть включати періоди карантину або вакцинації, що значно впливають на тренди. Фільтрування шуму спрямоване на згладжування часового ряду, усунення випадкових флуктуацій і покращення точності аналізу основних трендів. Використання методів, таких як ковзне середнє, дозволяє зосередитися на довгострокових трендах і ігнорувати короткострокові коливання [4].

Сучасні великі мовні моделі (LLM), такі як GPT, Gemini, BERT, Claude, T5 і Bloom, відкривають нові можливості для видобування додаткових ознак із текстових джерел. Наприклад, вони здатні аналізувати текстові звіти, соціальні медіа та наукові публікації для виявлення прихованих трендів і патернів, які неможливо визначити зі стандартних числових даних. Це може включати аналіз настроїв, ідентифікацію ключових подій або визначення тематичних трендів у динамічних умовах, таких як кризи чи пандемії [5].

Такі моделі здатні використовувати Prompt Engineering для налаштування на специфічні завдання, до прикладу, класифікації подій за впливом або ідентифікації ключових індикаторів. Зокрема, GPT-4 може бути корисним для визначення кореляцій між соціальними подіями та економічними показниками, а BERT ефективно знаходить ключові фрази для точнішої класифікації текстових даних. Використання стратегії RAG (Retrieval-Augmented Generation) дозволяє моделям

інтегрувати актуальну інформацію з зовнішніх баз знань, що значно підвищує їхню ефективність [6].

Інтеграція цих підходів дозволяє підвищити цінність даних, створюючи синтезовані ознаки, що враховують як числові, так і текстові показники. Такий підхід значно розширює можливості аналізу часових рядів, відкриваючи нові горизонти для точніших прогнозів.

Таким чином, інтеграція LLM у процес інженерії ознак відкриває нові горизонти для аналізу часових рядів. Використання цих підходів у моделях, таких як Prophet, дозволяє враховувати як традиційні часові ряди, так і додаткові фактори, що забезпечує точніші та адаптивніші прогнози.

Аналіз структури звітів прогнозів розвитку епідемії COVID-19

Аналітичні звіти, що описують епідеміологічну ситуацію COVID-19 в Україні, складалися протягом 2020—2022 років [7]. Вони є цінним джерелом даних для аналізу, оскільки містять багатий спектр інформації, необхідної для досліджень і прогнозування. Структура звітів зазвичай охоплює кілька ключових розділів, кожен з яких висвітлює специфічні аспекти епідеміологічної ситуації.

Розглянемо зміст одного з таких звітів [8]. Розділ «Вступ» надає загальну характеристику епідемічної динаміки, відображаючи основні тенденції та ключові фактори, які впливають на поширення вірусу. Він також містить основні показники, такі як загальна кількість нових випадків, одужань та летальних випадків за визначений період. Ця інформація є базою для подальшого детального аналізу. У розділі «Загальнонаціональна епідемічна динаміка» викладено комплексний аналіз динаміки захворюваності, тестування та летальних випадків. Розділ деталізується у підрозділах, які висвітлюють динаміку госпіталізацій, навантаження на лікарні та специфіку інфікування серед вакцинованих. Наприклад, у підрозділі 1.2 розглянуто показники заповнення лікарень і тенденції госпіталізацій, що можуть бути корисними для оцінки ефективності протиепідемічних заходів. Демографічні дані аналізуються у спеціальному розділі, де акцентується на віковій структурі інфікованих, рівні смертності та рівні одужання серед різних вікових груп. Це дозволяє оцінити вразливість окремих демографічних категорій до вірусу.

Окремо в звітах подано регіональний аналіз, де порівнюються дані за різними адміністративними одиницями. Розділ «Аналіз затримок оприлюднення даних» зосереджується на проблемах, пов’язаних із затримкою у зборі та публікації даних. Це питання є критичним для точності прогнозування, адже несвоєчасні дані можуть спотворювати результати моделі.

Прогнозування розвитку епідемії, зокрема за допомогою моделі Prophet, подано у розділі 6. Ця модель використовується для аналізу часових рядів і дозволяє оцінити майбутні тренди на основі історичних даних. Наприклад, прогноз на основі даних за 2022 рік містив оцінку впливу нових хвиль пандемії та передбачав потенційні піки захворюваності.

Також у цих аналітичних звітах відбувалося порівняння прогнозів за попередні періоди дослідження, що дає змогу оцінити якість прогнозів в динаміці.

Компартментні моделі, розглянуті у розділі 8, надають додатковий інструмент для оцінки епідемічної ситуації. Вони дозволяють враховувати репродуктивне число та інші ключові показники, необхідні для розробки сценаріїв подальшого розвитку пандемії.

Висновки кожного звіту підсумовують основні результати аналізу та надають рекомендації для подальших дій. Зазвичай ці висновки включають відносні значення, що дозволяє оцінити ефективність вжитих заходів.

Для подальшого дослідження звіти можуть бути корисними для отримання специфічних ознак. Наприклад, інформація про епідемічну динаміку в останніх абзацах звітів містить числові показники нових випадків, одужань та летальних випадків. Дані про динаміку захворюваності та госпіталізацій у розділах 1.1—1.3 можуть бути представлені у відсоткових значеннях, що допомагає збагачувати датасети. Репродуктивне число, яке часто наводиться перед графіками у розділі 8, є критично важливим для моделювання.

Застосування методів RAG та Prompt Engineering для видобування ознак за допомогою LLM

Методи RAG (Retrieval-Augmented Generation) і Prompt Engineering відкривають широкі можливості для ефективного видобування ознак з текстових джерел за допомогою сучасних великих мовних моделей (LLM). Ці методи забезпечують інтеграцію текстової інформації з контекстними даними для створення точніших і адаптивних моделей аналізу, що є особливо актуальним для роботи з великими обсягами неоднорідних даних [9].

Метод RAG дозволяє моделі поєднувати генеративні можливості з доступом до зовнішніх джерел даних. Завдяки цьому підходу можна забезпечити точнішу обробку текстів, отриманих з різних джерел, до прикладу, архівів звітів, баз даних або веб-ресурсів. У рамках цього підходу, для

локального використання RAG завантажено звіти у форматах PDF та HTML, які згодом перетворено у векторні представлення (ембедінги) та збережено у векторному сховищі. Наприклад, платформа Ollama із застосуванням моделей Llama і ChatGPT надає можливість працювати з такими локальними даними, що включають епідеміологічні звіти. Цей підхід гарантує збереження конфіденційності та дозволяє зосередитися на використанні специфічної інформації, що підвищує точність моделі [10].

Prompt Engineering, зі свого боку, є інструментом для тонкого налаштування моделей, що дозволяє адаптувати їх до конкретних завдань. За допомогою модифікованих промптів можна отримати необхідні ознаки зі складних текстових структур. Наприклад, для аналізу розділу «Епідемічна динаміка» у звітах COVID-19 можна створити промпт, що спрямовує модель на отримання числових значень, таких як кількість нових випадків, одужань і летальних випадків. Аналогічно, для розділів 1.1—1.3 можна налаштувати промпти на пошук тенденцій у відсоткових значеннях або оцінку змін навантаження на медичну систему [11].

Додатково, методи RAG і Prompt Engineering можуть використовуватися для виділення критично важливих показників, таких як репродуктивне число, що зазвичай наводиться перед графіками у звітах. Використання адаптованих промптів допомагає моделі коректно ідентифікувати це значення, навіть якщо воно розташоване серед значного обсягу текстової інформації. Діаграма діяльності, що описує метод RAG показана на рисунку.

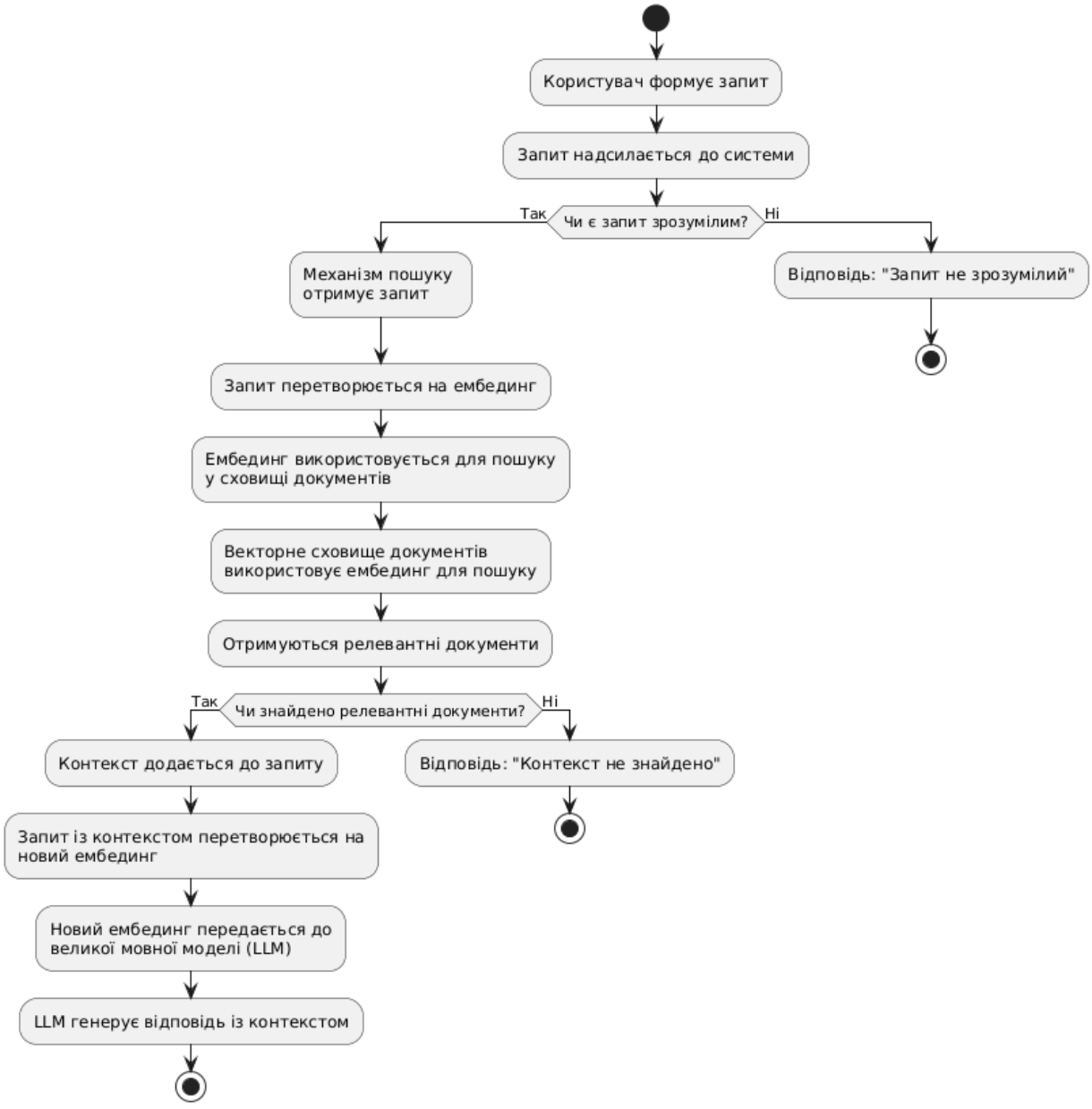

Рис. 1. Опис застосування методу RAG для розширення контексту LLM

До того ж, інтеграція методів RAG дозволяє використовувати актуальні дані з зовнішніх джерел у режимі реального часу, що значно розширює аналітичні можливості. Це особливо корисно для адаптації моделей до динамічних змін у контексті пандемії, наприклад, для аналізу даних про регіональні хвилі захворюваності чи вплив вакцинації.

Таким чином, комбінація методів RAG і Prompt Engineering забезпечує ефективне отримання ознак із текстових звітів, створюючи основу для вдосконалення аналізу та прогнозування. Їхнє застосування дозволяє глибше зрозуміти динаміку даних та адаптувати аналітичні моделі до потреб конкретного дослідження [12], [13].

Результати застосування методів RAG та Prompt Engineering для видобування додаткових ознак

У межах дослідження здійснено аналіз 39 аналітичних звітів, підготовлених Робочою групою з математичного моделювання проблем, пов'язаних з пандемією COVID-19 в Україні. Звіти охоплювали період з 2020 по 2022 роки та містили докладну інформацію про епідеміологічну ситуацію, включно з даними про епідемічну динаміку, демографічними показниками, навантаженням на медичні заклади та прогнозами. Комплексний характер цих звітів робив їх цінним джерелом для автоматизованого видобування ознак.

Для видобування ключових ознак використано методи RAG і Prompt Engineering, які дозволили врахувати наявність специфічних показників у звітах та забезпечити узгодженість між ними. Перед обробкою звіти структуровано, а дані — попередньо очищено. Це гарантувало високу якість результатів та дало змогу максимально ефективно застосувати сучасні інструменти аналізу тексту.

Серед видобутих ознак враховано кілька ключових параметрів, отриманих з різних розділів звітів. Розділ «Епідемічна динаміка» (розділ 1) надав числові показники, такі як кількість нових випадків інфікування, одужань та летальних випадків. Ці дані, зазвичай зазначені в останньому абзаці кожного звіту, забезпечили основу для аналізу загальної динаміки епідемії.

Розділи 1.1, 1.2 та 1.3, які стосувалися динаміки захворюваності, навантаження на лікарні та госпіталізацій серед вакцинованих, містили відносні значення у відсотках. Такі дані важливі для оцінки змін у системі охорони здоров'я та мали потенціал для значного розширення датасету, доповнюючи числові параметри якісними оцінками. Розділ 8, щодо прогнозування на основі компартментної моделі, включав значення репродуктивного числа. Цей параметр особливо корисний, оскільки часто наводився перед графіками, що дозволяло інтегрувати його у прогнозні моделі для точнішого визначення трендів епідемії.

Усі ці параметри оброблено, структуровано та включено до загального CSV файлу, що дозволило інтегрувати їх у часовий ряд. Це значно розширило можливості аналітичних моделей, забезпечуючи гнучкість у роботі з різноманітними типами даних.

Зібрані дані збережено у форматі CSV, що забезпечує простоту інтеграції в наявні аналітичні системи. Файл оптимізовано для подальшої обробки, включно з тематичною категоризацією ознак, що спрощує їхнє використання в різних дослідницьких контекстах. Це дозволяє легко адаптувати дані до нових завдань та потреб користувачів [14], [15].

Методи, застосовані в дослідженні, також дозволили збагачувати моделі прогнозування новими параметрами. Наприклад, включення даних про відсоткові зміни завантаженості лікарень дозволило оцінювати короткострокові ризики, тоді як використання репродуктивного числа сприяло побудові довгострокових прогнозів. Такі ознаки значно підвищили адаптивність і точність аналітичних моделей.

Аналіз трендів у звітах виявив приховані залежності, які залишалися непомітними в процесі традиційного числового аналізу. Це, зокрема, стосувалося регіональних відмінностей у динаміці епідемії та впливу соціально-економічних факторів. Узагальнення ключових висновків звітів дозволило розробити рекомендації, спрямовані на вдосконалення заходів протидії пандемії.

Збереження результатів у структурованому форматі забезпечує гнучкість їхнього подальшого використання в аналітичних моделях. Це також відкриває можливості для створення нових прогнозних систем, орієнтованих на розв'язання специфічних задач, зокрема адаптації до локальних умов чи змін у політиці охорони здоров'я.

Застосування отриманих ознак за допомогою LLM в прогнозуванні часового ряду моделлю Prophet

У цьому розділі подані результати використання моделі Prophet для прогнозування епідеміологічної динаміки. Основну увагу приділено впливу отриманих ознак, інтегрованих у часовий ряд, на точність прогнозів. Здійснено порівняння трьох основних підходів: використання моделі Prophet з параметрами за замовчуванням, моделі Prophet з урахуванням святкових періодів як аномалій, а також моделі Prophet, доповненої ознаками, отриманими за допомогою методів RAG та Prompt Engineering.

Прогнозування охоплювало три ключові звітні періоди: 2020, 2021 та 2022 роки. Ці періоди вибрано на основі даних з аналітичних звітів Робочої групи, які відображали унікальні характеристики кожного року. До прикладу, 2020 рік характеризувався першими хвилями пандемії та впровадженням карантинних обмежень, тоді як 2021 рік супроводжувався масовими вакцинаціями та сплесками нових штамів. У 2022 році спостерігався спад активності пандемії, але залишалися регіональні аномалії. Цей підхід дозволив забезпечити всебічну оцінку ефективності моделей.

Модель Prophet з параметрами за замовчуванням слугувала базовим підходом, використовуючи лише часовий ряд без додаткових параметрів. Вона продемонструвала стабільність у прогнозуванні, проте не завжди точно враховувала аномальні події, такі як різкі хвилі захворюваності чи раптові спади. Це обмеження стало очевидним у періоди суттєвих коливань у динаміці епідемії. Додавання святкових періодів як аномалій до моделі Prophet дозволило врахувати сезонні ефекти, що впливають на епідеміологічну динаміку. Наприклад, у 2021 році включення святкових днів дозволило передбачити сплески інфікувань після масових зібрань. Хоча модель показала покращення в порівнянні з базовим підходом, вона все ще мала труднощі з прогнозуванням подій, не пов’язаних з сезонними факторами.

Модель Prophet, доповнена отриманими ознаками, забезпечила найкращі результати. Включення таких параметрів, як динаміка навантаження на лікарні, відсоткові зміни захворюваності та репродуктивне число, дозволило моделі враховувати більше контекстуальних змін. Наприклад, під час третьої хвилі пандемії в 2021 році ці параметри допомогли значно зменшити відхилення прогнозів від реальних даних. Врахування додаткових ознак сприяло більшій адаптивності моделі до нетипових змін у динаміці захворюваності.

Графіки порівняння прогнозів за трьома моделями показано для кожного із зазначених періодів (рис. 2—4).

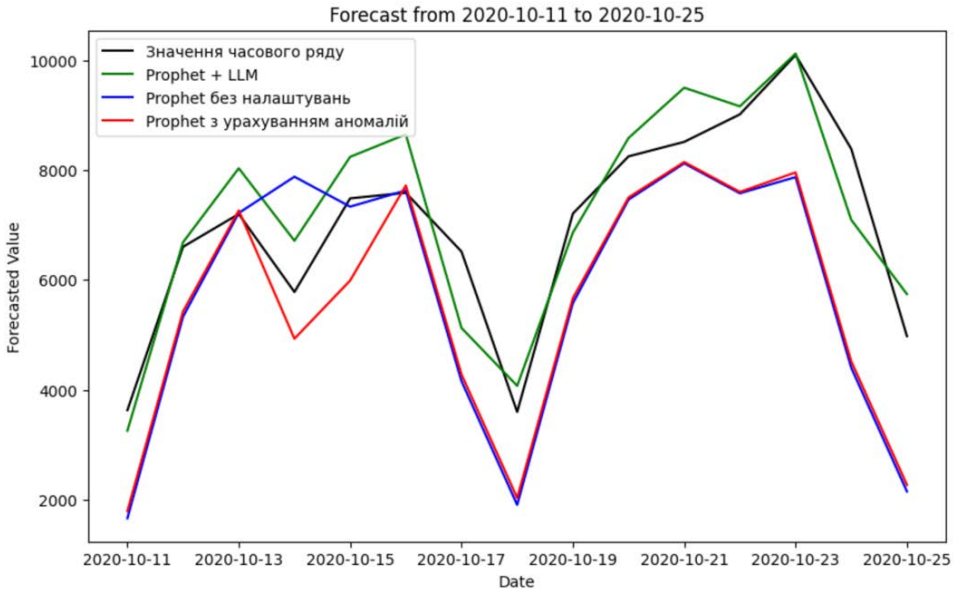

Рис. 2. Порівняння прогнозів за допомогою моделі Prophet без додаткових налаштувань, моделі Prophet з урахуванням свят як аномалій, та моделі Prophet з розширеним датасетом з ознаками, отриманими за допомогою LLM за період 11.10. —25.10.2020

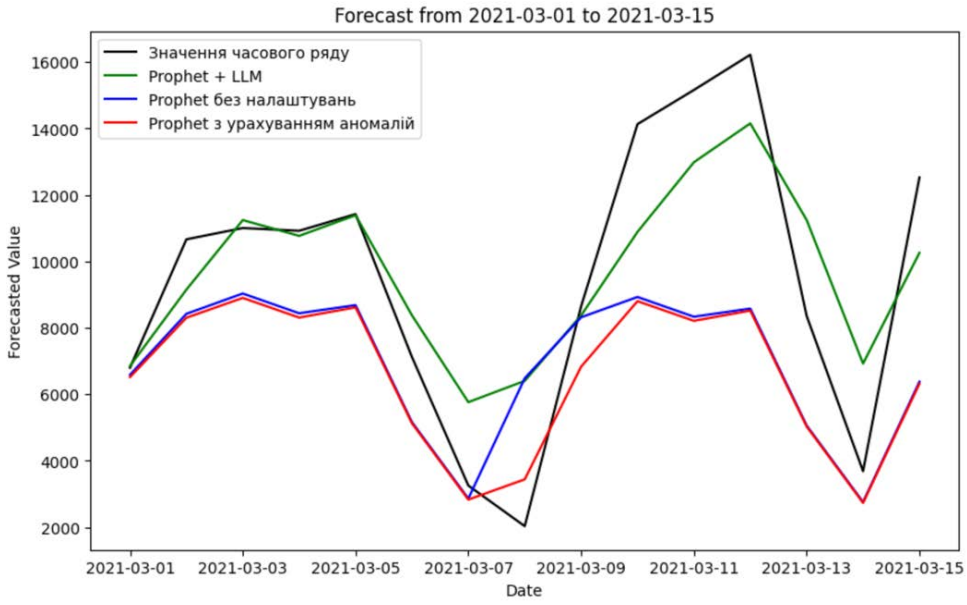

Рис. 3. Порівняння прогнозів за допомогою моделі Prophet без додаткових налаштувань, моделі Prophet з урахуванням свят як аномалій, та моделі Prophet з розширеним датасетом з ознаками, отриманими за допомогою LLM за період 01.03. —15.03.2021

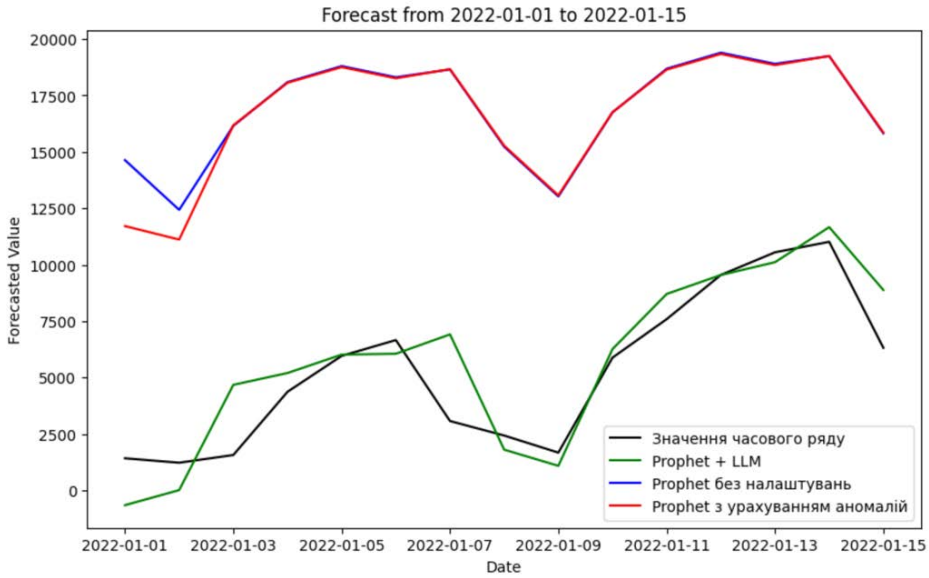

Рис. 4. Порівняння прогнозів за допомогою моделі Prophet без додаткових налаштувань, моделі Prophet з урахуванням свят як аномалій, та моделі Prophet з розширеним датасетом з ознаками, отриманими за допомогою LLM за період 01.01. —15.01.2022

З метою порівняння точності прогнозу моделі Prophet без додаткових налаштувань, та моделі Prophet з даними, отриманими за допомогою LLM, обчислимо значення метрик MSE (Mean Squared Error, середньоквадратична похибка), та метрики MAE (Mean Absolute Error, середня абсолютна похибка). Значення наведених метрик такі:

– MSE для моделі Prophet без додаткових налаштувань: 975676786,26.

– MSE для моделі Prophet з даними, отриманими за допомогою LLM: 309394597,17.

– MAE для моделі Prophet без додаткових налаштувань: 30043,39.

– MAE для моделі Prophet з даними, отриманими за допомогою LLM: 15320,84.

Згідно з отриманими значеннями, дані, отримані за допомогою LLM та застосовані з моделлю Prophet дають приріст точності на 48,98 % (за метрикою MAE) та 68,30 % (за метрикою MSE).

Значення метрик точності прогнозу та наведені графіки порівнянь чітко ілюструють переваги використання моделі Prophet з отриманими ознаками. Зокрема, модель з додатковими ознаками показала найбільшу точність у періоди різких змін, що підкреслює її релевантність у контексті

прогнозування складних систем.

Таким чином, результати дослідження демонструють, що інтеграція отриманих ознак за допомогою LLM у модель Prophet значно підвищує її точність. Це підкреслює важливість використання сучасних аналітичних методів для поліпшення прогнозування та прийняття обґрунтованих рішень у складних епідеміологічних умовах.

Висновки

Проведений аналіз показав, що використання методів RAG та Prompt Engineering дозволяє автоматизувати процес отримання релевантних ознак з текстових звітів. Усього оброблено 39 аналітичних звітів, які охоплювали різні аспекти епідеміологічної ситуації в Україні за період 2020—2022 років. Отримані ознаки, такі як кількість нових випадків, рівень одужань, летальність, репродуктивне число та інші ключові параметри, збережено у структурованому вигляді та інтегровано в часовий ряд.

Порівняння ефективності трьох підходів до прогнозування, а саме моделі Prophet з параметрами за замовчуванням, моделі Prophet з урахуванням святкових періодів як аномалій та моделі Prophet з доданими ознаками, продемонструвало суттєві переваги останнього підходу. Зокрема, включення додаткових ознак дозволило значно підвищити точність прогнозів, особливо у періоди різких змін динаміки захворюваності.

Результати аналізу підтвердили, що інтеграція видобутих ознак дозволяє моделям гнучкіше реагувати на змінні умови та краще враховувати контекстні фактори. Це стало можливим завдяки тому, що ознаки, отримані з текстових звітів, забезпечували додаткову інформацію про структуру та тренди даних, які не завжди доступні в числових рядах.

Додатково, використання видобутих ознак дозволило покращити прогнозування довгострокових трендів та врахувати сезонні та регіональні відмінності. Наприклад, завдяки інклюзії параметрів, що відображали навантаження на медичну систему та динаміку вакцинацій, вдалося значно підвищити точність у періоди пікових навантажень.

Ґрунтуючись на порівнянні метрик точності прогнозу моделей та аналізі графіків порівняння прогнозів, зауважимо, що модель Prophet з отриманими ознаками забезпечує найкращі результати серед розглянутих підходів. Найбільше поліпшення спостерігалося під час аномальних періодів, таких як хвилі пандемії чи регіональні сплески захворюваності. Це підтверджує ефективність інтеграції текстових ознак у контексті моделювання складних систем, що містять як числові, так і якісні параметри.

Отже, результати дослідження демонструють, що методи отримання ознак з текстових джерел за допомогою LLM є перспективним підходом до підвищення якості прогнозування. Їхнє впровадження дозволяє створювати адаптивніші моделі, здатні враховувати як числові, так і текстові параметри, що відкриває нові горизонти у застосуванні машинного навчання для аналізу складних часових рядів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] В. Б. Мокін, А. В. Лосенко, і А. Р. Ящолт, «Інформаційна технологія аналізу та прогнозування кількості нових випадків хвороби на коронавірус SARS-COV-2 в Україні на основі моделі Prophet», *Вісник Вінницького політехнічного інституту*, № 5, с. 71-83, 2020. https://doi.org/10.31649/1997-9266-2020-152-5-71-83 .

[2] В. Б. Мокін, А. В. Лосенко, і А. Р. Ящолт, «Інформаційна технологія аналізу та прогнозування багатохвильової кількості нових випадків захворювань на коронавірус COVID-19 на основі моделі Prophet», *Вісник Вінницького політехнічного інституту*, № 6, с. 65-75, 2020. https://doi.org/10.31649/1997-9266-2020-153-6-65-75 .

[3] В. Б. Мокін, М. В. Дратований, А. В. Лосенко, С. О. Жуков, «Прогнозування хвиль коронавірусу на основі відновленої когнітивної карти міжрегіонального впливу,» *Інформаційні технології та комп'ютерна інженерія*, т. 52, вип. 3, с. 86-94, 2021.

[4] A. Vartholomaios, S. Karlos, E. Kouloumpris, and G. Tsoumakas, "Short-term Renewable Energy Forecasting in Greece using Prophet Decomposition and Tree-based Ensembles," *arXiv*, Jul. 2021. [Electronic resource]. Available: https://arxiv.org/abs/2107.03825 . Accessed: 23 Nov. 2024.

[5] dos Santos Junior, J. C. Hu, R. Song, and Y. Bai, "Domain-Driven LLM Development: Insights into RAG and Fine-Tuning Practices," in Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August. 2024, pp. 6416-6417. https://doi.org/10.1145/3637528.3671445 .

[6] M. Arslan, S. Munawar, and C. Cruz, "Business insights using RAG–LLMs: a review and case study," *Journal of Decision Systems*, pp.1-30, 2024. https://doi.org/10.1080/12460125.2024.2410040 .

[7] Інститут проблем математичних машин і систем НАН України, *Звіти робочої групи з математичного моделю-*

вання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні, [Електронний ресурс]. Режим доступу: https://old.nas.gov.ua/UA//Activity/covid/Pages/wg.aspx . Дата звернення: 23 листопада. 2024.

[8] Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні, *Прогноз розвитку епідемії COVID-19 в Україні на 23 лютого – 8 березня 2022 року («Прогноз РГ-62»).* [Електронний ресурс]. Режим доступу: https://old.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=8716 . Дата звернення: 23 листопада. 2024.

[9] H. Tang, et al., "Time series forecasting with llms: Understanding and enhancing model capabilities," *arXiv*, 2024. [Electronic resource]. Available: https://arxiv.org/abs/2402.10835. Accessed: 23 листопад 2024.

[10] P. Cawood, and T. L. van Zyl, "Feature-weighted Stacking for Nonseasonal Time Series Forecasts: A Case Study of the COVID-19 Epidemic Curves," *arXiv*, Aug. 2021. [Electronic resource]. Available: https://arxiv.org/abs/2108.08723. Accessed: 23 Nov. 2024.

[11] B. VanBerlo, M. A. S. Ross, and D. Hsia, "Univariate Long-Term Municipal Water Demand Forecasting," *arXiv*, May 2021. [Electronic resource]. Available: https://arxiv.org/abs/2105.08486. Accessed: 23 Nov. 2024.

[12] J. Heaton, "An Empirical Analysis of Feature Engineering for Predictive Modeling," *arXiv*, Apr. 2019. [Electronic resource]. Available: https://arxiv.org/abs/1701.07852. Accessed: 23 Nov. 2024.

[13] B. S. Shaw, "False Prophet: Feature Engineering for a Homemade Time Series Regression," *Towards Data Science*, Dec. 2020. [Electronic resource]. Available: https://towardsdatascience.com/false-prophet-feature-engineering-for-a-homemade-time-series-regression-1b3f7a1b1c7e. Accessed: 23 Nov. 2024.

[14] H. Xue, and F. D. Salim, "Promptcast: A new prompt-based learning paradigm for time series forecasting," *IEEE Transactions on Knowledge and Data Engineering*, 2023. https://doi.org/10.1109/TKDE.2023.3342137 .

[15] B. S. Shaw, "Integrating Feature Engineering and Prophet for Enhanced Time Series Predictions," *Towards Data Science*, Nov. 2020. [Electronic resource]. Available: https://towardsdatascience.com/integrating-feature-engineering-and-prophet-for-enhanced-time-series-predictions-cfd62a5d6351. Accessed: 23 Nov. 2024.

Рекомендована кафедрою системного аналізу та інформаційних технологій ВНТУ

Стаття надійшла до редакції 14.12.2024

Лосенко Арсен Володимирович — д-р філософії, асистент кафедри системного аналізу та інформаційних технологій, e-mail: arsenlosenko@protonmail.com ;

Крижановський Євгеній Миколайович — канд. техн. наук, доцент, доцент кафедри системного аналізу та інформаційних технологій, e-mail: kruzhan@gmail.com ;

Штельмах Ігор Миколайович — канд. техн. наук, асистент кафедри системного аналізу та інформаційних технологій, e-mail: igor.shtelmakh@vntu.edu.ua ;

Варчук Ілона В'ячеславівна — канд. техн. наук, доцент кафедри системного аналізу та інформаційних технологій, e-mail: ilonavarchuk@gmail.com .

Вінницький національний технічний університет, Вінниця

A. V. Losenko[1]
Ye. M. Kryzhanovskyi[1]
I. M. Shtelmakh[1]
I. V. Varchuk[1]

LLM-based Feature Extraction Technology for Patient Testing from Textual Reports to Enhance Covid-19 Case Forecasting

[1]Vinnytsia National Technical University

The article focuses on the application of modern large language models (LLMs) to automate the extraction of essential features from analytical textual reports on the COVID-19 pandemic in Ukraine during 2020–2022. These reports encompass a broad spectrum of data, including regional morbidity indicators, testing dynamics, vaccination outcomes, and demographic characteristics of patients. The study explores the integration of these extracted features into time series models to improve the accuracy of epidemic forecasts.

Central to the research is the use of the Prophet model, which was enhanced to account for seasonal changes and anomalies in the data. The study addressed challenges such as the multi-wave nature of the COVID-19 time series, incorpo-

rating sharp increases and decreases in cases. Adjustments were made for anomalies caused by changes in quarantine measures, testing policies, and vaccination campaigns, particularly during winter surges.

Optimizing the Prophet model involved advanced parameter tuning using methods such as grid search and stochastic optimization, tailored to the specific epidemiological context in Ukraine. Additionally, the study evaluated the potential of neural network models, including LSTM (Long Short-Term Memory), to analyze time series data. LSTM's ability to capture nonlinear relationships and process multiple input variables complements traditional methods, providing deeper insights into long-term trends and interdependencies in the data.

The goal of this study is to develop an effective forecasting tool that integrates LLM-extracted features with advanced modeling techniques. By combining Prophet with enhancements and neural network approaches like LSTM, the research aims to significantly improve the accuracy of short- and long-term forecasts. This is particularly crucial for timely decision-making in public health during periods of epidemiological uncertainty.

Keywords: information technology, feature engineering, time series forecasting, Prophet, artificial intelligence, large language models, COVID-19.

***Losenko Arsen V.* —** PhD, Assistant of the Chair of System Analysis and Information Technologies, e-mail: arsenlosenko@protonmail.com ;

Kryzhanovskyi Yevhenii M. — Cand. Sc. (Eng.), Associate Professor of the Chair of System Analysis and Information Technologies, e-mail: kruzhan@gmail.com ;

Shtelmakh Ihor M. — Cand. Sc. (Eng.), Assistant of the Chair of System Analysis and Information Technologies, e-mail: igor.shtelmakh@vntu.edu.ua ;

Varchuk Ilona V. — Cand. Sc. (Eng.), Associate Professor of the Chair of System Analysis and Information Technologies, e-mail: ilonavarchuk@gmail.com