DOI:**10.33099/2311-7249/2024-50-2-111-117**

УДК: 004.8

Миколайчук Роман Антонович *(доктор технічних наук, доцент)* [1]
Миколайчук Аліса Іванівна *(кандидат філологічних наук, доцент)* [2]

[1] ***Національний університет оборони України, Київ, Україна***
[2] ***Київський національний університет імені Тараса Шевченка, Київ, Україна***

ВИКОРИСТАННЯ ТЕХНОЛОГІЙ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ АВТОМАТИЗАЦІЇ ПРОЦЕСУ ОБРОБКИ ДОКУМЕНТІВ

В умовах постійного зростання обсягу інформації та необхідності обробки великих текстових масивів, традиційні методи виявилися недостатньо ефективними та потребують значних людських ресурсів. Стаття присвячена розкриттю важливих аспектів використання сучасних технологій штучного інтелекту для оптимізації процесу створення звітів і надає практичні рекомендації для подальших досліджень у цій області. Метою статті є висвітлення підходу до використання великих мовних моделей, технології генерації розширеного пошуку та автономних агентів для автоматизації процесу обробки текстових даних на основі множини документів й оцінювання ефективності застосування таких технологій для забезпечення точності, релевантності та здатності до узагальнення інформації. Під час проведення дослідження було застосовано методи аналізу, синтезу, моделювання та експерименту. Зазначений методологічний підхід дав змогу розкрити підхід до використання технологій штучного інтелекту для автоматизації процесів розроблення і написання звітів на основі множини документів у захищених середовищах та провести експеримент щодо оцінювання ефективності використання розглянутих моделей і технологій. Запропоновано використовувати великі мовні моделі, що дають змогу перетворювати текст у векторні представлення для підвищення точності та релевантності згенерованих звітів, технологію генерації розширеного пошуку для поєднання пошуку релевантних документів та генерації тексту, що дало змогу значно покращити якість узагальнених звітів, автономних агентів для автоматизації збору, аналізу та обробки даних, що знижує потребу в постійному втручанні людини. Проведено експериментальні дослідження, результати яких свідчать, що використання великих мовних моделей спільно із технологією генерації розширеного пошуку та автономними агентами дає змогу значно покращити якість та ефективність автоматизації створення звітів. Проведене оцінювання за допомогою відомих метрик (BLEU, ROUGE та METEOR) підтвердили високу точність, релевантність та здатність до узагальнення згенерованих текстів. Зазначено переваги, недоліки запропонованого підходу, зокрема, підкреслюється важливість впровадження технологій штучного інтелекту в захищені середовища для забезпечення високого рівня безпеки даних. Теоретична значущість полягає у розширенні розуміння можливостей технологій штучного інтелекту під час обробки текстових даних та створення звітів. Практичною значущістю визначено потенціал застосування технологій штучного інтелекту у сфері безпеки і оборони та технічних галузях для підвищення ефективності моніторингу та управління. Науковою новизною дослідження є впровадження сучасних методів обробки природної мови та генерації тексту для створення звітів на основі множини документів.

Ключові слова: *штучний інтелект, автоматизація, обробка документів, обробка природної мови, великі мовні моделі, генерація розширеного пошуку, агенти штучного інтелекту.*

Вступ

У сучасному світі обсяги інформації зростають з неймовірною швидкістю, що ставить виклики стосовно її ефективної обробки та використання. Кожного дня створюються величезні масиви даних, що містять текстові документи, звіти, наукові статті, інформацію з соціальних медіа тощо. Цей інформаційний потік вимагає швидкої обробки, аналізу та узагальнення для забезпечення можливості приймати обґрунтовані рішення на основі актуальних даних.

Традиційні методи обробки текстової інформації, з використанням ручної праці, вже не відповідають сучасним вимогам щодо швидкості та точності виконання. Ручна обробка документів потребує значних людських ресурсів, часу та може призводити до помилок через людський фактор. В умовах, коли своєчасність та точність інформації є критичними, такі методи стають неефективними. Особливо актуальною ця проблема постає у сфері безпеки та оборони, де обробка великих обсягів інформації має бути не лише швидкою та точною, але й безпечною. Дані, що обробляються, часто є конфіденційними або мають обмежений доступ, що вимагає використання захищених середовищ для їх обробки. Будь-який виток інформації може мати серйозні наслідки для національної безпеки.

Використання технологій штучного інтелекту

(далі – ШІ) для автоматизації обробки текстових даних у сфері безпеки та оборони має враховувати необхідність роботи в захищених середовищах без доступу до мережі інтернет. Це дає змогу мінімізувати ризики витоку даних та забезпечити конфіденційність інформації. Відсутність доступу до інтернету під час обробки даних вимагає використання локальних рішень, що можуть працювати автономно та забезпечувати високу ефективність обробки.

Отже, в сучасних умовах стрімкого зростання обсягів інформації та підвищених вимог до її обробки, використання технологій штучного інтелекту стає необхідним для забезпечення ефективності та точності. Особливо важливим є застосування цих технологій у захищених середовищах без доступу до мережі інтернет, що забезпечує безпеку та конфіденційність оброблюваних даних у сфері безпеки та оборони.

Постановка проблеми. У контексті розвитку технологій штучного інтелекту та обробки природної мови (Natural Language Processing) (далі – NLP) [1], з'являються нові можливості для створення автоматизованих систем, здатних ефективно генерувати звіти на основі великого обсягу документів. Проте, такі системи стикаються з численними викликами, пов'язаними з потребою обробки великого обсягу різнорідної інформації, забезпеченням точності та релевантності згенерованих текстів.

Центральним аспектом цієї проблеми є розроблення моделей, що дають змогу автоматично обробляти текстові дані та створювати узагальнені звіти з множини документів. Така автоматизація вимагає від систем не тільки здатності до обробки великих обсягів даних, але й інтеграції інформації з різних джерел, забезпечення її точності та релевантності до запитів користувачів. Одним з можливих підходів для вирішення зазначеної проблеми є використання великих мовних моделей (Large Language Models) (далі – LLM), що побудовані на основі машинного навчання та обробки природної мови (NLP). Великі мовні моделі по суті є алгоритмом штучного інтелекту, що використовує методи глибокого навчання та величезні набори даних для розуміння, узагальнення, створення та прогнозування нового вмісту [2]. Наприклад моделі Llama3 та Mixtral здатні перетворювати текст у векторні представлення та використовувати їх для генерації нових текстів. Додатково, технологія генерації розширеного пошуку (Retrieval-Augmented Generation) (далі – RAG) дає змогу поєднувати пошук релевантних документів з генерацією тексту, що підвищує якість узагальнених звітів [3]. Водночас технологія штучного інтелекту може забезпечити автоматичний збір, аналіз та обробку даних, знижуючи потребу в постійному втручанні людини [4].

Отже, актуальність дослідження полягає у зростаючій потребі в автоматизації та підвищенні ефективності систем обробки текстових даних. Сучасні виклики, такі як необхідність швидкого реагування на непередбачувані зміни у середовищі та забезпечення конфіденційності даних, вимагають від систем бути більш гнучкими та інтелектуальними. Використання технологій LLM, RAG та агентів штучного інтелекту може забезпечити не лише автоматизацію, але й високу точність, релевантність і здатність до узагальнення інформації, що є важливим для ефективної обробки текстових даних у сфері безпеки та оборони.

Аналіз останніх досліджень і публікацій. Питанням використання технологій ШІ для обробки документів присвячена низка робіт. Так, у роботі [1] розглядаються методології та виклики, пов'язані з впровадженням чат-ботів, що використовують технології NLP. Огляд охоплює період з 1999 по 2022 рік і підкреслює важливість використання NLP для покращення взаємодії людина – комп'ютер. Результати досліджень свідчать, що чат-боти, які застосовують глибоке навчання, такі як моделі Seq2Seq (Sequence-to-Sequence) і BERT (Bidirectional Encoder Representations from Transformers), забезпечують більш природну та ефективну взаємодію з користувачами.

У статті [3] описується концепція технології RAG, що поєднує пошук інформації та генерацію тексту. Це дає змогу моделям LLM, бути більш ефективними у створенні узагальнених звітів з великих обсягів даних. Технологія RAG може значно підвищити точність та релевантність згенерованих текстів, що є критично важливим для автоматизації створення звітів.

У дослідженнях [4] проведено аналіз розвитку автономних агентів, які використовують можливості великих мовних моделей (LLM) для автоматизації складних завдань. Автономні агенти або агенти штучного інтелекту (artificial intelligence) (далі – AI-агенти), можуть планувати, виконувати та оптимізувати завдання, що дає змогу їм працювати незалежно та адаптуватися до змінних умов. Підкреслюється, що такі агенти можуть бути корисними для автоматизації робочих процесів у різних сферах, зокрема безпеки та оборони, де конфіденційність даних є критичною.

Незважаючи на значний прогрес у розробленні технологій обробки природної мови (NLP), великих мовних моделей (LLM) та AI-агентів, залишаються невирішені питання, що підкреслюють актуальність подальших досліджень стосовно: інтеграції технологій LLM, RAG та AI-агентів для виконання спільних завдань; підвищення ефективності AI-агентів під час обробки документів.

Особливу увагу слід приділити можливостям використання зазначених технологій у захищених середовищах без доступу до мережі інтернет, що є критично важливим для обробки конфіденційної інформації у сфері безпеки та оборони. Виконання зазначених завдань дасть змогу оцінити ефективність наведених технологій та визначити перспективи їх впровадження у сфері безпеки та оборони.

Метою статті є висвітлення підходу до використання великих мовних моделей, технології генерації розширеного пошуку та автономних агентів для автоматизації процесу обробки текстових даних на основі множини документів й оцінювання ефективності застосування таких технологій для забезпечення точності, релевантності та здатності до узагальнення інформації.

Виклад основного матеріалу дослідження

Великі мовні моделі (LLM) стали одними з найважливіших інструментів під час обробки природної мови (NLP), що дають змогу автоматично розуміти, аналізувати та генерувати текст на основі великої кількості вхідних даних. Моделі, такі як GPT-3, BERT, Llama3 та інші, базуються на архітектурі трансформерів, що забезпечує ефективну обробку контексту та взаємодії між словами в тексті. Використання таких моделей значно покращує точність та релевантність результатів у різних завданнях, від машинного перекладу до створення узагальнених звітів [5]. Великі мовні моделі (LLM) здатні навчатися на великих обсягах даних, що дає змогу їм адаптуватися до різних контекстів і забезпечувати високу якість текстів.

Розглянемо процес обробки тексту на основі трансформерів. Процес починається з перетворення тексту на послідовність токенів, що є основними одиницями обробки тексту і можуть бути окремими словами, частинами слів або навіть символами. Наприклад, речення «The quick brown fox» може бути перетворене на токени [«The», «quick», «brown», «fox»]. Кожен токен у послідовності перетворюється (трансформується) у векторне представлення. Це дає змогу моделі LLM обробляти текст у числовій формі, що є необхідним для подальшої обробки за допомогою алгоритмів машинного навчання. Векторні представлення токенів отримуються за відомими методиками мовного моделювання та навчання ознак в обробці природної мови (word embedding) з попередньо навчених ембедінгів, що зберігають семантичну інформацію про слова.

Оскільки трансформери не мають вбудованої інформації про порядок слів, до кожного векторного представлення токенів додається позиційний вектор. Цей вектор містить інформацію про позицію токена у реченні, що дає змогу моделі враховувати послідовність слів. Це досягається шляхом додавання позиційних ембедінгів до векторів токенів.

Для посилення деяких частин вхідних даних і применшення інших доцільно використовувати механізм уваги (Self-Attention), що дає змогу моделі LLM обчислювати вагові коефіцієнти для кожного токена щодо всіх інших токенів у реченні. Це враховує взаємозв'язки між словами та контекст, в якому вони вживаються. Кожен токен оцінюється з точки зору його важливості для інших токенів у реченні, що дає змогу моделі визначати, які частини тексту є найбільш релевантними для обробки в цей момент.

Вихідні дані з механізму уваги нормалізуються і до них додається початкове значення, що допомагає стабілізувати процес навчання та знижує ризик перенавчання. Цей процес забезпечує послідовність обробки даних і підвищує точність моделі.

Після обробки механізмом уваги дані проходять через двошаровий перцептрон (штучний нейрон), що обробляє кожен вектор незалежно від інших [6]. Цей етап включає нелінійні перетворення, що дають змогу моделі навчатися складнішим залежностям у даних.

Після проходження через декілька таких шарів, модель формує вихідне векторне представлення для кожного токена. Це представлення може бути використане для різних завдань NLP, зокрема класифікація тексту, генерація тексту, відповіді на запитання тощо. Кожен з цих етапів забезпечує ефективну обробку тексту та дає змогу моделі враховувати як локальний контекст кожного слова, так і глобальну структуру тексту.

Тому, архітектура трансформерів є ключовою складовою сучасних великих мовних моделей, забезпечуючи високу точність та гнучкість у генерації та обробці тексту.

Retrieval-Augmented Generation (RAG) – це сучасна технологія в обробці природної мови, що поєднує можливості пошуку інформації та генерації тексту та дає змогу створювати більш релевантні та точні тексти на основі великих обсягів даних, що особливо важливо для завдань автоматизації написання звітів [3].

Перший етап технології RAG містить пошук релевантних документів на основі вхідного запиту. Запит перетворюється на векторне представлення, що дає змогу обчислити його схожість з векторами документів у базі даних. Цей процес здійснюється за допомогою косинусної подібності (sim) [7], що визначається за виразом:

$$\text{sim}(q, d_i) = \frac{q \cdot d_i}{\|q\| \|d_i\|} \quad (1)$$

де q – вектор запиту,

d_i – вектор i-го документа.

Документи з найбільшою схожістю обираються для подальшої обробки.

На другому етапі RAG використовує знайдені документи для створення нового тексту, що відповідає на запит користувача або генерує звіт. Генеративна модель отримує як вхідний контекст обрані документи та запит і створює узагальнений текст.

Мовне моделювання використовує статистичні та ймовірнісні методи для визначення ймовірності появи певної послідовності слів у реченні. Тому ймовірність появи наступного слова у тексті обчислюється на основі попереднього контексту [8]:

$$P(w_{t+1} \mid w_1, w_2, \ldots, w_t, C), \quad (2)$$

де C – контекст, що містить знайдені документи;

w_t – коефіцієнти моделі;

t – кількість слів у реченні

Поєднання пошуку та генерації дає змогу створювати більш точні та релевантні тексти, оскільки модель використовує інформацію з багатьох джерел. Технологія RAG може бути використана у різних контекстах, від відповіді на запитання до автоматизації створення звітів. Використання попередньо натренованих векторних представлень дає змогу швидко знаходити релевантну інформацію та зменшити час обробки запитів.

У випадку, коли конфіденційність даних є критично важливою, технологія RAG може бути використана у захищених середовищах без доступу до мережі інтернет, на основі використання локально збереженої множини документів. Це дає змогу мінімізувати ризики витоку даних та підвищити рівень безпеки оброблюваної інформації.

Отже, RAG є потужною технологією, що поєднує пошук і генерацію тексту для створення високоякісних та релевантних текстових матеріалів. Завдяки своїм можливостям, зазначена технологія є особливо корисною для автоматизації процесів розроблення та написання звітів на основі множини документів у захищених середовищах, що робить її незамінною у сферах, де безпека даних є критично важливою.

Агенти штучного інтелекту є автономними програмними системами, які здатні виконувати завдання без постійного втручання людини. Вони можуть приймати рішення, виконувати певні дії, взаємодіяти з іншими агентами та користувачами, а також адаптувати свою поведінку на основі зворотного зв'язку. У контексті обробки текстових даних та автоматизації процесів написання звітів, AI-агенти стають важливими інструментами, які здатні підвищити ефективність і точність обробки інформації. Вони можуть автоматично збирати дані з різних джерел, таких як веб-сайти, бази даних та документів. AI-агенти здатні фільтрувати, аналізувати та структурувати інформацію для подальшого використання.

AI-агенти використовують методи обробки природної мови (NLP) для аналізу текстових даних, що містить класифікацію текстів, виявлення сутностей, аналіз тональності та інші завдання. Вони можуть створювати новий текст на основі наявних даних. Це може бути автоматичне написання звітів, відповідей на запитання, резюме документів тощо.

AI-агенти здатні автоматизувати різні завдання, такі як управління електронною поштою, планування зустрічей, обробка запитів від користувачів тощо. Вони можуть інтегруватися з іншими системами та інструментами для забезпечення безперервності робочих процесів.

Для забезпечення безпеку даних та мінімізації ризиків витоку інформації AI-агенти можуть працювати у захищених середовищах без доступу до мережі інтернет.

До переваг від використання AI-агентів:

Автономність. AI-агенти можуть працювати самостійно, виконуючи завдання без постійного втручання людини.

Ефективність. Вони здатні обробляти великі обсяги даних швидко та точно, що підвищує загальну ефективність робочих процесів.

Безпека. Можливість роботи у захищених середовищах без доступу до інтернету забезпечує високу безпеку оброблюваних даних.

Тому, AI-агенти є суттєвими інструментами для автоматизації обробки текстових даних і підвищення ефективності робочих процесів. Вони відіграють важливу роль у сфері безпеки та оборони, де конфіденційність даних і швидкість обробки інформації є критично важливими. Використання AI-агентів дає змогу значно підвищити точність та ефективність аналізу, забезпечуючи при цьому достатній рівень безпеки даних.

Наведені технології надають потужні інструменти для автоматизації обробки текстових даних та створення звітів. Однак, для оцінювання їхньої ефективності та виявлення практичних переваг і недоліків, потрібно провести низку експериментальних досліджень. Далі наведено методологію та результати експериментів, що демонструють, як ці технології можуть бути застосовані на практиці для написання звітів на основі множини документів.

Для дослідження були використані оперативні зведення, опубліковані на офіційному Telegram-каналі Генерального штабу Збройних Сил України. Ці зведення містять детальну інформацію про бойові дії, втрати противника та стан оперативної обстановки. Щоденні звіти збиралися та аналізувалися для створення тижневих узагальнених звітів за різними тематичними напрямами.

На першому етапі використовувалася технологія RAG для пошуку релевантних частин документів. Запити були перетворені на векторні представлення, що дало змогу обчислити схожість між запитами та документами. На другому етапі зібрані документи використовувалися для генерації узагальнених звітів. Кожен текст був перетворений на векторне представлення. Використовуючи такі векторні представлення, генеративні моделі створювали нові тексти, що відповідали на запит користувача або узагальнювали інформацію з різних документів. Звіти генерувались за допомогою різних LLM, характеристика яких наведена у таблиці 1.

Для оцінювання ефективності використання LLM для автоматизації написання звітів важливо застосовувати чіткі та об'єктивні метрики. Нижче наведені ключові метрики, що використовувались під час експериментальних досліджень.

Таблиця 1

Основні характеристики великих мовних моделей

Модель	Розробник	Кількість параметрів (млрд. шт.)	Кількість вхідних токенів (шт.)
Llama3 8b	Meta	8	8192
Llama3 70b	Meta	70	8192
Mixtral 8x7b	Mistral	8x7	32768
Gemma 7b	Google	7	8192

Точність визначає, наскільки правильно модель передає фактичну інформацію з оригінальних документів у згенерованих звітах. Для оцінювання точності використано метрику *BLEU* (Bilingual Evaluation Understudy – двомовне навчання з оцінювання), що вимірює відповідність між згенерованим текстом та еталонними текстами, що оцінюють точність переданих фактів [9]:

$$BLEU = BP \cdot \exp(\sum_{n=1}^{N} w_n \log p_n), \quad (3)$$

де BP – пеналізація за довжину;

w_n – ваги *N*-грам;

p_n – точність *N*-грам;

N-грама – безперервна послідовність із *n* елементів із заданого зразка тексту [10].

Релевантність визначає, наскільки добре зміст згенерованих звітів відповідає запиту або темі, для якої вони були створені. Для оцінювання кількості збігів між згенерованими звітами та еталонними текстами використано метрику *ROUGE* (Recall-Oriented Understudy for Gisting Evaluation – орієнтоване на запам'ятовування дослідження для оцінки суті) [11]:

$$ROUGE = \frac{\sum_{S \in R} \sum_{g \in S} \min\left(C(g), C_{g \in e}(g)\right)}{\sum_{S \in R} \sum_{g \in S} C(g)}, \quad (4)$$

де, R – еталонний текст;

S – набір уніграм в еталонному тексті;

g – уніграма (елемент);

$C(g)$ – кількість поява уніграми g в еталонному тексті;

$C_{gen}(g)$ – кількість поява уніграми g в згенерованому тексті.

Узагальнення визначає, наскільки добре система здатна узагальнювати інформацію з різних джерел, створюючи зв'язний та інформативний звіт. Використано метрику *METEOR* (Metric for Evaluation of Translation with Explicit Ordering), що оцінює узагальнення шляхом вимірювання семантичної схожості між згенерованим текстом та еталонними текстами (зважене середнє значення точності та повноти з урахуванням синонімів і перестановок) [12]:

$$METEOR = \frac{10AF}{9A+F}, \quad (5)$$

де, A – точність першої уніграми, що обчислюється як відношення кількості уніграм у згенерованому тексті до загальної кількості уніграм в еталонному тексті;

F – повнота.

Результати експериментальних досліджень щодо узагальнення інформації з різних джерел наведено у таблиці 2.

Таблиця 2

Результати узагальнення інформації з великих мовних моделей

Модель	Точність	Релевантність	Узагальнення
Llama3 8b	82.3	79.5	81.2
Llama3 70b	89.1	86.7	88.5
Mixtral 8x7b	84.7	77.8	83.4
Gemma 7b	80.2	77.8	79.1

Відповідно до даних, наведених у таблиці 2, слід зазначити таке. Модель Llama3 8b демонструє високу точність та здатність до узагальнення інформації, проте поступається моделям з більшим об'ємом параметрів у деталізації та релевантності. Хоча модель Gemma 7b менш потужна порівняно з іншими, водночас вона демонструє прийнятну точність і здатність до узагальнення інформації для базових задач. Модель Mixtral 8x7b добре справляється із задачами, що вимагають великого контексту, демонструючи високі результати у всіх метриках. Найкраща серед розглянутих моделей – Llama3 70b. Висока точність та релевантність згенерованих текстів робить її ідеальною для складних задач із великим обсягом контексту.

Отже, перехід від опису технологій до експериментальних досліджень дає змогу оцінити практичну ефективність LLM та RAG для автоматизації створення звітів на основі множини документів. Результати експериментів свідчать, що використання технологій штучного інтелекту значно покращує якість та точність узагальнених звітів, забезпечуючи високий рівень релевантності та здатності до узагальнення інформації.

Висновки й перспективи подальших досліджень

На основі проведених досліджень можна зробити висновок, що використання технологій штучного інтелекту, таких як великі мовні моделі, генерації розширеного пошуку та автономні агенти доцільно використовувати для автоматизації створення тематичних звітів на основі множини документів. Експериментальне моделювання процесу генерації текстів за допомогою великих мовних моделей (LLM) та технології генерації розширеного пошуку (RAG) показало значний потенціал у покращенні точності, релевантності та узагальнення текстів. Результати демонструють, що такі системи можуть ефективно використовувати інформацію з різних джерел для створення якісних та інформативних звітів.

Наукова новизна викладеного у статті зводиться до впровадження передових методів великих

мовних моделей та технологій генерації розширеного пошуку для автоматизації процесу обробки текстових даних і створення звітів у захищених середовищах без доступу до інтернету. Теоретичною значущістю є розширення розуміння можливостей великих мовних моделей та технології генерації розширеного пошуку в контексті обробки природної мови, а практичною значущістю – потенціал застосування розроблених методів у сферах безпеки та оборони для підвищення ефективності управління і моніторингу інформації.

Практичні рекомендації зосереджуються на впровадженні моделей Llama3 та Mixtral для завдань генерації звітів, що забезпечить високу точність та релевантність тексту. Використання локальних моделей дає змогу знизити ризики витоку даних і забезпечити високий рівень безпеки оброблюваної інформації. Інтеграція технології генерації розширеного пошуку до систем обробки даних покращить якість та ефективність створення звітів на основі множини документів. Впровадження автономних агентів для автоматизації збору та аналізу даних дає змогу знизити навантаження на людину та підвищити ефективність процесу обробки інформації й розроблення документів. Крім того, використання агентів у захищених середовищах забезпечить безпеку даних.

Подальші дослідження можуть бути спрямовані на розроблення та впровадження більш складних великих мовних моделей з покращеними алгоритмами обробки контексту для підвищення точності та релевантності текстів. Важливим напрямом є також удосконалення технології генерації розширеного пошуку для інтеграції з більшою кількістю джерел даних та покращення алгоритмів пошуку. Розроблення нових агентів штучного інтелекту з покращеною автономністю та здатністю до адаптації на основі нових даних дасть змогу підвищити ефективність автоматизації обробки текстових даних.

Список бібліографічних посилань

1. Lin C. C., Huang A. Y. Q., Yang S. J. H. A Review of AI-Driven Conversational Chatbots Implementation Methodologies and Challenges (1999–2022). Sustainability 2023, 15(5). DOI: https://doi.org/10.3390/su15054012. **2. Wyndham A**. 10 Large Language Models That Matter to the Language Industry. Data & Indexes, Technology. 2024. URL: https://slator.com/10-large-language-models-that-matter-to-the-language-industry (Accessed: 29 July 2024). **3. Zhao P., Zhang H., Yu Q., Wang Z., Geng Y., Fu F., Yang L., Zhang W., Jiang J., Cui B.** Retrieval-Augmented Generation for AI-Generated Content: A Surve. arXiv: 2402.19473 (cs.CV) 2024. URL: https://arxiv.org/abs/2402.19473 (Accessed: 29 July 2024). **4. Sanjay P., Li W., Xiaoming Ch., Hua L.** GPT was Only the Beginning: Autonomous Agents are Coming. Boston Consulting Group. 2023. [онлайн] URL: https://www.bcg.com/publications/2023/gpt-was-only-the-beginning-autonomous-agents-are-coming (Accessed: 29 July 2024). **5. Bisen Sh. K.** Large Language Models (LLM): Difference between GPT-3 & BERT. Published in Bright AI. 2022. URL: https://medium.com/bright-ai/nlp-deep-learning-models-difference-between-bert-gpt-3-f273e67597d7 (Accessed: 29 July 2024). **6. Гороховатський О. В., Передрій О. О.** Багатошаровий персептрон як інструмент первинної кластеризації зображень. *Реєстрація, зберігання і обробка даних*. 2016. Т. 18, № 4. С. 33–43. URL: http://dspace.nbuv.gov.ua/handle/ 123456789/131626 (дата звернення: 29.07.2024). **7. Singhal A.** Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.* 2001. № 24 (4). P. 35–43. URL: http://singhal.info/ieee2001.pdf. (дата звернення: 30 July 2024). **8. Гулівата І. О., Гусак Л. П., Радзіховська Л .М.** Вища та прикладна математика: Теорія Ймовірностей: навчальний посібник. Вінниця: Видавничо-редакційний відділ ВТЕІ КНТЕУ, 2018. 208 с. URL: https://ir.vtei.edu.ua/g.php?fname=25604.pdf (дата звернення: 30.07.2024). **9. Papineni K., Roukos S., Ward T., Zhu W. J.** BLEU: a method for automatic evaluation of machine translation (PDF). ACL-2002: *40th Annual meeting of the Association for Computational Linguistics.* 2002. P. 311–318. URL: https://aclanthology.org/P02-1040.pdf (дата звернення: 30.07.2024). **10. Lin Ch.-Y., Hovy E.** Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In: *Proceedings of 2003 Language Technology Conference* (HLT-NAACL 2003), Edmonton, Canada, May 27 –June 1, 2003. 71–78 URL: https://aclanthology.org/N03-1020.pdf (Accessed: 30 July 2024). **11. Lin Ch.-Y.** ROUGE: A package for automatic evaluation of summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25–26, 2004. URL: https://aclanthology.org/W04-1013.pdf (Accessed: 30 July 2024). **12. Banerjee S., Lavie A.** METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, June 2005. URL: https://aclanthology.org/W05-0909.pdf (дат Accessed: 30 July 2024).

USING ARTIFICIAL INTELLIGENCE TECHNOLOGIES FOR DOCUMENT PROCESSING AUTOMATION

***Mykolaichuk Roman** (Doctor of Technical Sciences, Associate Professor)* [1]
***Mykolaichuk Alisa** (Candidate of Philological Science, Associate Professor)* [2]

[1] *The National Defence University of Ukraine, Kyiv, Ukraine*
[2] *Taras Shevchenko National University of Kyiv, Kyiv, Ukraine*

Formulation of the problem in general. *Given the constant growth in the amount of information and the need to process large text arrays, traditional methods have proven to be insufficiently effective and require significant human resources. The article is devoted to the disclosure of important aspects of using modern artificial intelligence technologies to optimize the process of creating reports and provides practical recommendations for further research*

in this area. The purpose of the article is to highlight an approach to using large language models, advanced search generation technology and autonomous agents to automate the process of processing textual data based on a set of documents and to assess the effectiveness of using such technologies to ensure accuracy, relevance and ability to generalize information. The methods of analysis, synthesis, modeling, and experimentation were used in the course of the study. This methodological approach makes it possible to reveal the approach to using artificial intelligence technologies to automate the processes of developing and writing reports based on a set of documents in secure environments and to conduct an experiment to evaluate the effectiveness of using the models and technologies under consideration.

Analysis of recent researches and publications *A number of studies have examined the methodologies and challenges associated with the implementation of artificial intelligence technologies for information processing for text generation. Despite significant progress in the development of natural language processing technologies, large language models, and autonomous agents, there are still unresolved issues that emphasize the relevance of research on the integration of technologies to perform common tasks.*

Presenting the main material *It is proposed to use large language models that allow converting text into vector representations to improve the accuracy and relevance of generated reports, advanced search generation technology to combine search for relevant documents and text generation, which significantly improves the quality of summarized reports, and autonomous agents to automate data collection, analysis and processing, which reduces the need for constant human intervention. Experimental studies have been conducted, the results of which show that the use of large language models in conjunction with advanced search generation technology and autonomous agents can significantly improve the quality and efficiency of automated report generation. Evaluation using well-known metrics (BLEU, ROUGE, and METEOR) has confirmed the high accuracy, relevance, and generalization ability of the generated texts. The advantages and disadvantages of the proposed approach are noted, in particular, the importance of implementing artificial intelligence technologies in secure environments to ensure a high level of data security is emphasized.*

Elements of scientific novelty. *The scientific novelty of the study is the introduction of modern methods of natural language processing using artificial intelligence and text generation technologies to create reports based on a set of documents.*

Practical significance of the article *The practical significance lies in the potential of applying artificial intelligence technologies in the security and defense sector and technical industries to improve the efficiency of monitoring and management.*

Conclusion and the perspectives of future researches. *It is advisable to use artificial intelligence technologies to automate the creation of thematic reports based on a set of documents. Experimental modeling of the text generation process with the help of large language models and advanced search generation technology has shown significant potential for improving the accuracy, relevance, and generalization of texts. Further research could focus on developing and implementing more sophisticated models with improved context processing algorithms and improving the technology for generating advanced searches to integrate with more data sources.*

Keywords*: artificial intelligence, automation, document processing, natural language processing, large language models, advanced search generation, artificial intelligence agents.*

References

1. Lin, C. C., Huang, A. Y. Q., Yang, S. J. H., (2023). A review of AI-driven conversational chatbots implementation methodologies and challenges (1999–2022). Sustainability, 15(5). DOI: https://doi.org/10.3390/ su15054012. **2. Wyndham, A.,** (2024). 10 Large Language Models That Matter to the Language Industry. Data & Indexes, Technology [online]. Avaliable at: https://slator.com/10-large-language-models-that-matter-to-the-language-industry [Accessed: 29 July 2024]. **3. Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J., Cui, B.,** (2024). Retrieval-Augmented Generation for AI-Generated Content: A Surve. *Arxiv* [online]. Avaliable at: https://arxiv.org/abs/2402.19473 [Accessed: 29 July 2024]. **4. Sanjay, P., Li, W., Xiaoming, Ch., Hua, L.,** (2023). GPT was Only the Beginning: Autonomous Agents are Coming. Boston Consulting Group [online]. Avaliable at: https://www.bcg.com/publications/2023/gpt-was-only-the-beginning-autonomous-agents-are-coming [Accessed: 29 July 2024]. **5. Bisen, Sh. K.,** (2022). Large Language Models (LLM): Difference between GPT-3 & BERT. Published in Bright AI [online]. Avaliable at: https://medium.com/bright-ai/nlp-deep-learning-models-difference-between-bert-gpt-3-f273e67597d7 [Accessed: 29 July 2024]. **6. Horokhovat·s′kyy, O., Peredriy, O.,** (2016). Multilayer perceptron as a tool for primary image clustering. *Reyestratsiya, zberihannya i obrobka danykh*, 18(4), 33–43 [online]. Avaliable at: http://dspace.nbuv.gov.ua/handle/123456789/131626 [Accessed: 29 July 2024]. **7. Singhal, A.,** (2001). Modern Information Retrieval: A Brief Overview [online]. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering,* 24 (4), 35–43. Avaliable at: http://singhal.info/ieee2001.pdf [Accessed: 29 July 2024]. **8. Hulivata, I., Husak, L., Radzikhovs′ka, L.,** (2018). Higher and applied mathematics: Probability theory. Navchal′nyy posibnyk. Vinnytsya: Vydavnycho-redaktsiynyy viddil VTEI KNTEU, 208 [online]. Avaliable at: https://ir.vtei.edu.ua/g.php?fname=25604.pdf [Accessed: 30 July 2024]. **9. Papineni, K., Roukos, S., Ward, T., Zhu, W. J.,** (2002). BLEU: a method for automatic evaluation of machine translation (PDF). ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, 311–318 [online]. Avaliable at: https://aclanthology.org/P02-1040.pdf [Accessed: 30 July 2024]. **10. Lin, Ch.-Y., Hovy, E.,** (2003). Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In: *Proceedings of 2003 Language Technology Conference* (HLT-NAACL 2003), Edmonton, Canada, May 27 –June 1, 2003. 71–78 [online]. Avaliable at: https://aclanthology.org/N03-1020.pdf [Accessed: 30 July 2024]. **11. Lin, Ch.-Y.,** (2004). ROUGE: A package for automatic evaluation of summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25–26 [online]. Avaliable at: https://aclanthology.org/W04-1013.pdf [Accessed: 30 July 2024]. **12. Banerjee, S., Lavie, A.,** (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, June [online]. Avaliable at: https://aclanthology.org/W05-0909.pdf [Accessed: 30 July 2024].