УДК 004.8:004.738.5 **doi: 10.26906/SUNZ.2025.3.87-95**

В. В. Заіка, О. В. Чуєв, О. І. Морозова, Т. С. Нікітіна

Національний аерокосмічний університет «Харківський авіаційний інститут», Харків

ІНТЕЛЕКТУАЛЬНІ ВЕБСИСТЕМИ ДЛЯ АВТОМАТИЗАЦІЇ ОБРОБКИ ТА МОНІТОРИНГУ ІНФОРМАЦІЙНИХ ПОТОКІВ

Анотація. Актуальність. У сучасних умовах різкого зростання обсягів цифрових даних та ускладнення мережевої інфраструктури зростає потреба в інтелектуальних рішеннях для обробки та моніторингу інформаційних потоків. Традиційні засоби стають недостатньо ефективними для забезпечення швидкого реагування на інциденти та витягнення значущої інформації з неструктурованих джерел. **Об'єкт дослідження:** інтелектуальні вебсистеми для автоматизації обробки текстових даних і моніторингу серверної інфраструктури. **Мета статті:** аналіз сучасних підходів до побудови інтелектуальних вебсистем, що поєднують технології розпізнавання тексту та автоматичного моніторингу інфраструктури, а також розробка концептуальної моделі їх інтеграції. **Результати дослідження.** У статті представлено теоретичні основи інтелектуальних вебсистем, проаналізовано сучасні методи розпізнавання тексту на основі глибинного навчання, а також засоби моніторингу інфраструктури, зокрема системи збору метрик, логів і виявлення аномалій. Запропоновано концептуальну модель інтеграції таких компонентів у єдину вебсистему. **Висновки.** Інтелектуальні вебсистеми демонструють високий потенціал для автоматизації ключових процесів аналізу даних та контролю IT-інфраструктури. Їх впровадження сприяє підвищенню ефективності, швидкості реагування та зниженню ризиків, пов'язаних з обробкою великих обсягів інформації. **Сфера використання отриманих результатів**: вебзастосунки для обробки документів, платформи моніторингу серверів і хмарних сервісів, системи підтримки прийняття рішень у бізнесі та кібербезпеці.

Ключові слова: інтелектуальні вебсистеми, автоматизація, оптичне розпізнавання символів (OCR), моніторинг інфраструктури, машинне навчання, обробка природної мови (NLP), SIEM, Semantic Web.

Вступ

У сучасному цифровому середовищі інформаційні потоки та обсяги даних постійно збільшуються, що створює необхідність автоматизованої обробки для оперативного отримання знань та прийняття ефективних рішень. Інтелектуальні вебсистеми поєднують традиційну вебархітектуру з передовими методами штучного інтелекту (ШІ) та машинного навчання (МН), що дозволяє здійснювати більш гнучкий аналіз даних і забезпечувати адаптивну взаємодію з користувачами. Завдяки цим технологіям як приватні особи, так і організації можуть швидше знаходити необхідну інформацію, оптимізувати свої процеси та приймати раціональні рішення.

Концепція Web Intelligence передбачає трансформацію Інтернету у "Мережу мудрості" (The Wisdom Web), де розумні алгоритми аналізують потоки даних з різних джерел, включаючи соціальні мережі [1]. Це відкриває можливість створення самонавчальних систем, здатних генерувати персоналізовані рекомендації та прогнозувати події на основі аналізу великих масивів інформації, але також потребує відповідної системи, що забезпечить швидку, якісну і стійку роботу таких процесів. Значний розвиток отримали технології автоматизації обробки текстової інформації, зокрема оптичне розпізнавання символів (OCR) і обробка природної мови (NLP). Ці методи дають змогу перетворювати неструктуровані тексти у формат, доступний для машинного аналізу. Попит на OCR нового покоління, що використовує технології нейронних мереж, демонструє стрімке зростання, і очікується, що його розвиток продовжиться протягом наступних десятиліть [2]. Одночасно, прогрес у сфері NLP сприяв появі генеративного ШІ, що охоплює здатність великих мовних моделей (LLM) вести діалог та моделей ге-нерації зображень – інтерпретувати запити користувачів [3]. Це свідчить про актуальність цих технологій у глобальному інформаційному просторі.

Важливим аспектом є моніторинг серверної інфраструктури, який забезпечує стабільність роботи вебдодатків завдяки оперативному виявленню відмов та аномалій. Надійний моніторинг критично важливий, оскільки затримки у роботі додатків або зниження якості обслуговування можуть призвести до відтоку користувачів, фінансових втрат та погіршення репутації компаній [4]. Таким чином, інтеграція передових методів текстової аналітики, прогнозування подій та моніторингу інфраструктури створює комплексний підхід до управління інформаційними потоками.

Метою даної статті є систематизація сучасних методів та підходів у сфері інтелектуальних вебсистем, аналіз алгоритмів розпізнавання тексту та моніторингу серверів, а також представлення концептуальної моделі їх інтеграції. У дослідженні розглядаються ключові аспекти, включаючи теоретичні основи інтелектуальних вебтехнологій, огляд методів автоматизованої обробки інформації, аналіз проблем та обмежень, що постають перед цими технологіями, формування висновків та визначення перспектив подальших наукових розробок.

Теоретичні основи інтелектуальних вебсистем

Інтелектуальна вебсистема – це сучасний вебдодаток або комплекс сервісів, який інтегрує передові методи штучного інтелекту (ШІ), включаючи експертні системи, нейронні мережі, машинне навчання та онтології, для автоматичної обробки, аналізу та інтерпретації великих обсягів даних [5]. Такі системи створюють умови для швидкої та ефективної взаємодії користувачів з інформацією, дозволяючи отримувати релевантні результати та підтримувати прийняття обґрунтованих рішень.

Основними компонентами інтелектуальної вебсистеми є модулі збору та фільтрації даних, які забезпечують накопичення інформації з різних джерел, включаючи бази знань, сенсорні пристрої, соціальні мережі та вебресурси [6]. Крім того, аналітичні алгоритми, що включають методи прогнозування, сегментації та розпізнавання шаблонів, дозволяють здійснювати поглиблений аналіз отриманих даних. Важливу роль відіграють інтерфейси взаємодії з користувачем, які забезпечують зручний доступ до обробленої інформації, адаптивний контент та персоналізовані рекомендації. Ключовими особливостями інтелектуальних вебсистем є здатність до навчання на основі отриманих даних, що дозволяє удосконалювати алгоритми обробки та підвищувати точність прогнозів [6]. Важливою характеристикою також є самоадаптація – можливість змінювати моделі аналізу відповідно до зміни умов або нових викликів. Крім того, такі системи здатні працювати з неоднорідними джерелами інформації, включаючи структуровані та неструктуровані дані, що забезпечує їхню універсальність і гнучкість.

З погляду архітектури, інтелектуальні вебсистеми зазвичай реалізуються за принципом розподілених або хмарних сервісів, що забезпечує їхню масштабованість, гнучкість та надійність. Найчастіше використовується мікросервісна або сервіс-орієнтована архітектура (SOA), яка дозволяє ефективну взаємодію між компонентами за допомогою стандартних протоколів REST/HTTP. Для зберігання великих обсягів даних використовуються різні типи баз:

- SQL-бази – для структурованих даних;
- NoSQL-бази – для гнучкого управління нереляційними та великими масивами даних;
- графові бази даних – для роботи з онтологіями та знаннями.

Обробка потокових даних здійснюється за допомогою таких технологій, як Kafka, Flink, Spark Streaming, а для аналізу неструктурованої інформації застосовуються платформи Big Data, включаючи Hadoop та Elasticsearch. Інформаційний потік у таких системах може включати текстову інформацію, таку як документи, новини або аналітичні звіти, та телеметричні дані, як метрики серверної інфраструктури чи логи роботи додатків.

Принципово важливим є поділ потоків даних на семантичний рівень (аналіз змісту) та технічний рівень (моніторинг серверної інфраструктури). Семантична складова потребує застосування методів NLP, таких як морфологічний аналіз, тематичне моделювання та розпізнавання сутностей. Технології Semantic Web, такі як RDF, OWL та SPARQL, та графи знань (knowledge graph) допомагають формалізувати знання, що дозволяє системі не лише аналізувати дані, а й генерувати рекомендації або автоматизовані висновки. Однією з ключових переваг інтелектуальних вебсистем є їхня адаптивність та здатність до самонавчання. Це досягається завдяки використанню сучасних моделей машинного навчання, зокрема:

- нейронних мереж – для розпізнавання складних закономірностей;
- методів ансамблів – для підвищення точності прогнозів;
- методу опорних векторів (SVM) – для класифікації та моделювання даних.

Системи інтелектуального моніторингу інфраструктури часто застосовують аналітику потокових метрик, що дозволяє заздалегідь виявляти аномалії та прогнозувати навантаження на сервери [7]. Використання "мікросервісної" моделі гарантує високу гнучкість: кожен окремий сервіс, наприклад, OCR-аналізатор, система зберігання метрик або модуль семантичного аналізу, може розвиватися незалежно.

Підсумовуючи, інтелектуальні вебсистеми поєднують класичні принципи веброзробки (клієнт-серверна архітектура, API, протоколи) з сучасними методами обробки даних та алгоритмами ШІ. Це створює універсальну платформу, здатну аналізувати великі масиви неоднорідної інформації в реальному часі, ефективно прогнозувати зміни в цифровому середовищі та автоматично адаптуватися до нових умов.

Огляд сучасних підходів до розпізнавання тексту

Розпізнавання тексту охоплює як обробку документів (сканованих зображень з друкованим чи рукописним текстом), так і виявлення текстових фрагментів на зображеннях (наприклад, вуличні знаки). Традиційні методи OCR спираються на попередню обробку зображень (бінаризація, шумозаглушення, вичленення контурів символів), після чого застосовують шаблонне чи контурне розпізнавання. Раніше широко використовувалися евристичні алгоритми та статистичні моделі, наприклад методи на основі прихованих марковських моделей (HMM) для моделювання послідовностей символів [8]. Однак такі підходи часто потребували ручної адаптації під специфічні шрифти чи структури документів.

Серйозний прорив у розпізнаванні тексту принесли методи глибокого навчання. Згорткові нейронні мережі (CNN) ефективні у витяганні ознак із зображень тексту, а рекурентні мережі (RNN, LSTM) враховують послідовний характер символів. Згортково-рекурентні моделі, так звані CRNN, комбінують ці підходи для побудови end-to-end системи розпізнавання [9]. Варто зауважити, що сучасні системи розпізнавання зазвичай доповнюють мовними моделями чи словниками для корекції помилок і підвищення точності. Також активно розвивається оптичне розпізнавання рукописного тексту (HTR). Міжнародні дослідження вказують, що основною складністю HTR є велика варіабельність почерку (різні стилі, розмір букв, похил тощо). Саме тому сучасні HTR-системи базуються на поєднанні CNN та RNN, часто з попереднім навчанням на великих корпусах рукописів, а потім донавчанням для конкретних мов або наборів символів.

У 2023–2024 роках з'явились системи на основі архітектури Transformer, які змінили підходи до розпізнавання тексту. Наприклад, модель TrOCR використовує трансформерні мережі як для обробки зображення документа, так і для генерації тексту, демонструючи вищу точність порівняно з попередніми CNN–RNN підходами. Відзначається, що "існуючі підходи для розпізнавання зазвичай базуються на CNN для розуміння зображень і RNN для генерації тексту", тоді як нові

рішення з тренованими трансформерними моделями перевершують попередні результати як для друкованого, так і для рукописного тексту [10]. Варто відзначити, що за такі якісні результати роботи доводиться розплачуватися високою обчислювальною складністю та потребою у великих моделях та GPU.

Ще однією ключовою віхою стала поява мовної моделі BERT (Bidirectional Encoder Representations from Transformers), запропонованої дослідниками Google [11]. Хоча BERT першочергово призначена для обробки тексту, її принципи трансформерної архітектури лягли в основу більшості сучасних моделей, зокрема і в OCR-системах. BERT дозволяє враховувати контекст слова з обох боків речення (ліворуч і праворуч), що забезпечує значно глибше розуміння семантики порівняно з попередніми підходами. У суміжних задачах, як-от доопрацювання розпізнаного тексту, класифікація документів або витяг інформації, інтеграція BERT-моделей показує значне покращення якості результатів. Крім того, ідеї, закладені в BERT, використовуються в мультимодальних рішеннях, які об'єднують зображення і текст для комплексної інтерпретації даних. Таким чином, сучасні підходи до OCR можна згрупувати так:

- класичні алгоритми (фільтрація, сегментація, шаблонне розпізнавання, HMM) для чітко структурованих документів;
- методи нейронних мереж (CNN, RNN/LSTM, CRNN) з попередньою обробкою зображень для різноякісних документів і форм;
- трансформери і великі мовні моделі (напр., TrOCR, BERT) для автоматичного виділення ознак і прогнозування символів на рівні слова або рядка;
- комерційні та open-source системи (Tesseract з LSTM-модулями, ABBYY FineReader, Google Cloud Vision, Microsoft Cognitive Services), які реалізують багато з перерахованих методів у готових API та інструментах.

Серед мовних і бібліотечних платформ для розпізнавання тексту варто зазначити відкрите програмне забезпечення (Tesseract OCR, Kraken) і хмарні сервіси (Google Vision API, Azure OCR). Багато з них використовують навчання на великих синтетичних та реальних наборах даних, що дозволяє підвищувати якість розпізнавання. Незважаючи на прогрес, залишаються важливі завдання: покращення точності розпізнавання рукопису, багатомовної підтримки та обробки складних макетів документів. Дослідники також працюють над розширенням систем для безперервних потоків тексту (наприклад, розпізнавання тексту у відео чи живих трансляціях).

Огляд сучасних підходів до моніторингу серверної інфраструктури

Моніторинг серверної інфраструктури є критично важливим елементом забезпечення безпеки, стабільності та ефективності критичних систем, таких як дата-центри, телекомунікаційні мережі та хмарні платформи. Сучасні підходи до моніторингу поєднують традиційні методи аналізу даних із передовими технологіями штучного інтелекту (ШІ) та обробки великих даних, що дозволяє своєчасно виявляти інциденти, прогнозувати потенційні загрози та оптимізувати роботу систем. Цей розділ аналізує ключові методи, інструменти та виклики, пов’язані з моніторингом серверної інфраструктури, з акцентом на їхню роль у створенні інтелектуальних вебсистем.

Аналіз серверних логів є основою моніторингу критичних систем, оскільки логи містять детальну інформацію про події, такі як системні помилки, спроби несанкціонованого доступу або пікові навантаження. Сучасні методи аналізу логів можна поділити на три основні категорії. По-перше, rule-based системи використовують заздалегідь визначені правила для ідентифікації відомих шаблонів, таких як повторювані помилки чи сигнатури кібератак. Ці системи прості у впровадженні, швидкі для обробки стандартних сценаріїв і широко застосовуються в інструментах, таких як Zabbix. Проте їхня ефективність обмежена при зіткненні з новими або нестандартними загрозами, оскільки вони не здатні адаптуватися до змінних умов без оновлення правил. Моделі машинного навчання (ML), такі як Random Forest, Support Vector Machines (SVM) або рекурентні нейронні мережі (LSTM), пропонують більш гнучкий підхід. Ці моделі навчаються на історичних даних, що дозволяє їм виявляти аномалії, які не відповідають нормальній поведінці системи, або прогнозувати майбутні інциденти. Наприклад, LSTM ефективно аналізує часові ряди логів, що допомагає виявляти послідовності подій, характерні для атак типу "розподілений відмова в обслуговуванні" (DDoS) [12-13].

Однак ML-моделі потребують значних обчислювальних ресурсів і якісних даних для навчання, що може ускладнити їхнє впровадження в системах із обмеженими ресурсами. Методи виявлення аномалій зосереджені на ідентифікації відхилень від нормальної поведінки системи. Наприклад, алгоритми кластеризації можуть групувати нормальні події та виділяти аномальні, такі як незвично високе навантаження на сервер. Інструменти, такі як Splunk і ELK Stack (Elasticsearch, Logstash, Kibana), забезпечують комплексну обробку логів, включаючи їх агрегацію, візуалізацію та аналіз у реальному часі. Основними викликами для цих методів є обробка великих обсягів даних, що генеруються сучасними системами, та зменшення кількості хибних спрацьовувань, які можуть відволікати операторів від реальних загроз.

Системи управління подіями та інформацією безпеки (SIEM) відіграють ключову роль у сучасних стратегіях кібербезпеки, забезпечуючи централізований підхід до збору, аналізу та реагування на інциденти безпеки в масштабі всієї IT-інфраструктури. Такі рішення, як IBM QRadar, Micro Focus ArcSight та Splunk Enterprise Security, дозволяють організаціям в режимі реального часу отримувати події з широкого спектра джерел - від серверів і робочих станцій до мережевих пристроїв, систем керування і прикладного програмного забезпечення. Основна функціональність SIEM зосереджена навколо трьох компонентів: централізованої агрегації логів, кореляції подій із метою виявлення складних шаблонів атак і створення умов для оперативного реагування. Наприклад, у ситуації з потенційною атакою нульового дня SIEM

може поєднати низку подій, таких як неуспішні спроби автентифікації, аномальний об'єм вихідного трафіку з певного сегмента мережі або запуск нетипових процесів, і на основі цього сформувати тригер сповіщення для аналітиків безпеки [14]. Сам механізм кореляції ґрунтується на заздалегідь визначених або адптивно створених правилах, які встановлюють логічні зв'язки між подіями, що на перший погляд можуть бути не пов'язаними між собою. Наприклад, якщо в логах контролера домену фіксується повторювана помилка входу для облікового запису адміністратора, а на міжмережевому екрані реєструється зовнішній доступ із підозрілої IP-адреси до того самого сегмента мережі, система може автоматично позначити ситуацію як індикатор компрометації.

Крім детектування, сучасні SIEM-платформи включають можливості автоматизованого реагування - від надсилання сповіщень до запуску скриптів, які можуть ізолювати вузол чи заблокувати обліковий запис, що підозрюється в компрометації. Однак, попри високу функціональність, впровадження та ефективне використання SIEM-систем пов'язане з низкою викликів. Розробка й оптимізація правил кореляції часто вимагає глибокого розуміння архітектури підприємства, а також специфіки бізнес-процесів і можливих векторів атак. Особливо це стосується складних гетерогенних середовищ, де джерела подій мають різну структуру, обсяг і частоту генерації даних. Крім того, класичні SIEM-рішення в основному орієнтовані на структуровані дані - логи, часові ряди, системні події - і мають обмежену здатність до обробки неструктурованої інформації, такої як звіти про вразливості, вільно сформульовані повідомлення користувачів або технічні описи інцидентів. Це створює потребу в доповненні SIEM-середовища інструментами, що використовують методи обробки природної мови (NLP), здатними витягувати цінну інформацію з текстів, класифікувати інциденти за контекстом і допомагати в прийнятті рішень на основі аналізу мови. Таким чином, інтеграція традиційних SIEM-рішень із сучасними аналітичними інструментами дозволяє не лише покращити точність виявлення, але й зробити реагування на інциденти більш гнучким, інформативним і адаптованим до реальних загроз. З огляду на зростання обсягу та складності даних, які генеруються сучасними інформаційними системами, SIEM-рішення еволюціонують у напрямку більшої автоматизації, масштабованості та інтеграції зі сторонніми джерелами аналітики. Це включає підтримку хмарних платформ, контейнеризованих середовищ і мікросервісної архітектури, де джерела подій можуть активно з'являтися й зникати, що вимагає від SIEM-гравців гнучких підходів до обробки нестатичних даних. Одним із напрямків розвитку є поєднання SIEM із системами розширеного виявлення та реагування (XDR), які дозволяють ширше охоплювати поверхню атаки, включаючи кінцеві точки, поштові системи, мережі та хмарні сервіси в єдиній аналітичній площині. Це особливо актуально для організацій, що працюють у моделі гібридної інфраструктури або активно впроваджують DevOps-підходи, де традиційні засоби контролю безпеки часто не встигають за швидкістю змін у середовищі.

Обробка даних у реальному часі стала не просто бажаною функціональністю, а обов'язковою умовою для забезпечення стійкості та надійності критичних інформаційних систем, де навіть незначна затримка в реакції на інцидент може спричинити серйозні наслідки для бізнесу або інфраструктури. У цьому контексті технології обробки потокових даних, зокрема Apache Kafka, Apache Flink або AWS Kinesis, набули широкого поширення завдяки своїй здатності обробляти величезні обсяги інформації в режимі, наближеному до реального часу. Так, Kafka часто використовується як високопродуктивний брокер повідомлень, який здатен забезпечити передачу і зберігання мільйонів подій на секунду, водночас гарантуючи послідовність, стійкість до відмов і масштабованість, необхідну для обробки телеметричних даних від тисяч серверів, додатків або IoT-пристроїв. На додачу, Flink, орієнтований саме на обчислення поверх потоку даних, дозволяє не просто фільтрувати чи орієнтувати події, а й реалізовувати складні аналітичні операції, як-от виявлення аномалій, агрегації за часовими вікнами або реалізацію стано-орієнтованих автоматів без перерви на зберігання даних у сховищах. У середовищах з постійним інформаційним навантаженням, таких як хмарні інфраструктури або дата-центри, це відкриває можливість виявляти інциденти ще до того, як вони позначаються на кінцевих користувачах. Для прикладу, телеком-оператор може в реальному часі виявляти пікове навантаження в окремому сегменті мережі, миттєво активуючи автоматизовані скрипти переспрямування трафіку або масштабування ресурсів, що дає змогу уникнути перебоїв у наданні послуг. Проте реальна цінність такої оперативної аналітики розкривається лише тоді, коли вона поєднується з прогнозним аналізом, який дозволяє дивитися не тільки на те, що відбувається зараз, а й передбачати, що може трапитись у майбутньому. Моделі типу ARIMA [15] або Prophet [16] використовуються для обробки часових рядів, таких як завантаженість процесора, кількість запитів до API або температурні показники обладнання, що дозволяє формувати прогнози з урахуванням сезонності, циклічності чи раптових сплесків. Це надзвичайно важливо в системах, де планування потужностей має бути випереджальним: скажімо, у банківській інфраструктурі, де обсяги транзакцій суттєво зростають у конкретні дні або години, можливість заздалегідь масштабувати обчислювальні ресурси дозволяє уникнути збоїв і втрати клієнтів. Ще одним рівнем складності є інтеграція структурованих даних - як-от метрик чи логів - із неструктурованими джерелами, такими як текстові звіти про інциденти, електронні листи або описові записи аналітиків. Наприклад, за допомогою алгоритмів обробки природної мови (NLP) можна класифікувати типи атак, виявити повторювані сценарії чи формулювання, що вказують на наявність трендів, які варто врахувати в прогнозуванні [17].

Поєднання текстових описів з показниками систем дозволяє формувати багатовимірні моделі оцінки ризиків, де кожен новий сигнал аналізується не ізольовано, а в контексті історичного досвіду. Водночас ключовими технічними викликами залишаються підтримка мінімальної затримки в обробці великих обсягів

даних, забезпечення цілісності та послідовності потоків у складних архітектурах мікросервісів, а також необхідність ретельного збирання, нормалізації та зберігання історичних даних, які є критично важливими для навчання моделей прогнозування. Без якісного датасету жодна аналітична система не зможе дати адекватного прогнозу, навіть якщо сама по собі є надсучасною з точки зору архітектури чи алгоритмів.

Концептуальна модель інтеграції вебсистем

Запропонована концептуальна модель інтеграції вебсистем поєднує модулі розпізнавання тексту та моніторингу інцидентів у єдину платформу, що дозволяє ефективно обробляти різнорідні інформаційні потоки. Ця модель спрямована на створення синергії між обробкою неструктурованих текстових даних (звітів, логів, документації) та структурованих даних (метрик, подій), що підвищує точність, швидкість і ефективність реагування на інциденти в критичних системах. Архітектура інтегрованої вебсистеми складається з двох основних модулів, які взаємодіють через чітко визначені потоки даних. Модуль розпізнавання тексту відповідає за обробку неструктурованих даних, таких як текстові звіти, логи, документація чи повідомлення. Цей модуль використовує технології обробки природної мови (NLP), такі як моделі BERT, GPT або RoBERTa, для витягнення семантичної інформації, класифікації тексту та аналізу настроїв. Наприклад, модель BERT може бути використана для категоризації звітів про кібератаки за рівнем загрози. Крім того, оптичне розпізнавання символів (OCR), реалізоване через інструменти, такі як Tesseract або ABBYY FineReader, дозволяє оцифровувати відскановані документи або зображення, що містять текстову інформацію, для подальшого аналізу [18].

Модуль моніторингу обробляє структуровані дані, такі як метрики продуктивності (навантаження CPU, використання пам’яті), системні події або логи. Цей модуль використовує методи аналізу аномалій, моделі машинного навчання (наприклад, Random Forest або LSTM) та SIEM-системи для виявлення інцидентів. Наприклад, SIEM-система може виявити аномальний мережевий трафік, а ML-модель може передбачити потенційний збій на основі історичних даних.

Потоки даних між модулями забезпечують інтеграцію: витягнена з текстових джерел інформація, наприклад, ключові слова чи описи подій, передається до модуля моніторингу для кореляції зі структурованими даними. Наприклад, якщо текстовий звіт містить інформацію про нову вразливість, модуль моніторингу може перевірити логи на наявність відповідних подій, таких як спроби експлуатації цієї вразливості. Ця архітектура підтримує хмарні технології, такі як AWS або Azure, для забезпечення масштабованості та обробки великих обсягів даних.

Інтеграція модулів розпізнавання тексту та моніторингу створює синергію, яка значно підвищує ефективність обробки інформаційних потоків. По-перше, швидше виявлення інцидентів досягається завдяки автоматичній обробці текстових даних. Наприклад, NLP-модель може проаналізувати звіт про нову кібератаку, витягнути ключові характеристики (тип атаки, уразливі системи) і передати цю інформацію модулю моніторингу для негайної перевірки відповідних логів. Це зменшує час, необхідний для ідентифікації загрози, порівняно з ручним аналізом. По-друге, більш повний аналіз забезпечується завдяки збагаченню структурованих даних текстовими. Наприклад, текстовий аналіз може ідентифікувати згадки про нову вразливість у звітах безпеки, які потім корелюються з метриками серверів для виявлення спроб її експлуатації. Це дозволяє створювати більш точну картину подій, враховуючи як кількісні (метрики), так і якісні (текстові описи) дані. По-третє, прогнозування загроз стає можливим завдяки аналізу трендів у текстових даних. Наприклад, частота згадок певного типу атак у звітах може вказувати на зростання їхньої популярності, що дозволяє модулю моніторингу налаштувати правила для раннього виявлення таких загроз. Алгоритми глибокого навчання, такі як трансформери, можуть аналізувати історичні текстові дані для прогнозування майбутніх інцидентів, що доповнює традиційні методи прогнозного аналізу, такі як ARIMA.

Ця синергія дозволяє інтегрованій системі не лише реагувати на поточні інциденти, але й проактивно запобігати майбутнім загрозам, що є ключовою перевагою для критичних систем.

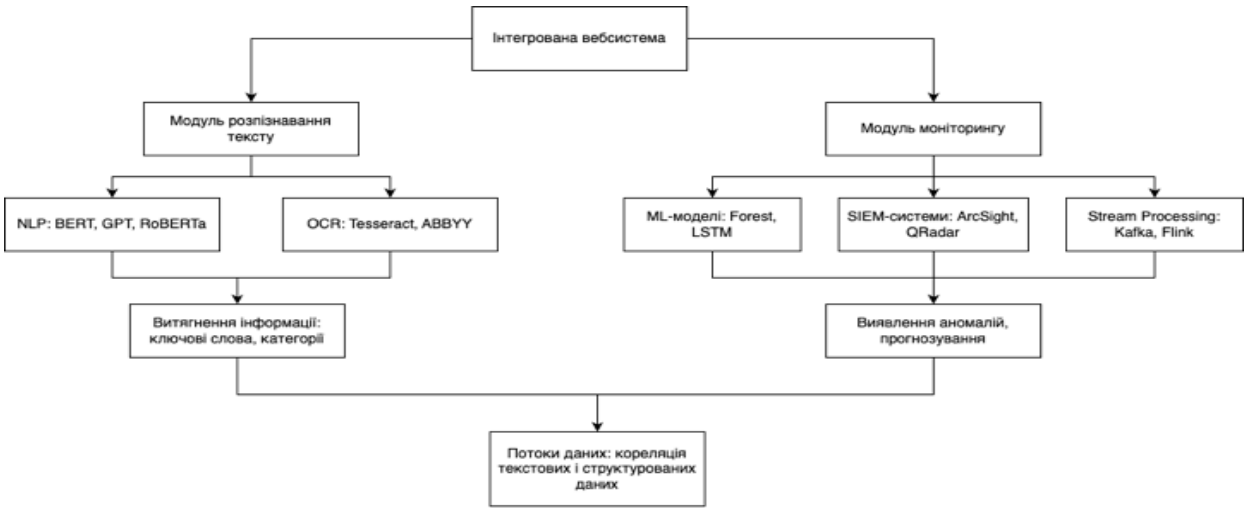

Рис. 1. Модель інтеграції вебсистем

Інтегрована вебсистема, що поєднує розпізнавання тексту та моніторинг, має широкий спектр застосувань у різних сферах, основні сценарії включають:

- кібербезпека: Інтегрована система може аналізувати текстові звіти про вразливості (наприклад, з баз CVE) і корелювати їх із логами серверів для виявлення спроб атак. Наприклад, витягнення інформації про нову вразливість у програмному забезпеченні може спонукати систему перевірити, чи використовується це ПЗ на серверах, і вжити превентивних заходів. SIEM-системи, інтегровані з NLP, можуть автоматично класифікувати текстові описи інцидентів, що зменшує час реагування;
- адміністрування серверів: Система може обробляти текстову документацію, таку як інструкції з конфігурації серверів, і автоматично порівнювати її з поточними налаштуваннями для виявлення невідповідностей. Наприклад, OCR може оцифрувати паперові інструкції, а NLP-модель може витягнути ключові параметри для порівняння з метриками серверів. Це спрощує адміністрування та знижує ризик людських помилок;
- розумні міста: У контексті розумних міст інтегрована система може аналізувати текстові звіти від датчиків, камер або операторів (наприклад, описи аварій на дорогах) і корелювати їх із структурованими даними, такими як трафік чи показники якості повітря. Наприклад, текстовий аналіз звітів про стан доріг може допомогти системі управління трафіком оптимізувати маршрути для аварійних служб [19].

Кожна з цих сфер має унікальні вимоги, але інтегрована система пропонує універсальний підхід до обробки різнорідних інформаційних потоків. Наприклад, у кібербезпеці акцент робиться на швидкості реагування та точності кореляції, тоді як у розумних містах важлива масштабованість і інтеграція з IoT-пристроями.

Виклики та обмеження

Розробка та впровадження інтелектуальних вебсистем, які інтегрують модулі розпізнавання тексту (NLP, OCR) і моніторингу (ML-моделі, SIEM-системи, stream processing), відкривають значні можливості для автоматизації обробки інформаційних потоків у таких сферах, як кібербезпека, адміністрування серверів і розумні міста. Проте реалізація такої системи стикається з численними викликами та обмеженнями, що охоплюють технічні, організаційні, етичні та масштабувальні аспекти. Ці перешкоди впливають на ефективність, доступність і застосовність системи, вимагаючи комплексного підходу до їхнього подолання. У цьому розділі детально розглядаються ключові виклики, їхній вплив на функціонування системи, а також пропонуються стратегії для їхнього вирішення, щоб забезпечити стабільність, точність і універсальність інтегрованих вебсистем.

Технічні виклики є одними з найкритичніших, оскільки інтелектуальні вебсистеми обробляють величезні обсяги різнорідних даних у реальному часі. Одним із основних викликів є обробка великих обсягів даних, що генеруються серверними логами, метриками продуктивності, текстовими звітами та документацією. Наприклад, сучасні дата-центри можуть створювати терабайти логів щодня, тоді як обробка текстових звітів про кіберзагрози за допомогою моделей глибокого навчання, таких як BERT чи GPT-4, вимагає значних обчислювальних ресурсів [20]. Ця проблема ускладняється необхідністю аналізу даних у реальному часі для швидкого виявлення інцидентів, таких як кібератаки чи збої обладнання. Технології stream processing, як Apache Kafka або Apache Flink, хоча й ефективні, можуть зазнавати затримок під час пікових навантажень, що знижує оперативність реагування. Крім того, підтримка такої інфраструктури, особливо у хмарних середовищах, як AWS, Google Cloud чи Microsoft Azure, потребує значних фінансових витрат, що може бути непосильним для малих організацій або тих, що працюють у країнах із обмеженим доступом до високопродуктивних обчислень. Для вирішення цього виклику пропонується використання edge computing [21], що дозволяє попередньо обробляти дані на локальних пристроях, зменшуючи навантаження на центральні сервери. Також ефективними є методи оптимізації моделей ШІ, такі як дистилювання нейронних мереж, які знижують обчислювальні вимоги без значної втрати точності, та алгоритми стиснення даних для зменшення обсягів передачі інформації через мережі.

Іншим значним технічним викликом є інтеграція різнорідних систем, що входять до складу інтелектуальної вебсистеми. Модуль розпізнавання тексту, який використовує NLP для аналізу звітів і OCR для оцифрування документів, генерує неструктуровані текстові дані, тоді як модуль моніторингу працює зі структурованими подіями, як метрики чи логи у форматі CEF, що підтримуються SIEM-системами, такими як ArcSight чи QRadar. Несумісність форматів даних, протоколів і API ускладнює їхню взаємодію. Наприклад, векторні представлення тексту, отримані з NLP-моделей, потребують трансформації для кореляції з подіями в SIEM, що може викликати затримки або втрату інформації. Крім того, OCR-системи, як Tesseract, часто видають неточні результати при обробці документів низької якості, що знижує надійність витягненої інформації. Для подолання цього виклику рекомендується розробка middleware-шарів, які уніфікують дані у стандартних форматах, як JSON або XML, а також використання відкритих стандартів обміну даними, таких як STIX/TAXII [22] для кібербезпеки. Контейнеризація за допомогою Docker і оркестрація через Kubernetes також можуть спростити розгортання та інтеграцію модулів, забезпечуючи їхню модульність і масштабованість.

Точність моделей ШІ та проблема хибних спрацьовувань є ще одним технічним обмеженням, бо такі моделі машинного навчання, як Random Forest чи LSTM, що використовуються для виявлення аномалій у логах, можуть неправильно класифікувати нормальні події як загрози (хибнопозитивні) або пропускати реальні інциденти (хибнонегативні) через недостатню якість тренувальних даних або складність сучасних кібератак. Аналогічно, NLP-моделі можуть інтерпретувати тексти з помилками, особливо якщо вони містять технічну термінологію чи багатозначні вирази. Наприклад, звіт про "фішинг" може бути неправильно класифікований як нешкідливий, якщо модель не врахує контекст. Це знижує довіру до системи та

може призвести до перевантаження операторів через необхідність перевірки хибних сигналів. Для підвищення точності пропонується використання ансамблевих методів, які комбінують кілька моделей для кращої класифікації, регулярне оновлення тренувальних наборів даних із урахуванням нових типів загроз, а також впровадження інтерфейсів для зворотного зв'язку, що дозволяють операторам коригувати помилки моделей у реальному часі [23, 24].

Організаційні виклики також суттєво впливають на впровадження інтелектуальних вебсистем, одним із них є нестача кваліфікованих фахівців, здатних розробляти, налаштовувати та підтримувати складні системи, що поєднують ШІ, кібербезпеку та обробку великих даних. Наприклад, налаштування кореляційних правил у SIEM-системах вимагає глибокого розуміння кіберзагроз, тоді як оптимізація NLP-моделей потребує знань у глибокому навчанні та обробці природної мови. У багатьох організаціях, особливо в економічно слабо розвинених країнах, таких фахівців бракує, що може затримати впровадження або знизити ефективність системи. Для вирішення цього рекомендується організація внутрішніх тренінгів, співпраця з університетами для підготовки спеціалістів, а також використання інструментів автоматизації, таких як AutoML, які спрощують створення та налаштування ML-моделей для некваліфікованих користувачів.

Опір змінам у організаціях є ще одним організаційним викликом, бо персонал, звиклий до традиційних методів, як ручний аналіз логів або rule-based системи, може чинити опір переходу до ШІ-базованих рішень через їхню складність і непрозорість. Наприклад, моделі глибокого навчання часто сприймаються як "чорний ящик", що ускладнює інтерпретацію їхніх рішень. Це може викликати недовіру, особливо в критичних сферах, як кібербезпека, де помилки мають серйозні наслідки. Для подолання цього пропонується проводити пілотні проєкти, які демонструють переваги системи, розробляти інтуїтивно зрозумілі інтерфейси користувача для спрощення взаємодії, а також залучати працівників до процесу розробки, щоб підвищити їхню довіру та розуміння технології.

Відповідність регуляторним вимогам є важливим організаційним обмеженням, особливо для систем, що обробляють чутливі дані, як звіти про кібератаки чи метрики IoT-пристроїв у розумних містах. Міжнародні стандарти, такі як GDPR (ЄС), CCPA (Каліфорнія) або ISO/IEC 27001, вимагають суворого захисту персональних даних і прозорості обробки. Наприклад, текстовий аналіз звітів може включати імена чи IP-адреси, що підпадають під регуляції. Недотримання стандартів може призвести до штрафів або втрати репутації. Для вирішення цього необхідно впроваджувати механізми анонімізації даних, використовувати шифрування (наприклад, AES-256 для зберігання, TLS для передачі) і проводити регулярні аудити безпеки, щоб забезпечити відповідність вимогам [25].

Упередженість моделей ШІ є серйозною проблемою, оскільки тренувальні дані можуть бути незбалансованими або містити історичні упередження. Наприклад, якщо набір даних для NLP складається переважно з англомовних звітів про кіберзагрози, система може погано обробляти тексти іншими мовами, такими як українська чи китайська, що обмежує її застосовність у глобальному контексті. Упередженість може призвести до несправедливих рішень, як-от ігнорування загроз, характерних для певних регіонів. Для вирішення цього пропонується використовувати різноманітні набори даних, застосовувати методи fairness-aware machine learning і проводити регулярне тестування моделей на упередженість за допомогою метрик, як-от demographic parity.

Інтеграція текстових і структурованих даних підвищує ризик витоку інформації, особливо якщо система обробляє чутливі дані, як звіти про вразливості чи персональні метрики IoT. Наприклад, вразливості в модулі OCR можуть дозволити зловмисникам отримати доступ до оцифрованих документів. Для захисту даних рекомендується впроваджувати диференційну приватність, яка додає шум до результатів обробки, зберігаючи їхню корисність, а також використовувати захищені протоколи, як HTTPS і SSH. Регулярне тестування на проникнення (penetration testing) також допоможе виявити та усунути вразливості.

Соціальні наслідки автоматизації, зокрема скорочення робочих місць, є важливим обмеженням, бо автоматизація моніторингу може замінити ручну працю, особливо для низькокваліфікованих аналітиків логів, що викликає соціальну напругу. Наприклад, у великих організаціях перехід до ШІ може призвести до протестів або зниження мотивації працівників. Для пом'якшення цього пропонується перекваліфікація персоналу для роботи з новими системами, створення ролей, пов'язаних із наглядом за ШІ, і акцент на гібридному підході, де людина і машина співпрацюють, а не конкурують.

Масштабованість системи також має свої обмеження, отже адаптація до різних доменів, таких як охорона здоров'я чи фінанси, є складною через специфіку даних і вимог. Наприклад, аналіз медичних звітів потребує обробки складної термінології, тоді як фінансові дані підпадають під регуляції, як PCI DSS. Для забезпечення універсальності рекомендується розробка модульної архітектури, яка дозволяє підключати спеціалізовані моделі, і використання transfer learning для адаптації NLP-моделей до нових доменів із мінімальними зусиллями. Обмеження в реальному часі, пов'язані з обробкою пікових навантажень, також ускладнюють масштабування. Наприклад, складні текстові запити, як аналіз багатосторінкових звітів, можуть перевантажити систему, знижуючи її продуктивність. Для цього пропонується використовувати розподілені обчислення, пріоритезацію критичних завдань (наприклад, виявлення атак) і кешування даних для зменшення затримок. Впровадження інтелектуальних вебсистем для автоматизації обробки та моніторингу інформаційних потоків стикається з комплексними викликами, які охоплюють технічні (обробка даних, інтеграція, точність), організаційні (кваліфікація, регуляції, опір змінам), етичні (упередженість, конфіденційність) і масштабувальні (адаптація, реальний час) аспекти. Подолання цих обмежень вимагає поєднання технологічних інновацій, таких як edge computing і стандартизація даних, організаційних

заходів, включаючи тренінги та аудити, і етичних практик, як анонімізація та fairness. Лише через комплексний підхід можна реалізувати повний потенціал таких систем, забезпечуючи їхню надійність, ефективність і універсальність у критичних сферах, як кібербезпека, адміністрування серверів і розумні міста.

Висновки

Інтелектуальні вебсистеми, що поєднують методи штучного інтелекту, машинного навчання, обробки природної мови (NLP) та оптичного розпізнавання символів (OCR), є інноваційним рішенням для автоматизації обробки та моніторингу інформаційних потоків. Проведений аналіз демонструє їх значний потенціал у таких сферах, як кібербезпека, адміністрування серверів та управління інфраструктурою розумних міст. Ці системи забезпечують швидке виявлення інцидентів, прогнозування загроз і оптимізацію процесів завдяки інтеграції текстових і структурованих даних, що сприяє підвищенню ефективності та надійності критичних інформаційних систем.

Ключовими перевагами інтелектуальних вебсистем є їхня адаптивність, здатність до самонавчання та можливість обробки великих обсягів різнорідних даних у реальному часі. Вони дозволяють автоматизувати складні завдання, такі як аналіз логів, розпізнавання тексту та кореляція подій, зменшуючи час реагування на інциденти та знижуючи ризик людських помилок. Інтеграція модулів розпізнавання тексту та моніторингу створює синергію, яка забезпечує більш повний аналіз і проактивне запобігання загрозам, що є критично важливим для сучасних цифрових середовищ.

Однак реалізація таких систем пов’язана з низкою викликів, включаючи обробку великих обсягів даних, інтеграцію різнорідних модулів, забезпечення точності моделей ШІ та відповідність регуляторним вимогам. Технічні обмеження, такі як висока обчислювальна складність і ризик хибних спрацьовувань, потребують оптимізації алгоритмів і використання технологій edge computing. Організаційні виклики, зокрема нестача кваліфікованих фахівців і опір змінам, підкреслюють необхідність тренінгів і пілотних проєктів. Етичні аспекти, такі як упередженість моделей і захист даних, вимагають впровадження механізмів fairness і анонімізації.

Перспективи подальших досліджень включають розробку модульних архітектур для підвищення масштабованості, вдосконалення алгоритмів глибокого навчання для обробки багатомовних і неструктурованих даних, а також інтеграцію з новими технологіями, такими як XDR і IoT. Подолання зазначених викликів дозволить реалізувати повний потенціал інтелектуальних вебсистем, забезпечуючи їхню універсальність і ефективність у різних сферах застосування.

Список літератури

1. Liu, J., Zhong, N., Yao, Y. et al. The Wisdom Web: New Challenges for Web Intelligence (WI). Journal of Intelligent Information Systems 20, 5–9 (2003). https://doi.org/10.1023/A:1020945620934
2. Market Research Intellect: Online OCR Software Market Size And Forecast (2025); https://www.marketresearchintellect.com/product/global-online-ocr-software-market-size-and-forecast/
3. C. Stryker, J. Holdsworth: What is NLP (natural language processing)? (2024); https://www.ibm.com/think/topics/natural-language-processing
4. IBM Inc.: What is infrastructure monitoring? (2023): https://www.ibm.com/think/topics/infrastructure-monitoring
5. R. Guha, "Toward the Intelligent Web Systems," 2009 First International Conference on Computational Intelligence, Communication Systems and Networks, Indore, India, 2009, pp. 459-463, doi: 10.1109/CICSYN.2009.25.
6. Boudabous Maroua, Pappa Anna. WebT-IDC: A Web Tool for Intelligent Dataset Creation A Use Case for Forums and Blogs, Procedia Computer Science, Volume 192, 1051-1060 (2021). https://doi.org/10.1016/j.procs.2021.08.108.
7. Noetzold, D., Rossetto, A. G. D. M., Leithardt, V. R. Q., & Costa, H. . J. de M.: Enhancing Infrastructure Observability: Machine Learning for Proactive Monitoring and Anomaly Detection. (2024); https://doi.org/10.5753/jisa.2024.4509
8. C. Garrido-Munoz, A. Rios-Vila, J. Calvo-Zaragoza. Handwritten Text Recognition: A Survey (2025). https://doi.org/10.48550/arXiv.2502.08417
9. S. Rakesh, P. Kushal Reddy, V. Prashanth and K. Srinath Reddy. Handwritten text recognition using deep learning techniques: A survey. International Conference on Multidisciplinary Research and Sustainable Development (2024). https://doi.org/10.1051/matecconf/202439201126
10. M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, F. Wei. TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models (2025) https://doi.org/10.48550/arXiv.2109.10282
11. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019). https://doi.org/10.48550/arXiv.1810.04805
12. Popa, Sorin. "WEB Server monitoring." Annals of University of Craiova-Economic Sciences Series 2.36 (2008): 710-715. https://ideas.repec.org/a/aio/aucsse/v2y2008i11p710-715.html
13. Zeng, Wenxian, and Yue Wang. "Design and implementation of server monitoring system based on SNMP." 2009 International Joint Conference on Artificial Intelligence. IEEE, 2009. https://doi.org/10.1109/JCAI.2009.34
14. Mail, I. D. "The future of SIEM in a machine learning-driven cybersecurity landscape." Turkish Journal of Computer and Mathematics Education Vol 14.03 (2023): 1309-1314. https://doi.org/10.61841/turcomat.v14i03.14392
15. Teggi, P., Malakreddy, B., Teggi, P. P., & Harivinod, N. (2021). AIOPS prediction for server stability based on ARIMA Model. Int. J. Eng. Res. Technol, 10, 128-134. https://doi.org/10.17577/IJERTV10IS120055
16. Pasquadibisceglie, V., Scaringi, R., Appice, A., Castellano, G., & Malerba, D. (2024). PROPHET: explainable predictive process monitoring with heterogeneous graph neural networks. IEEE Transactions on Services Computing. https://doi.org/10.1109/TSC.2024.3463487
17. Sheeraz, M., Paracha, M. A., Haque, M. U., Durad, M. H., Mohsin, S. M., Band, S. S., & Mosavi, A. (2023). Effective security monitoring using efficient SIEM architecture. Hum.-Centric Comput. Inf. Sci, 13, 1-18. https://doi.org/10.22967/HCIS.2023.13.023

18. Tafti, Ahmad P., et al. "OCR as a service: an experimental evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym." Advances in Visual Computing: 12th International Symposium, ISVC 2016, Las Vegas, NV, USA, December 12-14, 2016, Proceedings, Part I 12. Springer International Publishing, 2016. https://doi.org/10.1007/978-3-319-50835-1_66
19. FERNOAGA, Vlad Paul, et al. "OCR-based solution for the integration of legacy and-or non-electric counters in cloud smart grids." 2018 IEEE 24th International Symposium for Design and Technology in Electronic Packaging (SIITME). IEEE, 2018. https://doi.org/10.1109/SIITME.2018.8599200
20. Koroteev, Mikhail V. "BERT: a review of applications in natural language processing and understanding." arXiv preprint arXiv:2103.11943 (2021). https://doi.org/10.48550/arXiv.2103.11943
21. Varghese, Blesson, et al. "Challenges and opportunities in edge computing." 2016 IEEE international conference on smart cloud (SmartCloud). IEEE, 2016. https://doi.org/10.1109/SmartCloud.2016.18
22. Barnum, Sean. "Standardizing cyber threat intelligence information with the structured threat information expression (stix)." Mitre Corporation 11 (2012): 1-22. https://www.mitre.org/sites/default/files/publications/stix.pdf
23. Sheta, Sagar Vishnubhai. "Developing efficient server monitoring systems using AI for real-time data processing." (2023). https://doi.org/10.2139/ssrn.5034125
24. Kairatuly, Akhmetzhanov Batyrzhan, et al. "Development of a Security and Monitoring System for Server Equipment Utilizing IoT Technologies." 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST). IEEE, 2024. https://doi.org/10.1109/SIST61555.2024.10629465
25. Massonet, Philippe, et al. "A monitoring and audit logging architecture for data location compliance in federated cloud infrastructures." 2011 IEEE international symposium on parallel and distributed processing workshops and PhD forum. IEEE, 2011. https://doi.org/10.1109/IPDPS.2011.304

Received (Надійшла) 16.05.2025
Accepted for publication (Прийнята до друку) 13.08.2025

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Заіка Владислав Віталійович – студент кафедри комп'ютерних систем, мереж і кібербезпеки, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна;
Vladyslav Zaika – student, Department of Computer Systems, Networks and Cybersecurity, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine;
e-mail: v.v.zaika@student.csn.khai.edu; ORCID Author ID: https://orcid.org/0009-0007-1750-9419.

Чуєв Олег В'ячеславович – студент кафедри комп'ютерних систем, мереж і кібербезпеки, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна;
Oleh Chuiev – student, Department of Computer Systems, Networks and Cybersecurity, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine;
e-mail: o.v.chuiev@student.csn.khai.edu; ORCID Author ID: https://orcid.org/0000-0002-3584-4380.

Морозова Ольга Ігорівна – доктор технічних наук, професор, професор кафедри комп'ютерних систем, мереж і кібербезпеки, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна;
Olga Morozova – Doctor of Technical Sciences, Professor, Professor of the Department of Computer Systems, Networks and Cybersecurity, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine;
e-mail: o.morozova@khai.edu; ORCID Author ID: https://orcid.org/0000-0001-7706-3155;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57194517520.

Нікітіна Тетяна Сергіївна – кандидат технічних наук, доцент кафедри комп'ютерних систем, мереж і кібербезпеки, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна;
Tetyana Nikitina – PhD in Technical Sciences, Associate Professor, Department of Computer Systems, Networks and Cybersecurity, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine;
e-mail: t.nikitina@khai.edu; ORCID Author ID: https://orcid.org/0009-0000-7549-1017;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=56559630000&origin=resultslist.

Intelligent web systems for automation of processing and monitoring of information flows

Vladyslav Zaika, Oleh Chuiev, Olga Morozova, Tetyana Nikitina

Abstract. In the conditions of rapid growth of digital data and information flows on the Internet, there is a growing need for automated methods of their processing and monitoring for effective data acquisition and making informed conclusions. In order to cope with this data flow, intelligent web systems are being created. They use artificial intelligence (AI), machine learning (ML) and other advanced technologies to automatically collect, analyze and present information. The main purpose of this article is a comprehensive review of intelligent web systems aimed at automating the processing and monitoring of information flows. The theoretical foundations of such systems are considered, including semantic technologies and artificial intelligence methods, as well as modern approaches to text recognition and analysis based on deep learning. An analysis of modern approaches to automated text recognition (OCR and deep learning methods) and modern server infrastructure monitoring tools is presented. A conceptual model of integration of components of text processing systems and infrastructure monitoring is proposed. Key challenges and limitations of this approach are highlighted, and promising directions for further research are outlined.

Keywords: intelligent web systems, information flows, automation, artificial intelligence, optical character recognition (OCR), infrastructure monitoring, machine learning, natural language processing (NLP), security information and event management (SIEM), Semantic Web.