

Шифр: **ФЕЙК**

Тема роботи: **Розробка детектора фейкових новин на основі  
штучного інтелекту**

## ЗМІСТ

|  |    |
|--|----|
| ВСТУП .....  | 3  |
| РОЗДІЛ 1 ТЕОРЕТИЧНІ ОСНОВИ ВИЯВЛЕННЯ ФЕЙКОВИХ НОВИН.....               | 4  |
| 1.1 Поняття дезінформації та фейкових новин.....                       | 4  |
| 1.2 Аналіз джерел за темою дослідження .....                           | 5  |
| 1.3 Метрики та оцінювання якості моделей.....                          | 8  |
| РОЗДІЛ 2 ДОСЛІДЖЕННЯ І АНАЛІЗ ДАНИХ.....                               | 10 |
| 2.1 Формування українського датасету .....                             | 10 |
| Таблиця 2.1 — Структура та опис датасету.....                          | 11 |
| 2.2 Попередня обробка .....  | 11 |
| 2.3 Архітектура моделі .....   | 11 |
| 2.3 Експерименти з моделями .....                                      | 16 |
| Таблиця 2.2 — Порівняння якості моделей за метрикою F1-score .....     | 16 |
| РОЗДІЛ 3 ПРОЄКТУВАННЯ ТА РЕАЛІЗАЦІЯ СИСТЕМИ.....                       | 18 |
| 3.1 Архітектура рішення .....  | 18 |
| 3.2 Пайплайн реалізації модулів: навчання, API та веб-інтерфейсу ..... | 18 |
| 3.4 Тестування, безпека та масштабування.....                          | 20 |
| ВИСНОВКИ.....  | 21 |
| ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ.....                                       | 22 |
| ДОДАТКИ.....   | 24 |
| Додаток А. Приклад JSON-запиту до API та відповіді                     |    |
| Додаток Б. Конфігурація тренування (скорочено)                         |    |
| Додаток В. Стаття  |    |
| Додаток Г. Тези  |    |

## ВСТУП

Поширення дезінформації та фейкових новин у цифровому середовищі стало одним із ключових викликів для суспільства, медіа та державних інституцій. Український інформаційний простір, що переживає наслідки повномасштабної війни, особливо вразливий до маніпуляцій і психологічних операцій. Запровадження автоматизованих інструментів аналізу достовірності інформації на основі сучасних методів обробки природної мови (NLP) та глибокого навчання дозволяє оперативно виявляти сумнівні твердження, формувати пояснювані висновки та підвищувати рівень медіаграмотності.

Мета цієї роботи — спроектувати й реалізувати прототип детектора фейкових новин, адаптований до українського контексту, із підтримкою україномовних даних, сучасних архітектур (BERT/BiLSTM з увагою), а також із можливістю розгортання веб-сервісу (API) для інтеграції з освітніми та медіа-платформами.

Об'єктом дослідження є процеси автоматизованого аналізу текстів новин і коротких повідомлень.

Предметом дослідження — методи NLP та архітектури нейронних мереж, придатні для класифікації достовірності.

Завдання: провести огляд сучасних методів виявлення дипфейків, сформувати локальний датасет, спроектувати архітектуру системи, провести експерименти з декількома моделями, оцінити їх точність і сформувати рекомендації для промислового розгортання.

## РОЗДІЛ 1 ТЕОРЕТИЧНІ ОСНОВИ ВИЯВЛЕННЯ ФЕЙКОВИХ НОВИН

### 1.1 Поняття дезінформації та фейкових новин.

Фейкові новини — це повідомлення, що містять неправдиві або маніпулятивні твердження, поширювані з метою впливу на громадську думку, емоційний стан або поведінку читача. Вони можуть мати різну природу:

- повністю вигадані тексти;
- напівправа;
- контекстні підміни;
- клікбейтні заголовки;
- навмисне перекручування фактів.



Рисунок 1.1 – Фейкові новини

У науковій літературі зазвичай розрізняють категорії: правда, переважно правда, напівправа, майже неправда, неправда, та «pants-on-fire». Моделі класифікації навчаються розпізнавати мовні патерни та контекст, які корелюють із цими класами.

## 1.2 Аналіз джерел за темою дослідження

Можна простежити еволюцію підходів до виявлення фейкових новин, починаючи з гібридних архітектур у 2019 році і закінчуючи домінуванням трансформерних моделей та мультимодальних систем у період 2020–2025 років.

### 1.2.1 Контекстуалізована увага з бічними даними.

У роботі "Fake News Detection by Learning Convolution Filters through Contextualized Attention" [1] представлено модель Fake-Net, яка використовує інноваційний на той час підхід.

Замість прямого аналізу тексту новин, модель використовує механізм уваги (attention mechanism), щоб врахувати додаткову інформацію (так звані "side information"), таку як тема, спікер, його посада, партійна приналежність, контекст та обґрунтування.

Архітектура - це гібридна модель, що поєднує двоспрямовану LSTM (BiLSTM) для обробки метаданих та згорткову нейронну мережу (CNN) для аналізу основного тексту твердження. Метадані формують "контекстний запит" (context query), який спрямовує увагу CNN на найважливіші слова у твердженні.

Модель тренувалася на наборі даних LIAR, який містить 12 836 коротких заяв з ресурсу POLITIFACT, класифікованих за шістьма категоріями правдивості.

Переваги:

- врахування контексту: Модель показала, що використання бічної інформації значно покращує якість виявлення фейків. Це дозволяє системі зрозуміти, хто і в яких обставинах зробив заяву.

- покращена точність: Порівняно з базовою моделлю без механізму уваги, Fake-Net показала приріст точності приблизно на 2% для класифікації на 6 класів і на 5% для бінарної класифікації (фейк/не фейк).

Недоліки та обмеження:

- застаріла архітектура: На 2019 рік це був прогресивний підхід, але автор сам зазначає, що новітні архітектури, такі як Трансформери (BERT, XLNet), є значно потужнішими, особливо за наявності великих наборів даних.
- слабкі початкові представлення: Вектори слів (embeddings) ініціалізувалися випадково. Використання попередньо навчених моделей, як-от GloVe або word2vec, могло б суттєво покращити результати.
- обмеженість даних: Ефективність моделі обмежувалася відносно невеликим розміром датасету LIAR-PLUS (близько 12 тис. зразків).
- відсутність мультимодальності: Підхід аналізує лише текст та метадані, ігноруючи візуальну складову (зображення, відео), яка часто є ключовим елементом дезінформації.

### 1.2.2 Трансформерні моделі та локалізація для українського контексту

У роботі "Розробка детектора фейкових новин на основі штучного інтелекту" [2] представлено огляд сучасних методів та спроектовано систему, адаптовану для України.

У період з 2020 року домінуючим підходом стали трансформерні архітектури, такі як BERT, RoBERTa, та їх багатомовні аналоги (mBERT, XLM-R). Їхня ключова перевага — здатність враховувати довгі контексти завдяки механізму само-уваги (self-attention). Такі моделі навчаються на великих корпусах текстів і здатні враховувати довгі контексти, що критично для аналізу новин. Паралельно використовуються BiLSTM-мережі з механізмами уваги для задач, де обсяг даних обмежений або потрібна інтерпретованість. В останні роки активний розвиток отримує мультимодальна детекція (текст+зображення), оскільки фейки часто супроводжуються візуальними маніпуляціями. Значну увагу приділяють стійкості моделей до атак і доменної адаптації.

Модель акцентується на важливості створення локалізованого українського датасету новин, зібраних з медіа, Telegram-каналів та

фактчекінгових ресурсів, оскільки інформаційний простір України має свою специфіку, особливо в умовах війни.

Експерименти показали, що донавчені трансформери значно перевершують класичні підходи (TF-IDF) та архітектури типу BiLSTM+Attention.

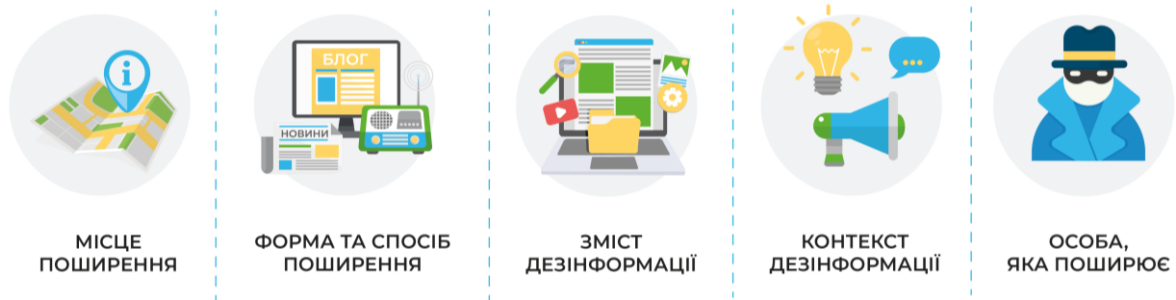


Рисунок 1.2 – Найважливіші критерії оцінки дезінформації

Переваги:

- висока точність: Доновчена на локальних даних модель XLM-R досягла F1-score 88%, що значно вище, ніж у BiLSTM (78%) та логістичної регресії (70%).

- адаптивність: Трансформери можна ефективно "донавчати" (fine-tune) на специфічних даних, що критично важливо для української мови та контексту.

- мультимодальність: Сучасні підходи активно розвивають багатомодальну детекцію (текст+зображення), оскільки фейки часто використовують візуальні маніпуляції, які неможливо виявити лише за текстом.

Недоліки та виклики:

- потреба в пояснюваності (Explainability): Трансформери є складними "чорними скриньками". Для практичного застосування важливо інтегрувати

методи пояснення рішень (наприклад, attention-карти, LIME/SHAP), щоб показати, які саме слова чи фрази вплинули на вердикт моделі.

- вимоги до ресурсів: Навчання та використання великих трансформерних моделей потребує значних обчислювальних ресурсів.

- залежність від даних: Якість роботи моделей безпосередньо залежить від обсягу та якості розміченого навчального датасету. Створення такого датасету є складним і трудомістким процесом.

### 1.2.3 Багатомодальні підходи (текст+зображення)

Мультимодальні моделі поєднують текстові ембеддинги з візуальними ознаками зображень або відео. Це актуально для новин із прикріпленими картинками, мемами та інфографікою, де текстова частина може виглядати правдоподібно, проте зображення маніпулює сприйняттям. Такі моделі застосовують двопотокові трансформери або крос-модальні уваги, покращуючи загальну точність класифікації.

## 1.3 Метрики та оцінювання якості моделей

Основні метрики: accuracy, precision, recall, F1-score, ROC-AUC. У задачах із нерівномірними класами використовують macro/micro-середні значення. В таблиці 1.1 наведено напрями використання цих метрик

Таблиця 1.1 — Напрями використання метрик оцінки якості моделі

| Метрика   | Що вимірює  | Для чого використовується   |
|-----------|---|---|
| 1         | 2   | 3   |
| Accuracy  | Частка правильних передбачень серед усіх прикладів.                 | Дає загальне уявлення про якість моделі, але може бути оманливою при незбалансованих даних.     |
| Precision | Скільки серед усіх передбачених фейкових новин насправді є фейками. | Важлива для зменшення кількості помилкових звинувачень (правдиві новини, визначені як фейкові). |
| Recall    | Скільки з усіх реальних фейкових новин модель змогла виявити.       | Важлива для мінімізації пропущених фейків.  |

## Продовження таблиці 1.1

| 1                | 2  | 3   |
|------------------|--|---|
| F1-score         | Гармонійне середнє між Precision і Recall                            | Дає збалансовану оцінку; часто основна метрика у задачі визначення фейкових новин           |
| Confusion Matrix | Кількість правильних та неправильних класифікацій для кожного класу  | Допомагає зрозуміти, які саме помилки робить модель (пропуски фейків чи хибні звинувачення) |
| ROC-AUC          | Площа під ROC-кривою: здатність відрізнити фейки від правдивих новин | Використовується для оцінки загальної якості моделі незалежно від порогу                    |

Важливо додавати пояснюваність (explainability): attention-карти, SHAP/LIME-аналіз, приклади зважувань токенів. Для порівнянності з іншими роботами потрібно описувати протоколи тренування, розподіл на train/val/test, та фіксувати випадкові зерна.

## Висновки до розділу 1

Проведений огляд демонструє доцільність використання трансформерів та уваги, а також потребу у локальному донавчанні для українського контексту.

## РОЗДІЛ 2 ДОСЛІДЖЕННЯ І АНАЛІЗ ДАНИХ

### 2.1 Формування українського датасету

Для створення локального датасету передбачено збір коротких новинних тверджень та заголовків з українських медіа, Telegram-каналів, офіційних сторінок органів влади, а також фактчекінгових ресурсів. Критерії відбору: новизна, суспільна значущість, наявність незалежних підтверджень або спростувань. Було обрано декілька датасетів для визначення фейкових новин, табл. 2.1

Таблиця 2.1 – Датасти для визначення фейкових новин

| Назва / джерело  | Що містить / особливості  | Чи підійде для виявлення фейкових новин / українська мова  | Посилання / деталі  |
|--|---|--|---|
| Multi-Fake-DetectiVE   | Датасет зі статтями та постами з соціальних мереж + візуальний компонент, що стосуються війни РФ-України. Є завдання “текст + зображення” та “відношення між ними”. | Так — багато матеріалів про війну, можливо частково українською. Дуже корисний як мультимодальний ресурс.                                | ( <a href="https://data.mendeley.com">data.mendeley.com</a> ) |
| “Ukrainian news / information truthfulness” — Kaggle (Ukrainian fake + true news)        | Містить тексти новин, позначених як фейкові та правдиві, стосуються війни РФ-України.   | Так — прямо українська + розмітка fake/true. Це один з найкращих “прямих” датасетів.   | ( <a href="https://selectdataset.com">selectdataset.com</a> ) |
| Detection and classification of manipulations in the text data of Ukrainian Social Media | Постановка на мульти-мітковість (пропагандистські техніки), Telegram-пости українською та російською.   | Так — може бути корисно, особливо для моделей, які треба виявляти різні типи маніпуляцій; але не завжди є чітка розмітка “фейк/не фейк”. | ( <a href="http://er.ucu.edu.ua">er.ucu.edu.ua</a> )          |

Датасети містить тексти, мітки достовірності (правда/напівправда/маніпуляція/неправда), метадані (джерело, дата, тема, тип каналу). Частина корпусу анотована вручну за допомогою настанов із міжекспертною згодою. Для підвищення узгодженості застосовано подвійне маркування та арбітраж спірних прикладів.

Таблиця 2.1 — Структура та опис датасету

| Клас        | Кількість | Приклади           | Джерела                 | Примітка                 |
|-------------|-----------|--------------------|-------------------------|--------------------------|
| Правда      | 2500      | офіційні заяви     | gov.ua, суспільні медіа | верифіковані джерела     |
| Напівправда | 2500      | змішані твердження | ЗМІ/соцмережі           | частково підтверджено    |
| Маніпуляція | 2500      | клікбейт/контекст  | соцмережі               | вибіркові факти          |
| Неправда    | 2500      | фейки              | анонімні канали         | спростовано фактчекерами |

## 2.2 Попередня обробка

Етапи: видалення HTML/URL, нормалізація скорочень, лематизація українською, видалення стоп-слів, токенізація під вибрану модель. Для трансформерів використовується спеціалізований токенайзер; для BiLSTM — словник із підрідкісними токенами (UNK). Виконуються аугментації: синонімічна заміна, випадкові видалення, перефразування для підвищення стійкості.

## 2.3 Архітектура моделі

Основна ідея моделі, полягає у використанні метаданих (додаткової інформації) для контекстуалізації аналізу самого тексту новини. Замість того, щоб аналізувати твердження ізольовано, модель використовує інформацію про автора, тему, контекст висловлювання тощо, щоб "підказати"

неймережі, на які саме слова та патерни в тексті слід звернути увагу. Цей процес реалізований через кілька ключових блоків, рис. 2.1.

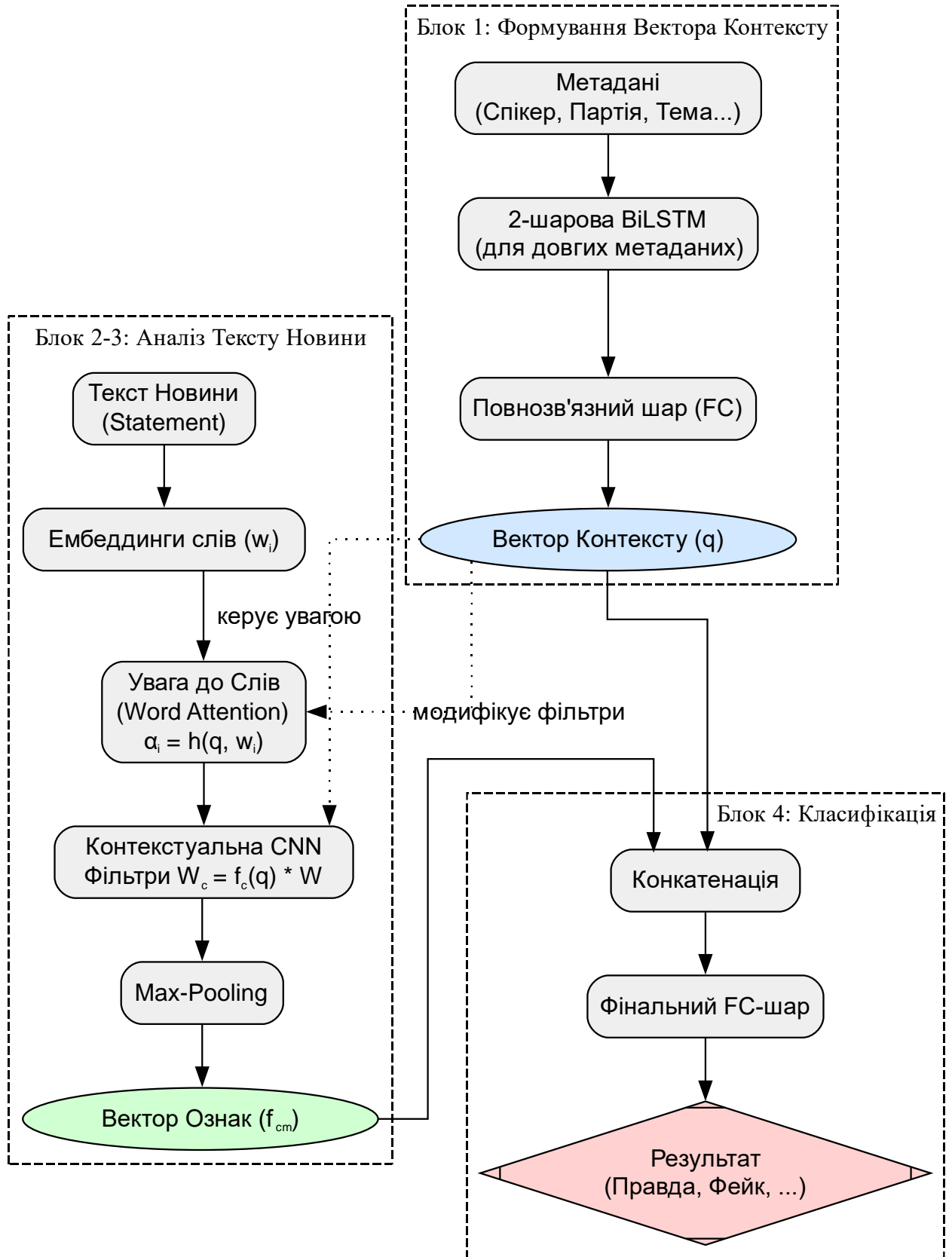


Рисунок 2.1 – Архітектура моделі































|  |  |
|--|--|
| <p style="text-align: center;"><b>МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ</b></p> <p style="text-align: center;"><b>МАТЕРІАЛИ</b></p> <p style="text-align: center;"><b>XVIII ВСЕУКРАЇНСЬКОЇ<br/>НАУКОВО-ТЕХНІЧНОЇ КОНФЕРЕНЦІЇ<br/>«СТАЛІЙ РОЗВИТОК МІСТ: ПОСТВОЄННИЙ<br/>ПЕРІОД»</b></p> <p style="text-align: center;"><b>ЧАСТИНА II</b></p> <p style="text-align: center;">2025</p> |  |
|  | <p>організація бажає мати комплексну стратегію безпеки, вона може використовувати ISO/IEC 27001 для загального управління безпекою та ISO/IEC 15408 для сертифікації своїх ІТ-рішень.</p> <p>FIPS 140: це набір стандартів, який визначає вимоги до криптографічних модулів, що використовуються для захисту інформації [5].</p> <p>CWA 14167: стандарт визначає вимоги до криптографічного модуля для послуг генерування ключів провайдером послуг сертифікації [6].</p> <p>Порівняльний аналіз стандартів FIPS 140 та CWA 14167 вказує, що FIPS 140 є більш жорстким стандартом, орієнтованим на криптографічні модулі та їх сертифікацію, тоді як CWA 14167 охоплює ширше коло питань довірчих послуг, включаючи управління ключами та цифровими підписами.</p> <p>FIPS 140 використовується переважно в США, тоді як CWA 14167 є частиною європейської системи довіри.</p> <p>CWA 14167 зосереджується на електронних підписах та довірчих службах, тоді як FIPS 140 акцентує увагу на фізичному захисті криптографічних модулів.</p> <p>Таким чином, вибір між цими стандартами залежить від специфіки застосування: FIPS 140 підходить для високозахисних урядових систем, тоді як CWA 14167 краще відповідає потребам цифрових сервісів у ЄС.</p> <p>У майбутньому очікується подальша інтеграція стандартів якості у процеси розробки ПЗ. Зокрема, розвиток технологій штучного інтелекту та машинного навчання сприятиме створенню більш адаптивних та проактивних систем кібербезпеки. Крім того, зростатиме роль міжнародних стандартів у встановленні єдиних вимог до безпеки ПЗ, що полегшить співпрацю між організаціями та підвищить загальний рівень захищеності цифрової інфраструктури.</p> <p>Список використаних джерел:</p> <ol style="list-style-type: none"> <li>1. <a href="https://visuresolutions.com/uk/blog/cybersecurity-engineering/?utm_source=D1%83">https://visuresolutions.com/uk/blog/cybersecurity-engineering/?utm_source=D1%83</a></li> <li>2. <a href="https://eba.com.ua/rol-shitchnogo-intelektu-v-kiberbezpeti-peredbachemnyata-zapobigannya-atakam/?utm_source">https://eba.com.ua/rol-shitchnogo-intelektu-v-kiberbezpeti-peredbachemnyata-zapobigannya-atakam/?utm_source</a></li> <li>3. ISO/IEC 27001:2013 <a href="https://itref.ir/uploads/editor/42890b.pdf">https://itref.ir/uploads/editor/42890b.pdf</a></li> <li>4. ISO/IEC 15408: <a href="https://cdn.standards.itech.ai/samples/72891/fbfa4f603f84c7ab0663ffc4c163f3f15O-IEC-15408-1-2022.pdf">https://cdn.standards.itech.ai/samples/72891/fbfa4f603f84c7ab0663ffc4c163f3f15O-IEC-15408-1-2022.pdf</a></li> <li>5. FIPS 140: <a href="https://en.wikipedia.org/wiki/FIPS_140">https://en.wikipedia.org/wiki/FIPS_140</a></li> <li>6. CWA 14167: <a href="https://www.evs.ee/en/cwa-14167-3-2004">https://www.evs.ee/en/cwa-14167-3-2004</a></li> </ol> |