УДК 004.8:004.056

МЕТОДИ ВИЯВЛЕННЯ ШКІДЛИВОГО КОНТЕНТУ, СТВОРЕНОГО ШІ

Садчиков В.О.
repvladika.ua@gmail.com
Черкаський державний фаховий бізнес-коледж
Захарова М.В.
м.Черкаси, Україна

Стрімкий розвиток генеративного штучного інтелекту дав змогу створювати реалістичні тексти, зображення, аудіо та відео, які дедалі важче відрізнити від справжніх. Це створює нові ризики для інформаційної безпеки, довіри до цифрових комунікацій і суспільної стабільності. Шкідливий контент, створений або змінений за допомогою ШІ, може використовуватися для дезінформації, шахрайства, підробки особи, маніпулювання громадською думкою та інших форм цифрового зловживання. Особливо небезпечним є те, що сучасні генеративні системи здатні імітувати зовнішність, голос, стиль мовлення та навіть окремі поведінкові риси людини. Унаслідок цього зловмисники можуть створювати переконливі фейкові повідомлення, аудіозаписи та відео, які важко відрізнити від справжніх[1]. Це створює серйозні фінансові, репутаційні та соціальні наслідки, тому розроблення та впровадження методів виявлення такого контенту є важливим завданням сучасної інформаційної безпеки.

Шкідливий контент, створений ШІ, можна розглядати як цифровий контент, що був згенерований або суттєво модифікований алгоритмами машинного навчання та може використовуватися для введення в оману, маніпуляції або поширення небезпечної інформації. Для протидії таким ризикам важливе значення має цифрова прозорість контенту, тобто можливість встановити, як саме було створено або змінено цифровий об’єкт, хто або що брало участь у його формуванні, а також чи були в нього внесені зміни після створення. У цьому контексті застосовують цифрові водяні знаки, метадані, цифрові підписи та інші засоби фіксації походження контенту[2].

Методи виявлення синтетичного контенту доцільно поділяти на три основні групи: виявлення за даними про походження, автоматизоване виявлення

на основі самого контенту та виявлення за участю людини. Ці підходи не виключають один одного й зазвичай дають найкращий результат у поєднанні. Водночас жоден із них не є абсолютно надійним, оскільки синтетичний контент може бути частково зміненим, проходити постобробку або спеціально створюватися з урахуванням уникнення детекції.

Виявлення шкідливого контенту, створеного або модифікованого за допомогою штучного інтелекту, є важливим напрямом забезпечення інформаційної безпеки та цифрової довіри. Розвиток генеративних технологій уможливив створення реалістичних текстів, зображень, аудіо та відео, що поряд із корисними застосуваннями підвищує ризики дезінформації, цифрового шахрайства, маніпуляцій і поширення небезпечних матеріалів. Жоден окремий технічний метод не є повним розв'язанням цієї проблеми, тому ефективність виявлення залежить від контексту використання, якості реалізації та належного контролю[3].

Першу групу становлять методи встановлення походження контенту. Вони базуються на записі та подальшій перевірці відомостей про джерело й історію цифрового об'єкта. До таких засобів належать цифрові водяні знаки, метадані та цифрові підписи. Цифрове водяне маркування передбачає вбудовування службової інформації безпосередньо у контент - зображення, відео, аудіо або текст. У водяному знаку можуть міститися відомості про походження, час створення чи інші характеристики матеріалу. Метадані, своєю чергою, дають змогу фіксувати автора, параметри створення та редагування, а цифрові підписи - перевіряти цілісність даних. Усе це допомагає підвищувати автентичність і достовірність цифрового контенту.

Водночас такі підходи мають обмеження. Метадані часто видаляються під час пересилання файлів або публікації на платформах, а водяні знаки можуть бути пошкоджені, видалені або підроблені. Крім того, наявність даних про походження ще не гарантує істинності змісту, а лише дає додатковий сигнал про джерело матеріалу.

Другу групу становлять автоматизовані методи контентного аналізу. Вони намагаються визначити, чи є контент синтетичним або модифікованим, аналізуючи сліди, що залишаються після генерації чи обробки. Для зображень і відео це можуть бути візуальні аномалії, неузгодженість освітлення, тіней і віддзеркалень, неприродні текстури, а також статистичні та форензічні сигнали, зокрема просторово-частотні характеристики. Для відео також аналізують рухи обличчя, моргання та інші поведінкові невідповідності.

Таблиця 1 – Інструменти та технології виявлення шкідливого контенту, створеного ШІ.

№	Назва інструмента	Домен	Модальність	Опис
1	ActiveFence	Виявлення	Текст, зображення, відео, аудіо	Інструмент виявлення AI-контенту для модерації та виявлення зловживань
2	AISEO	Виявлення	Текст	Визначає текст, написаний людиною, і текст, згенерований ШІ
3	Attestiv	Виявлення	Зображення, відео, документи	Валідація медіа та виявлення шахрайства
4	AudioSeal	Водяні знаки	Аудіо	Заснований на машинному навчанні інструмент для водяного маркування згенерованого мовлення, розроблений для ефективності
5	Azure AI Content Safety	Модерація контенту	Текст, зображення	Виявляє шкідливий користувацький і згенерований ШІ контент у застосунках і сервісах
6	Content Authenticity Initiative Signing Toolkit	Метадані	Зображення, відео, аудіо, документи	Інструмент для забезпечення автентичності контенту та відстеження його походження

Для аудіо виявлення ґрунтується на пошуку неприродних інтонацій, дефектів спектра, часових і частотних невідповідностей. Для тексту використовують інші характеристики: передбачуваність мовних конструкцій, стилістичну одноманітність, частоту повторів, а також показники як складність та нерівномірність тексту. Такі методи дають змогу знаходити ознаки штучної генерації безпосередньо у самому змісті матеріалу.

Третю групу формують підходи із залученням людини. У таких системах модератори, аналітики, фактчекери або інші фахівці перевіряють сумнівні результати та приймають рішення з урахуванням контексту. Це особливо важливо тоді, коли контент є частково синтетичним, коли потрібно оцінити спосіб його поширення, намір автора або наслідки можливої помилки. Поєднання автоматизованого аналізу з людською перевіркою зазвичай підвищує надійність оцінювання, хоча потребує більше часу та ресурсів[1].

Окреме місце в протидії шкідливому ШІ-контенту посідають превентивні та модераційні механізми. До них належать фільтрація навчальних даних, вхідних запитів і вихідного контенту, використання класифікаторів небезпечного вмісту, блок-листів ключових слів, а також хешування вже виявлених шкідливих матеріалів для їх повторного розпізнавання на різних платформах. Такі методи допомагають не лише виявляти небезпечний вміст, а й обмежувати його подальше поширення. Важливим при цьому є правильне налаштування порогів спрацювання, оскільки баланс між хибнопозитивними та хибнонегативними результатами безпосередньо впливає на ефективність системи.

Ще одним важливим аспектом є тестування та оцінювання систем виявлення. Для цього потрібно враховувати не лише формальні метрики точності, а й контекст використання, відтворюваність результатів, характер можливих помилок і стійкість до атак. Для систем із залученням людини важливо також оцінювати час виконання перевірки та складність роботи для користувача [4].

Таким чином, методи виявлення шкідливого контенту, створеного ШІ, охоплюють відстеження походження контенту, автоматизований аналіз його ознак та перевірку за участю людини. Найперспективнішим є не окремий метод, а їх поєднання: метадані й водяні знаки дають сигнал про походження, автоматизовані моделі масштабують аналіз, а людина допомагає врахувати контекст і зменшити ризик помилок. Водночас, жоден із цих підходів не є універсальним, тому подальший розвиток теми пов’язаний із удосконаленням

детекторів для різних мов і модальностей, виявленням частково синтетичного контенту, підвищенням стійкості до атак та розбудовою стандартів цифрової прозорості й міжплатформної взаємодії.

Список використаних джерел

1. «Шкідливий інтелект: Сценарії злочинного використання ШІ». URL:
2. https://my-itspecialist.com/adversarial-intelligence-malicious-scenarios-of-ai-usage (дата звернення: 18.03.2026).
3. «Reducing Risks Posed by Synthetic Content An Overview of Technical Approaches to Digital Content Transparency». URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-4.pdf (дата звернення: 18.03.2026).
4. «Виявлення контенту, створеного за допомогою ШІ». URL: https://gijn.org/ua/resurs-ua/viavlenna-kontentu-stvorenogo-za-dopomogou-si-posibnik-dla-zurnalistiv/ (дата звернення: 18.03.2026).
5. «Маркування ШІ-згенерованого контенту: як уряди та компанії посилюють прозорість використання штучного інтелекту в медіа та соцмережах». URL: https://dslua.org/publications/markuvannia-shi-zghenerovanoho-kontentu-iak-uriady-ta-kompanii-posyliuiut-prozorist-vykorystannia-shtuchnoho-intelektu-v-media-ta-sotsmerezhakh/ (дата звернення: 18.03.2026).