

Модуль: Убезпечення від неправдивих повідомлень



Практична вправа:

«Інструменти виявлення неправдивих повідомлень»



Навчальна мета заняття: навчитися перевіряти окремі відомості в мережі Інтернет на достовірність.



Час проведення: 1 год.

Місце проведення: комп'ютерний клас.



Устаткування:

персональний комп'ютер (ПК) зі встановленою операційною системою Windows 7 або вище та доступом до мережі Інтернет.

Порядок проведення заняття

Для перевірки повідомлень та інших матеріалів на предмет їх актуальності та достовірності можуть бути використані різні аналітичні методи. Для полегшення цього процесу також варто застосовувати низку технічних рішень.

Розширення **Fake Profile Detector (Deepfake, GAN)** – інструмент, який працює в браузері на базі Chrome або Chromium і дає змогу ідентифікувати зображення, створені з використанням штучного інтелекту (рис. 1). Саме тому, якщо обличчя реальної людини додано до іншого зображення, воно теж визначатиметься як справжнє. Розширення доступне за посиланням <https://chrome.google.com/webstore/detail/fake-profile-detector-dee/jbpcgcnhnmjmajjkgdaogpgfebnokpcc/related?hl=en-US>.

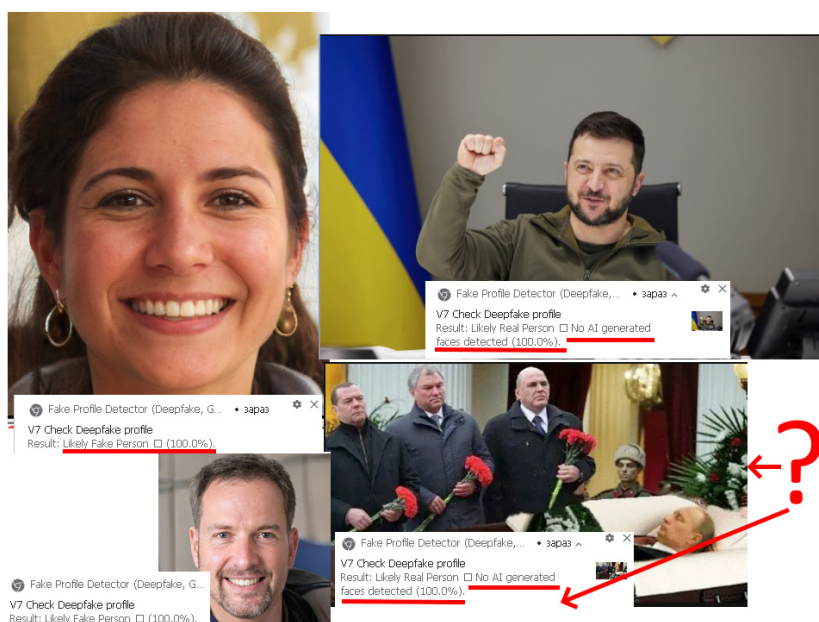


Рис. 1. Результат роботи розширення з виявлення дипфейків

Сервіси Sensity (platform.sensity.ai/login?redirect=%2Fdeepfake-detection), **Advanced AI Image Detector** (contentatscale.ai/ai-image-detector/), **Advanced AI Image Detector** (contentatscale.ai/ai-image-detector/), **Maybe's AI Art Detector** (huggingface.co/spaces/umm-maybe/AI-image-detector), **Is your image GAN generated?** (gan-detector-mayachitra.azurewebsites.net/), **AI or Not** (aiornot.optic.xyz/#home) та Illuminarty (app.illuminarty.ai/#/) виконують подібну функцію (рис. 2).

Сервіс **Forensically** (29a.ch/photo-forensics) дозволяє з використанням нескладних методів аналізу спробувати зрозуміти, чи вносилися до зображення якісь зміни.

Розширення **Deepfake Detection** (<https://github.com/deep2universe/DeepFakeChrome>) призначено для виявлення DeepFake-відео в сервісі YouTube. Також для ідентифікації фейкового відео призначено сервіс **Scan & Detect Deepfake Videos** (<https://scanner.deepware.ai>).

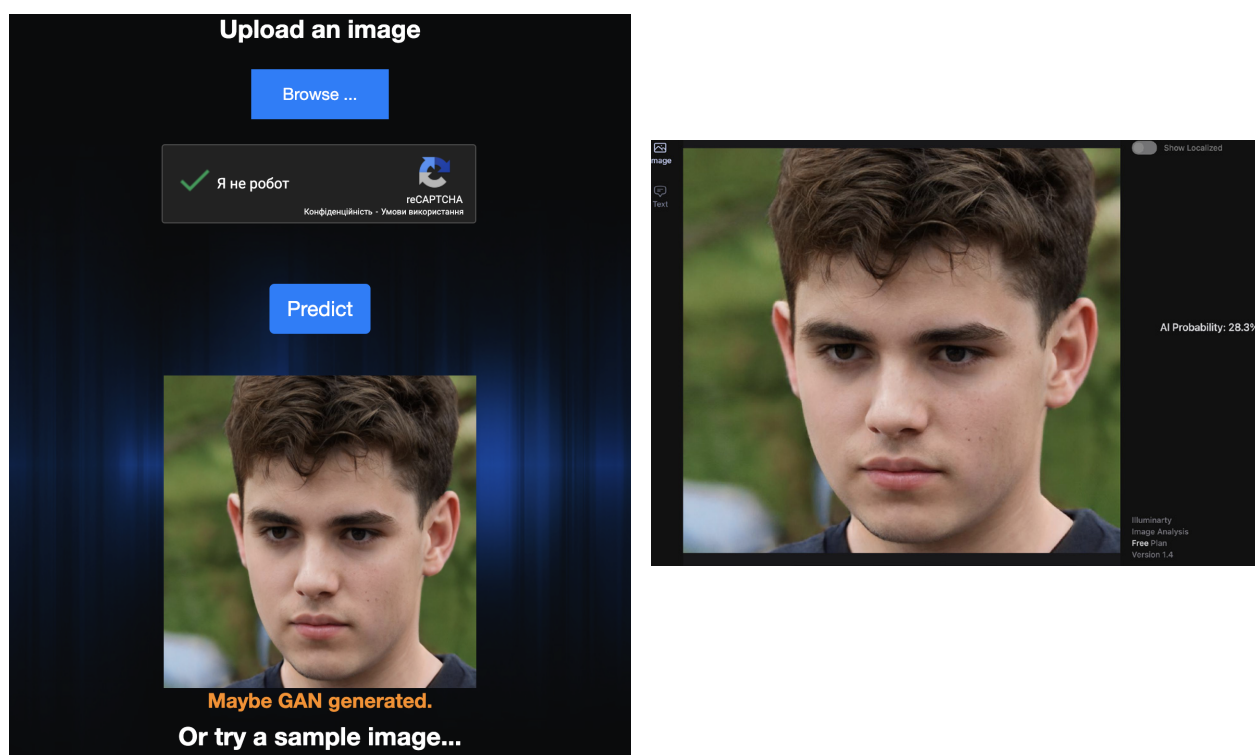


Рис. 2. Результат роботи сервісів з виявлення зображень, згенерованих за допомогою системи штучного інтелекту

Зображення можуть не мати ознак маніпуляцій, проте використовуватися у неправдивих повідомленнях у різних контекстах. Для того, щоб знайти першоджерело відповідних малюнків, можна використовувати розширення Who stole my pictures (рис. 3).

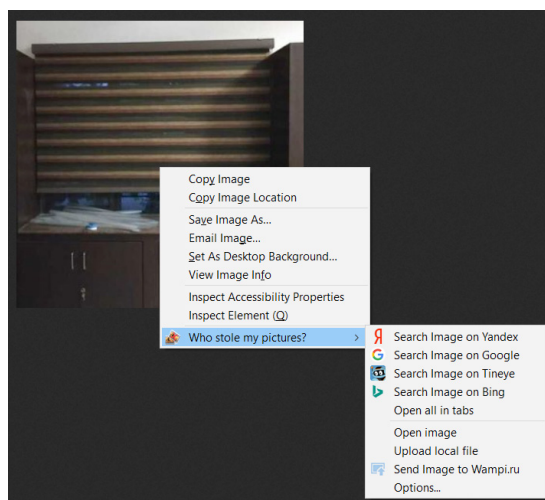


Рис. 3. Використання розширення для пошуку зображень

Завантажити описане розширення можна за адресами:

- ▶ для браузера Chrome (<https://chrome.google.com/webstore/detail/who-stole-my-pictures/mcdbnfhkikiofkkioppioekloflmaibd>);
- ▶ для браузера Firefox (<https://addons.mozilla.org/ru/firefox/addon/who-stole-my-pictures/>).

Подібними до описаного розширення є «**Fake news debunker by InVID & WeVerify**» (<https://www.invid-project.eu/tools-and-services/invid-verification-plugin/>), RevEye reverse image search (<https://chrome.google.com/webstore/detail/reveye-reverse-image-sear/keaaccljhehbbapnphnmpiklalfhelgf>), **Dessant Search by Image** (<https://github.com/dessant/search-by-image>).

Для аналізу тексту на предмет його створення за допомогою штучного інтелекту можна використати ресурс **GPTZero** (github.com/BurhanUITayyab/GPTZero, gptzero.me), **AI Detector** (contentatscale.ai/ai-content-detector/), **AI Content Detector** (copyleaks.com/ai-content-detector).

Створіть кілька зображень за допомогою ресурсу thispersondoesnotexist.com або generated.photos/faces та розмістіть їх у своєму профілі в соціальній мережі.

За допомогою описаних програмних продуктів спробуйте перевірити завантажені вами зображення на справжність. Спробуйте здійснити перевірку із зображеннями, опрацьованими Телеграм-ботом **@Pix2MixV2Bot**. Для різних фотознімків самостійно опрацюйте сервіс **Am I Real?** (<https://seintpl.github.io/AmIReal/>).

Перейдіть за посиланням olx.com та знайдіть кілька оголошень з продажу ігрових приставок. За допомогою розширення «Who stole my pictures» або «Fake news debunker by InVID & WeVerify» перевірте, чи немає серед фотографій у знайдених оголошеннях зображень, завантажених з інших сайтів.

Згенеруйте текст за допомогою однієї з систем генеративного штучного інтелекту, перевірте його за допомогою gptzero.me.