Reality Defender

\ Case Study

How a Tier-One Social Media Company Rooted Out Fake Users With Reality Defender

\ Case Study

How a Tier-One Social Media Company Rooted Out Fake Users With Reality Defender

Note: All names and PII have been removed to protect confidentiality.

Our client, one of the most-used social media platforms in existence, has hundreds of millions of accounts from individuals and brands alike posting at all hours from around the world. Due to the popularity of their platform and its relevancy beyond the internet (particularly in culture, news, and politics), our client dedicates a high double-digit percentages of its resources to combatting and preventing fraud.

\ 01

The Challenges

Fraud on our client’s social media platform appears in a myriad of ways. Individual account fraud sees unorganized or loosely-organized bad actors creating multiple “sock puppet” accounts for the purpose of artificially elevating a keyword or topic, all while posting under the guise of a real person who, in actuality, does not exist. The client’s platform also sees concentrated organized efforts from teams of bad actors — sometimes working on behalf of a nation-state or political entity — actioning hundreds of thousands of “bot” accounts (if not millions) to spread disinformation and otherwise harmful content.

Though the text used in the posts from these accounts is often created by humans, the images and visual content used to legitimize the fake account is often a visual deepfake, an artificial image or video generated by artificial intelligence (AI) for the purpose of deceiving fellow users. This visual content is often created using a myriad of models and tools, from off-the-shelf and open-source generative adversarial networks (GANs) to more advanced proprietary methods. The methods and approaches of creating these visual deepfakes number in the hundreds, with new and more advanced models being introduced weekly.

The majority of people can detect low quality deepfakes on sight and sound alone. Sound may not match up to a person speaking in a video, or the voice may not be in line with the speaker. Visual glitches and very obvious (or limited) facial twitches or expressions may appear that allow the viewer to discern the difference between a real speaker and an AI-generated mess.

> High-quality deepfakes are a problem for even the most astute viewers, especially on our client's platform.

A video or image generated using a high-quality deepfake method can often only be detected by deepfake detection platforms (like Reality Defender). This leaves viewers susceptible to manipulation, which can include political propaganda, demeaning marks made under the guise of a specific person, or the creation of entirely fictitious people purporting to be real.

Our client hired Reality Defender to run the profile photos of a select group of users on their platform that were believed to be bots and fake accounts. Using a deepfake or generated image for a user profile does not mean the user itself is guaranteed to be engaged in intentional duplicity.

(As of this writing, creating deepfaked avatars using images in the Lesna AI app is popular amongst millions of real users on all social networks.) That said, these select users routinely posted misinformation, had suspicious network activity, and made posts similar to other confirmed bot accounts. Scanning their profile pictures (all which featured human faces) for authenticity would only further confirm their intent to act maliciously on our client’s social media platform.

\ 02

The Solutions

Our client used the Reality Defender web app to scan 25 individual profile pictures, receiving the results in nanoseconds. Though the Reality Defender API allows any platform (including that of our client) to continuously and automatically bulk scan a near-infinite number of deepfakes, our client's small sample size of purported fakes required no API implementation.

> Our scans use an ensemble of models to not only test deepfakes against multiple existing and widely-used models, but models yet to be widely used or even seen elsewhere. By using this ensemble approach, the Reality Defender platform is able to better detect more nuances in deepfakes with more accurate results.

Upon uploading these profile pictures, our client received the results in seconds. Out of 25 images, 23 were suspected as fake, while two were "suspicious."

For each instance of analysis, our detector outputs a probability score, with higher scores denoting a higher chance that the image in question is fake or manipulated. By reporting a probability score as opposed to a simple yes or no, our systems informed the client of the amount of confidence we have in our decision on whether an image is fake or real.

Each scan generated an individual PDF report for the client detailing which scan-ning models were used, as well as a CSV containing the detailed scanning results. (It is worth noting that we also offer email alerts on detected fakes, which also detail and deliver the results in real time.)

\ 03

The Advantages

Our client was informed that their decisions on removing these accounts were based on the exported results.

> By showing the probability of the accounts' avatar authenticity (or lack thereof), our client had another datapoint available to use as evidence in the removal of these users, as users spreading disinformation, parroting other accounts taking similar actions, and acting under the guise of a fictitious entity with a convincing-enough avatar violated the company's terms of use/service.

During our discussions with our client, we mutually agreed that use of a deepfake as an avatar alone does not constitute intent to deceive or fraud both the platform and its users. Nonetheless, posting misinformation and content against the terms of use/service in concert with having a deepfake as an avatar showed intent to act in bad faith and against the intended use of the service.

\ 04

Conclusions

Our work with this client continued into endeavors with several other companies and platforms led by former members of the now-dispersed corresponding team from the aforementioned client. By utilizing Reality Defender's detection models, teams at platforms large and small are able to catch bad actors faster, make better informed decisions on punitive measures taken against them, and create an overall safer environment for users — one where sophisticated attacks featuring manipulated media may never see the light of day.

Reality
Defender

Reality Defender

realitydefender.com
ask@realitydefender.com