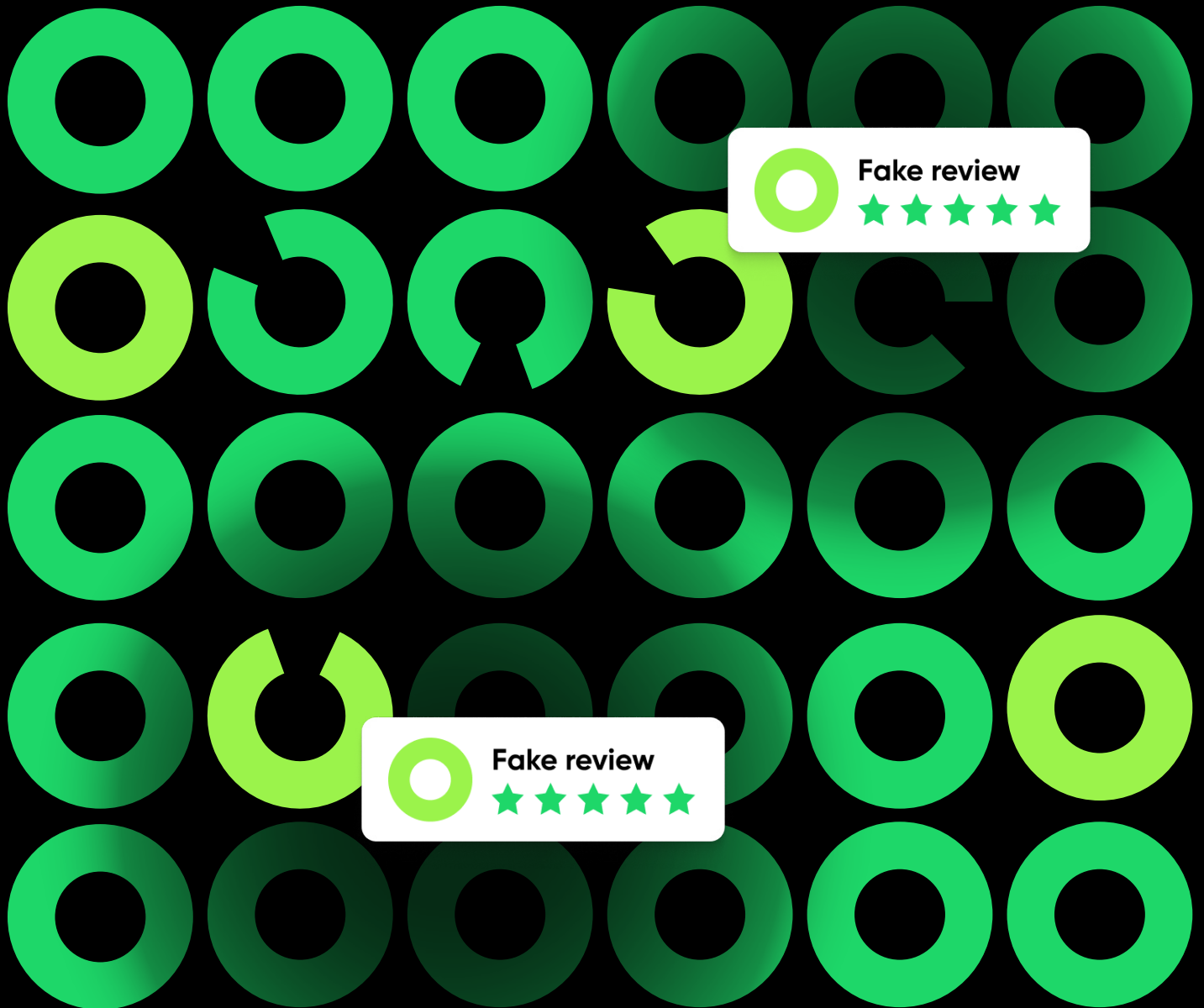




How to Tackle Fake Reviews in a Generative AI World



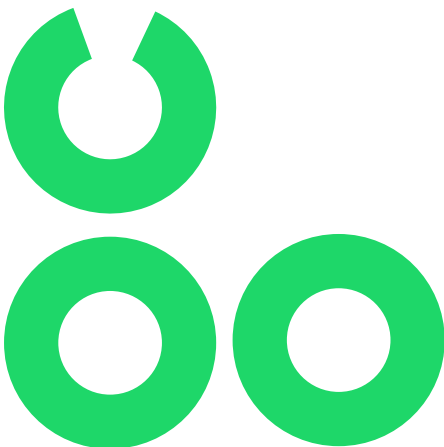
The recent explosion of generative AI tools, such as ChatGPT, Bard and Dall-E, has grabbed the world's attention. People are experimenting with them to see how good they are, or trying to find their limitations, and businesses are exploring how they can be used to improve productivity and enhance the user experience.

Despite being in the early days of scaling, these models are impressing users with their capabilities and are being adopted at an alarming rate. Whether it's to augment marketing creativity, improve customer service chatbots or write snippets of code, generative AI tools are already being used across many disciplines and a variety of industries. McKinsey recently estimated that it could add the equivalent of \$2.6 trillion to \$4.4 trillion annually in value to the global economy.

In the first five days of ChatGPT's release, over one million users logged onto the platform to try it out for themselves. In just four months, its total monthly website visits went from 266 million in December 2022, to 1.8 billion in April 2023!

And while many of us are excited by gen AI tools' applications, use cases and outputs, bad actors are finding ways to exploit them. Just as gen AI is creating new strategic paths and ways to build bonafide businesses, fraudsters are learning what they can do with this technology too. And it's unlikely red teaming will fully prevent it.

The problem is that the benefits gen AI tools offer are being weaponized by bad actors - primarily the ability to be more productive and increase the scale and speed at which they operate. From creating fake accounts on dating sites, to writing fake reviews, gen AI just made scammers lives a lot easier and the authenticity of content harder to determine.



Minimum effort, maximum reward

One of the biggest problems with these tools in fraudsters' hands is that with minimal effort, they can perpetrate their scams at a higher velocity than ever before. To illustrate, we can easily create a fictitious childcare service, generate a logo, and, with the right prompts, instruct a chatbot to write credible-sounding website copy and 50 positive reviews about a childminding business called 'Little Acorns'. In just a few minutes, our new business has 50 unique fake positive reviews to build some confidence for potential customers:

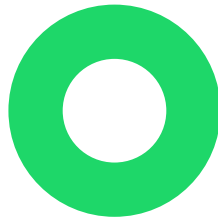


Little Acorns is an exceptional childminding service that I highly recommend. The caregivers are incredibly warm and nurturing, creating a welcoming environment for children. My child always comes home with a smile on their face, which is a testament to the love and care they receive at Little Acorns.



Generated by ChatGPT for a fictitious childcare service

The issue with this, of course, is that parents will come across reviews such as these when trying to choose childcare suitable for their needs, unaware they have been fabricated. Their decision could then be influenced by fake information, whereby they could leave their children with unvetted sitters in unsafe settings, with potentially detrimental effects.



So, what can platforms do to help keep their users safe and ensure they're providing honest and genuine reviews?






Fight AI with AI

As fraudsters use increasingly sophisticated AI tools to run their scams, so too must platforms to protect their customers from them. Fortunately, AI is designed to tackle high volumes very efficiently. As you scale, it becomes impossible to manually check every review on your platform. For example, Pasabi's technology combines AI, ML and behavioral analytics to process all your user profiles and review data and run it through a series of sophisticated checks to look for signals of fabrication or other suspicious activity.

Focus on account behaviors, not review content

The ability of large language models to generate images, videos, content and music from text inputs means authenticity is much harder to determine if focusing on outputs alone. Take, for example, the German [artist](#) who earlier this year won the Sony World Photography Award. He refused the award, revealing his submission was actually generated by AI.

This is where the behavioral signals around reviews and user accounts, rather than review content, yield more insights:

-  Where did the review come from?
-  Who posted it?
-  Who does it benefit?
-  When was it posted?
-  Who is the reviewer connected to?

Two examples of behavioral metrics that Pasabi uses as part of our model are geolocation and data spikes:



Location - one cue about authenticity

If someone has recently written reviews for a physiotherapy clinic in Portland, Oregon and café in Austin, Texas but their IP address indicates their location is 3,000 miles away, it seems highly suspicious and so is very likely to be a fake review. Pasabi's technology analyzes the geographic locations of reviewers and the businesses/products they are reviewing to start building a picture of what's happening on the platform.



Are there unusual spikes in the review data?

We also look at the timing of reviews posted to see if there are spikes in the review data which can indicate suspicious review behavior, such as incentivised schemes or review seller activity. Bad actors often post in patterns that we can spot over time.

Cluster technology reveals insightful patterns

In our experience, the worst offenders tend to work together. This could be a series of bots (automated systems using fake accounts) being deployed alongside real users, or a group of individuals working as one organization. Pasabi's technology analyzes all the data points we have collected, filters out the unconnected posts, accounts and businesses and looks for patterns in the dataset. Our cluster detection technology reveals these patterns, enabling us to identify accounts connected by suspicious behavior.

However, this alone does not provide the necessary evidence to act at scale. We combine the data with risk scoring, analyze reputation data and apply classifications to identify the type of review fraud at play. This can include positively-biased or negatively-biased fake reviews, incentivised reviews, gated reviews (holding back the negative reviews) or 'paid for' reviews.

This, in turn, allows you to customize your responses based on the type of review behavior you are seeing on your platform and build automated workflows to reduce your workload - for example, automatic removal, warning notices or cease & desist letters.

Connect suspicious behaviors to find review sellers

Review sellers or 'brokers' are a thriving industry on third party platforms, including social media apps. Brokers ask customers to write misleading or inflated reviews in exchange for free products, money or other incentives. They sell these fake reviews to unscrupulous businesses looking to boost their ratings.



Groups that solicit or encourage fake reviews violate our policies and are removed. We are working with Amazon on this matter and will continue to partner across the industry to address spam and fake reviews.

Meta



In a recent [article](#), Amazon stated they were a real problem for their platform, and that they have taken legal actions against 3 [major fake review brokers](#) targeting Amazon customers in the US, UK, Germany, France, Italy and Spain.

From working closely with our customers, such as Trustpilot, training our algorithms, and following our behavioral approach, we can identify connected accounts that are engaged in review seller activity. Focusing on the behaviors around the reviews being posted, not the review content alone, is key to their detection and gives our customers the confidence to enforce with automation at scale.



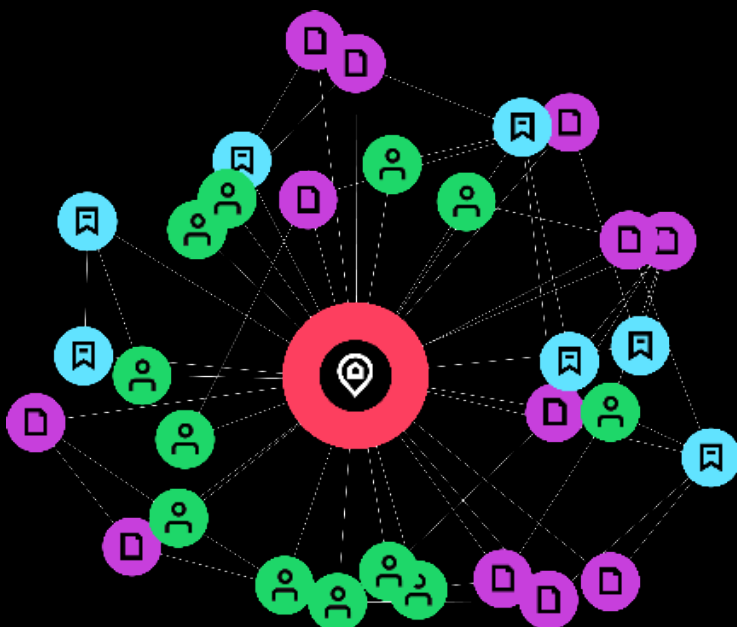
Pasabi's technology provides Trustpilot with the insight our legal, fraud and investigation teams need to strengthen existing capabilities, gather evidence against misbehaving companies and users, and drive our campaign against review sellers. Deploying Pasabi's technology to support our existing approach to manage this ongoing and growing challenge has been instrumental in delivering to consumers a clean, authentic and trusted user experience.

Being able to identify the scale of the issue through AI analytics software has made it possible for our legal team to be even more proactive in mitigating future risk, removing thousands of posts and companies.







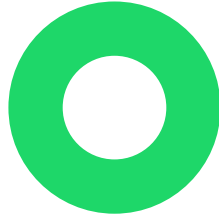
Trustpilot case study
Anoop Joshi, VP, Legal & Platform Integrity
Trustpilot

To help explain the process, the following is a simplified example of a cluster at the center of a behavioral detection model. We see patterns and associations between profiles and reviews. Multiple profiles share a single IP address. The device fingerprints are identifiers relating to our browsers and devices that allow internet users to be tracked and identified - which, in this example, show multiple profiles have them in common. The pattern of relatively few reviews to profiles is likely to indicate bot behavior rather than human reviewers.



Example showing a bot-driven reviews pattern

-  Profiles
-  Reviews
-  IP address
-  Device fingerprints



So, what does all this mean for a future powered by generative AI?

Gen AI tools promise greater advancements

No one can deny that AI-generated content currently mimics human writing very effectively. Well, these tools are only going to keep getting better - looking at how much they've improved in a short space of time, it's hard to predict where we'll be in even 6 months' time! What is certain, however, is that bad actors will continue to exploit them in ways we're not even aware of yet. We need to be mindful of this when developing fraud detection strategies.

Behavioral signals are a better indicator of fake reviews

While the quality of outputs from gen AI tools will likely improve, the behaviors of humans using the tools will be slower to change. Bad actors will continue to unwittingly leave behind digital fingerprints and cues that we can track to detect and stop them. So, a behavioral approach will be vital to continually tackle the challenge of fake reviews successfully.

Preserve authenticity to reinforce consumer trust

Authenticity matters when it comes to reviews. Without it reviews are worthless, and in the case of healthcare, childcare and pet services, potentially very damaging. As society and governments debate what guardrails or regulations are needed for generative AI tools, businesses can focus on using proven solutions that combine AI and behavioral signals, such as [Pasabi's Fake Review Detection](#), to keep their platforms free from harmful influence and as safe as possible for their users.

Contact us:



Chris Downie

As CEO, Chris provides the vision and direction for Pasabi. Passionate about product development and driven to produce great user experiences, Chris's expertise lies in strong people management and applying technology to solve real-world problems.

chris@pasabi.com



Pasabi's Trust & Safety platform enables marketplaces, platforms and online communities to detect multiple threat risks.

We use behavioral analytics, AI and clustering technology to identify fake accounts, fake reviews, fake listings, spam & scams and the bad actors behind them.

We focus on analyzing behavior over content to keep up with the evolving tactics of bad actors who use AI tools and other sophisticated techniques to scam and defraud users.

Pasabi, enabling trust and authenticity online.

www.pasabi.com