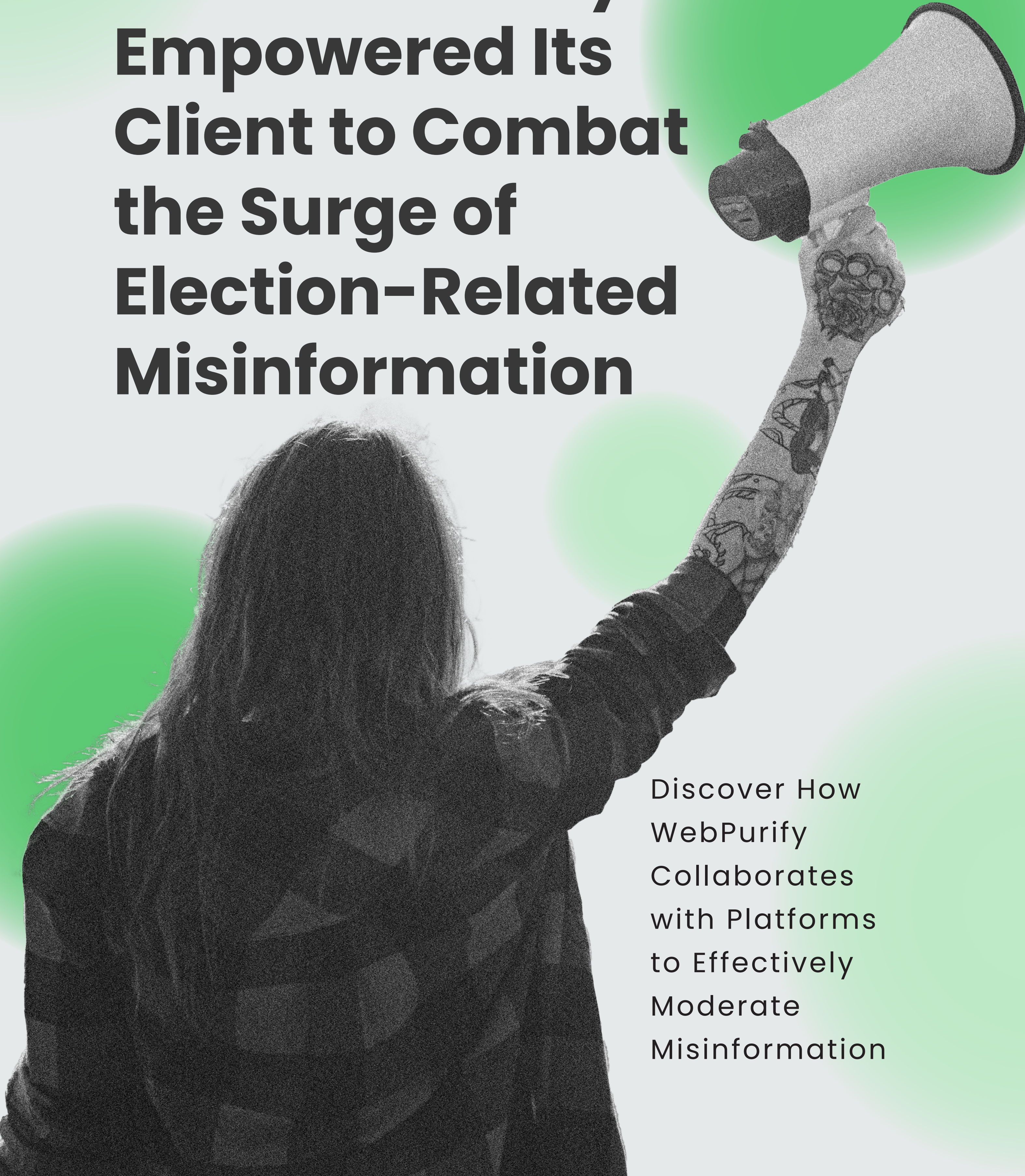
How WebPurify Empowered Its Client to Combat the Surge of Election-Related Misinformation



Discover How WebPurify Collaborates with Platforms to Effectively Moderate Misinformation

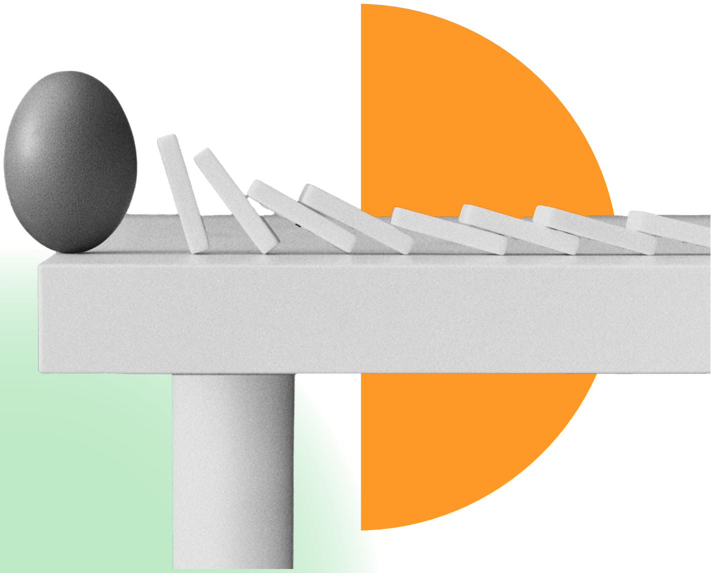
Moderating election-related misinformation might seem straightforward: find a false claim – like the fabricated story that US vice-presidential hopeful JD Vance confessed to intimate relations with a couch in Hillbilly Elegy – and remove it. But in reality, it's far more complex.

Platforms face an overwhelming volume of content to monitor, and AI, while powerful, can't tackle the problem alone. Effective moderation requires a blend of cutting-edge technology, skilled human moderators, and real-time fact-checking to keep up with the rapid evolution of misinformation tactics. These challenges are precisely why companies turn to WebPurify's expert moderation team and Trust & Safety consultancy for scalable, reliable solutions.

Read below to find out how one such company partnered with WebPurify to tackle their misinformation challenge. Note: To respect client confidentiality, the company's identity has been withheld.

The Challenge

2024 has been an unprecedented year for global elections, with more than 80 significant electoral events across continents – including major elections in the US, Europe, South Africa, Turkey, and India. In politically charged environments, misinformation thrives, so it's no surprise to see a surge of misleading content online this year. Such increases in misinformation jeopardize democratic processes, potentially influencing public opinion and voter decisions. Below are key challenges our client encountered in managing this complex issue:



Defining the Boundaries of Misinformation.

Where does misinformation begin, and how far does it go? Platforms must distinguish between trivial inaccuracies – like misreporting the color of Vice President Kamala Harris's pantsuit – and impactful misinformation that warrants action. While this example may seem inconsequential, these decisions are essential to establishing fair, enforceable policies.

Scaling Enforcement Mechanisms.

The sheer volume of user-generated content requires a balance between AI-driven solutions and human moderation. Determining when to deploy AI, what to prioritize for human review, and how best to train moderation teams are all ongoing challenges that affect response times and accuracy.

Keeping Pace with Emerging Trends.

Misinformation evolves constantly, with new tactics and themes appearing regularly. Monitoring these trends to prevent harmful narratives from taking root requires a proactive, adaptable approach.

The Solution

As Alexandra Popken, VP of Trust & Safety at WebPurify, explains, “This year, we collaborated with a prominent company to create a scalable framework for classifying misinformation online, ensuring alignment with industry best practices and applying a rigorous, objective methodology.” This client used WebPurify’s consulting service to deliver a number of outputs:

1 Defining the Framework

WebPurify defined the scope of enforceable misinformation by identifying key focus areas and informing a multi-level threshold to rank content risk. The team also conducted a cross-platform analysis and consulted with industry experts to ensure alignment with current standards. “Our client needed clarity on which types of misinformation to prioritize,” Alex explained. “We helped them identify key focus areas while excluding low-priority content. This ensured that resources were concentrated where the impact and stakes were the highest.”

2 Scaling Operational Processes

Despite operating with a specialized team, WebPurify successfully scaled its services to support the project by leveraging human expertise to validate and enhance the client’s AI-driven misinformation detection model. This included developing a streamlined workflow for moderators to produce high-quality, unbiased golden set data for AI training, designing a comprehensive fact-checking system, and implementing a process to monitor and address emerging misinformation trends.

3 Training

WebPurify played a crucial role in training the client’s in-house teams. “We trained their teams on the guidelines and positioned our own human moderation team as the ‘tiger team’ for quality control and surfacing new trends,” Alex explains. WebPurify continuously refined the guidelines using real-time data from content queues. “As moderators reviewed content, we leveraged their feedback and insights to fine-tune our policies, ensuring the misinformation detection process stayed adaptive and highly effective,” Alex added.

Examples of Misinformation Detected

WebPurify's human moderators have tackled a wide variety of misinformation.

Donald Trump Assassination Attempt

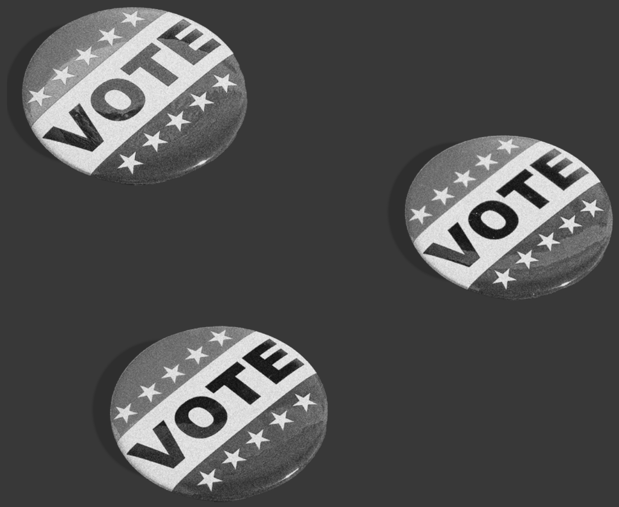
This case involved the assassination attempt on former US President Donald Trump by Thomas Matthew Crooks on July 13. Misinformation quickly spread online, with narratives claiming the event was staged by the Trump campaign, orchestrated by "the Democrats," or that President Joe Biden was responsible for an alleged lapse in Secret Service security. The WebPurify team promptly identified these claims as false and accurately labeled them as misinformation.

2024 Venezuelan Presidential Election

It was widely reported that Nicolás Maduro had manipulated the electoral process, and WebPurify's moderators categorized these reports as justified under the civic disruption category. By contrast, similarly contentious claims about the 2020 US Presidential Election were flagged as misinformation, given the absence of credible evidence supporting allegations of voter fraud. This distinction highlighted the crucial role of human moderators in evaluating the nuance and context of politically charged content.

"Misinformation isn't always black and white; many claims require *deep context* **and an understanding of nuance that algorithms alone can't provide. That's where human moderators, equipped with** *proper training and guidelines,* **can truly make a difference."**

ALEX POPKEN,
VP OF TRUST & SAFETY, WEBPURIFY



Results

WebPurify's involvement brought immediate, measurable improvements to the client's misinformation detection capabilities. The most critical outcome was the

99.24%

consensus rate achieved by WebPurify moderators in content classification. This means that two moderators independently categorized the same content with identical results. This high agreement rate provided the client with a strong validation signal, ensuring their data was reliable.

"This level of agreement is crucial when training AI models, as it ensures that the data fed into the system is accurate and consistent, ultimately enhancing the performance of the machine learning algorithms," Alex explains.

Additionally, WebPurify reported:

95% **quality on in-scope misinformation content:** This indicates that nearly all flagged content by moderators adhered to the predetermined guidelines for identifying in-scope misinformation.

99% **quality on out-of-scope content:** This indicates that content identified as 'not misinformation' was accurately categorized, minimizing false positives and preventing valuable content from being unnecessarily blocked.

Conclusion

In summary, WebPurify's collaboration with the client successfully addressed the complex issue of election-related misinformation by establishing a robust framework that combines AI capabilities with skilled human moderators. This partnership yielded a highly accurate moderation process, demonstrated by a 99.24% consensus rate and high-quality performance metrics. Through rigorous training, adaptive guidelines, and proactive trend monitoring, WebPurify helped the client navigate misinformation's evolving landscape, reinforcing the value of nuanced, human-led oversight in content moderation. This case underscores WebPurify's commitment to providing tailored, scalable solutions that enhance trust and transparency in digital spaces.

WebPurify

To learn more about how WebPurify's content moderation services can help your business navigate the challenges of misinformation, contact us today. WebPurify's expert team is here to support your brand's unique needs.

Discover the power of a comprehensive moderation solution:

- Customized content moderation strategies, including human-led fact-checking and misinformation detection.
- Scalable solutions that align with industry standards.
- Expertise in balancing technology with human oversight to ensure accuracy and transparency.

Get in touch with WebPurify to schedule a consultation and find out how we can help you safeguard your brand while maintaining trust in the digital ecosystem. Together, we can create a safer, more trustworthy environment for your platform.