El papel de la IA en la seguridad de correo electrónico de última generación de Hornetsecurity

La inteligencia artificial (IA) lleva años en boca de todos, aunque como disciplina académica existe desde 1956. Hoy en día, prácticamente todo el mundo conoce ChatGPT y sus múltiples variantes, usándolos tanto en su vida personal como profesional para generar texto, imágenes o incluso videos. Pero, ¿sabías que en Hornetsecurity llevamos años utilizando IA en nuestros productos? De hecho, es la base de muchas funciones diseñadas para mantener los mensajes maliciosos fuera de tu bandeja de entrada.

Entre las tecnologías que usamos en Hornetsecurity se encuentran las redes neuronales artificiales, que son muy útiles para **reconocer patrones, y el aprendizaje profundo**, que se apoya en varias capas de estas neuronas para mejorar la visión informática, el reconocimiento de voz, el procesamiento del lenguaje natural y la clasificación de imágenes.

El Machine Learning (ML), una de las bases de la IA, permite que los programas mejoren automáticamente en ciertas tareas. El Machine Learning no supervisado analiza flujos de datos para encontrar patrones sin necesidad de datos etiquetados. Por otro lado, el Machine Learning supervisado se entrena con datos etiquetados: si le enseñamos al modelo imágenes de plátanos y manzanas con sus respectivas etiquetas, será capaz de identificar correctamente nuevas imágenes que se le presenten. Finalmente, el aprendizaje por refuerzo funciona asignando recompensas por las respuestas correctas y penalizaciones por las incorrectas.

> Gracias a estas tecnologías, en Hornetsecurity conseguimos mantener el spam y las amenazas fuera de las bandejas de entrada, garantizando una experiencia de correo mucho más segura.

El papel de la IA en la seguridad de correo electrónico de última generación de Hornetsecurity

Como ocurre con muchos avances tecnológicos, los LLM no sólo ofrecen a los defensores herramientas más sofisticadas para protegerse, sino que también son utilizados por los atacantes para perfeccionar sus señuelos. Aunque es complicado obtener datos concretos sobre cómo los delincuentes emplean exactamente los LLM para mejorar sus correos electrónicos, se observa que ahora utilizan una gramática más precisa, traducciones a idiomas menos acostumbrados a amenazas de este tipo y asistencia de la IA para investigar objetivos o generar código de malware.

> Estas son algunas de las áreas donde aplicamos IA/ML para protegerte frente a amenazas avanzadas:

- **Análisis de intentos de fraude:** Verifica la autenticidad e integridad de los metadatos y el contenido de los correos.
- **Reconocimiento de suplantación de identidad:** Detecta y bloquea remitentes con identidades falsificadas.
- **Intent recognition system:** Alerting to content patterns that suggest malicious intent.
- **Sistema de reconocimiento de intenciones:** Identifica patrones de contenido que podrían revelar intenciones maliciosas.
- **Detección de espionaje:** Protege frente a ataques diseñados para robar información confidencial.
- **Identificación de hechos falsos:** analiza el contenido para detectar noticias o información falsificada, independientemente de su origen.
- **Detección de ataques dirigidos:** Identifica amenazas específicas contra personas con mayor riesgo.

Otra técnica muy útil es el agrupamiento de correos electrónicos mediante ML (conocido técnicamente como clustering). El phishing abarca un espectro de campañas. Por un lado, están los correos genéricos de bajo valor con objetivos mínimos, como "haga clic aquí para validar su dirección para su envío por FedEx". Estos deben enviarse en grandes cantidades para que los atacantes obtengan algún beneficio, ya que solo un pequeño porcentaje de los destinatarios cae en la trampa. Por otro lado, está el spear phishing, donde los atacantes investigan un poco sobre los objetivos y dedican esfuerzo a preparar correos personalizados, aumentando así las probabilidades de éxito. Por último, están los señuelos de correo electrónico altamente personalizados, a menudo dirigidos a ejecutivos de empresas. A este tipo de ataque se le conoce como phishing ejecutivo o "whaling". Estos correos se envían en pequeñas cantidades, pero están muy estudiados para maximizar su efectividad.

En las dos primeras categorías, identificar y clasificar automáticamente un correo individual puede ser un reto. Sin embargo, al analizar millones de correos electrónicos, empiezan a aparecer patrones que revelan claramente las campañas que los atacantes están utilizando. En este contexto, se emplean técnicas de Machine Learning no supervisado para combatir los ataques de phishing modernos. Estas técnicas agrupan correos electrónicos en función del contenido, el contexto, la dirección IP del remitente, el diseño del correo y muchos otros puntos de datos. Posteriormente, el sistema detecta valores atípicos dentro de estos grupos, lo que permite identificar posibles nuevas campañas de phishing.

Esto nos permite detectar rápidamente intentos de phishing sin depender exclusivamente de listas de reputación (que pueden tardar en actualizarse), heurísticas (que consumen muchos recursos al analizar cada correo) o firmas (que no siempre se actualizan con la rapidez que requieren las campañas de phishing modernas).

Aquí tienes algunos ejemplos de cómo funciona el clustering:

- Un aumento repentino de correos similares con ligeras variaciones en los nombres de dominio podría ser señal de una nueva campaña de phishing.
- Un incremento rápido de correos con contenido completamente distinto pero algunas características comunes, como nombres de archivos adjuntos o enlaces similares en las primeras líneas, también podría indicar una nueva campaña de phishing.

Este método también es útil cuando los atacantes extraen mensajes de sistemas infectados y los reutilizan para atacar a nuevas víctimas con pequeñas modificaciones. Esta técnica sigue siendo utilizada por varios botnets, como QakBot en 2023.

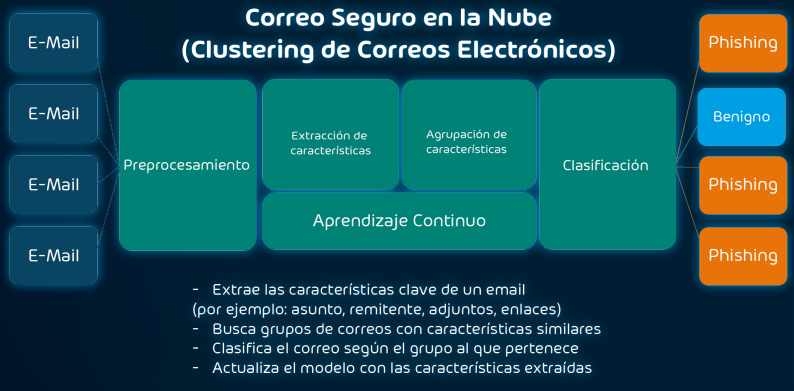
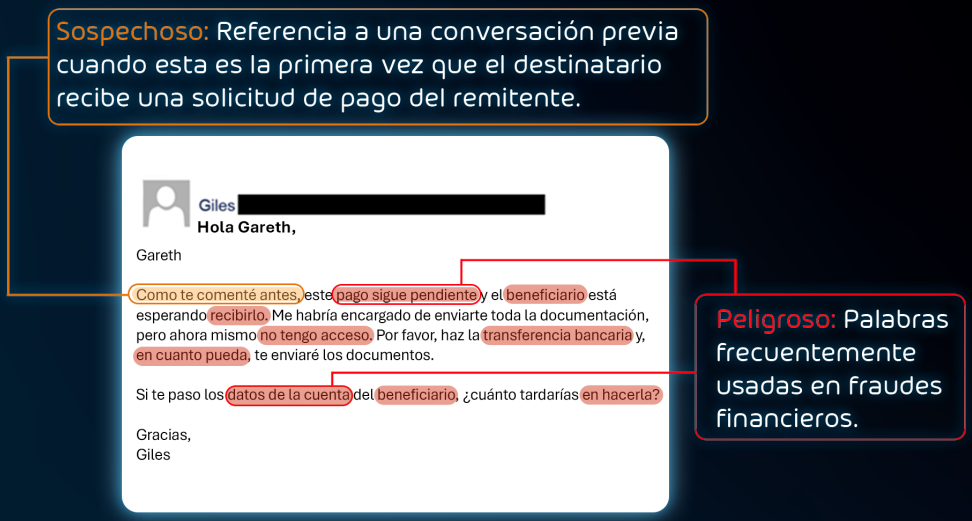


Diagrama de nuestra canalización de ML para identificar phishing mediante agrupamiento

Otra técnica útil de inteligencia artificial es el Procesamiento de Lenguajes Naturales (NLP), que analiza el texto de los correos electrónicos utilizando enfoques como el word embedding y el modelado de temas, extrayendo contexto y significado. Esto se puede combinar con análisis secuenciales o de series temporales para detectar patrones de comunicación anormales. Un ejemplo típico sería un directivo solicitando una transferencia financiera urgente, lo cual se marcaría como sospechoso si nunca antes había enviado un correo similar.



Ejemplo de cómo el NLP analiza texto y detecta señales

Como hemos mencionado antes, esto es un juego constante de adaptación: los delincuentes intentan sortear nuestras defensas mientras nosotros mejoramos continuamente nuestras detecciones para identificar nuevas variantes. Una de las ventajas de los modelos de ML es que aprenden con rapidez, y los mantenemos actualizados usando diversas fuentes de señales:

- El feedback de los usuarios finales son clave para identificar falsos positivos (cuando marcamos un correo como malicioso, pero no lo era) y falsos negativos (cuando no marcamos un correo sospechoso que en realidad sí lo era).
- Honeypots son simulaciones de objetivos o bandejas de correo que atraen ataques genéricos y dirigidos, y que utilizamos como datos de entrenamiento.

Esto permite que nuestros modelos se adapten continuamente al panorama de amenazas en constante evolución, y esta es la mejor forma de garantizar el éxito de una solución moderna como la nuestra.

Uno de los métodos que suelen usar los ciberdelincuentes es eliminar el contenido malicioso del correo electrónico y alojar la carga útil en un servidor externo, incluyendo únicamente un enlace en el mensaje. Para combatir esto, nuestra solución basada en IA, Secure Links, sustituye cada enlace en los correos entrantes por una versión que pasa por nuestro Secure Web Gateway (SWG).

Esta tecnología utiliza Machine Learning (ML) y aprendizaje profundo, combinando modelos supervisados y no supervisados, para analizar más de 47 características de los enlaces URL y los destinos de las páginas web. Esto incluye detectar comportamientos maliciosos, redirecciones sospechosas y técnicas de ofuscación. Además, aplica modelos de visión artificial para examinar imágenes, como logotipos de marcas, códigos QR y texto incrustado en imágenes. Gracias a este enfoque, podemos identificar contenido malicioso vinculado desde correos electrónicos, incluso en ataques rápidos y muy dirigidos. Otro método común es comprometer un sitio web sin modificar su contenido inicialmente, enviar una campaña de correos electrónicos y, una vez que estos han sido entregados, cargar el contenido malicioso en el sitio. Por eso, Secure Links analiza los enlaces de destino tanto en el momento de la entrega como en el momento en que el usuario hace clic en ellos.

Análisis de archivos adjuntos

A menudo, los atacantes mantienen sus cargas útiles fuera del texto del correo electrónico en sí, y no incluyen enlaces que se puedan escanear y, en su lugar, incluyen el contenido malicioso en un archivo adjunto. A diferencia de los correos electrónicos basados en texto que se pueden escanear con relativa facilidad, los archivos adjuntos vienen en muchos formatos de archivo diferentes y se pueden usar como arma de varias formas diferentes y, por supuesto, también pueden incluir enlaces a contenido malicioso.

Aquí, nuestro motor Sandbox, nuevamente basado en ML, abrirá los archivos adjuntos, identificará si son maliciosos y, si lo son, pondrá en cuarentena el correo electrónico. Este motor analizará el comportamiento del archivo, para ver si está tratando de identificar si se está ejecutando en un sandbox (una señal clara). También examina el sistema de archivos para ver si el archivo adjunto intenta crear archivos nuevos o modificar los existentes. El monitor de registro inspecciona el registro para ver si se crean valores inusuales, que a menudo se utilizan para persistir el malware después de reiniciar la PC. Y nuestro monitor de procesos detecta intentos de archivos PDF y Office maliciosos de iniciar procesos secundarios. También se inspecciona el tráfico de red del archivo adjunto, en busca de conexiones a servidores en Internet, nuevamente una táctica bastante sospechosa por parte de un documento adjunto. Finalmente, se inspecciona la memoria en el sandbox después de abrir el archivo adjunto, los tipos inusuales de acceso a la memoria son otra señal fuerte de malware.

En total, el motor de ML se basa en más de 500 indicadores en los archivos adjuntos en el motor Sandbox y los clasifica de manera fiable en archivos benignos o maliciosos rápidamente.

Análisis de correo saliente

Una de las soluciones más innovadoras de Hornetsecurity es el AI Recipient Validation (AIRV). Esta herramienta utiliza inteligencia artificial para analizar los patrones de comunicación por correo electrónico de cada usuario, aprendiendo de forma continua y detectando destinatarios no deseados, correos con información personal identificable (PII) y redacción inapropiada. Si se detecta algún problema, se marca para que el usuario lo revise, y las respuestas del usuario a estas advertencias se incorporan para mejorar en correos futuros.

AIRV alerta a los usuarios en los siguientes casos:

- Cuando se envía un correo a un destinatario que podría no ser deseado. (Por ejemplo, los honeypots, que son bandejas de entrada simuladas diseñadas para atraer y analizar ataques, se utilizan como datos de entrenamiento).
- Si falta un destinatario habitual de un grupo.
- Cuando se agrega o sustituye un usuario en un grupo existente.
- Al enviar correos a usuarios de organizaciones diferentes o a direcciones personales por primera vez.
- Si se responde a una lista de distribución muy amplia.
- Al enviar correos a destinatarios con los que no se ha tenido contacto previo.
- Cuando se envían correos con información sensible, como PII o datos de tarjetas de crédito.
- Si el contenido del correo incluye texto inapropiado.

Formación de usuarios mediante IA

Como ninguna solución para la higiene de correo electrónico es 100 % efectiva todo el tiempo, en ocasiones la última línea de defensa es el propio usuario, que debe estar alerta al recibir correos sospechosos. La solución Security Awareness Service de Hornetsecurity utiliza la IA como pieza clave, ajustando la formación a las necesidades de cada usuario. Se realizan campañas de phishing simuladas: los usuarios que hacen clic en enlaces o abren archivos adjuntos reciben más formación, mientras que los que no lo hacen no tienen que realizar entrenamientos adicionales. El motor de phishing selectivo basado en IA adapta el nivel de sofisticación de las simulaciones según ataques reales que hemos registrado, ayudando a los usuarios a detectar incluso las amenazas más avanzadas. Tal y como ocurre con los ataques reales, nuestros enlaces conducen a páginas de inicio de sesión falsas, los correos electrónicos forman parte de hilos auténticos y los archivos adjuntos incluyen macros „maliciosas".

La gran ventaja de este Security Awareness Service es que la IA lo gestiona de manera automática, liberando a los administradores de la necesidad de supervisar cada detalle de las simulaciones y la capacitación. Así, pueden enfocarse en tareas más estratégicas y productivas.

Conclusión

Hornetsecurity lleva años aprovechando la IA para afrontar distintos desafíos, afinando continuamente su enfoque para ofrecer una protección efectiva contra amenazas por correo electrónico y formando a los usuarios para detectar un ataque phishing con éxito.