The Role of AI in Hornetsecurity's Next Generation Email Security

Artificial intelligence (AI) has been in the spotlight for the last couple of years but has actually been around as an academic discipline since 1956. Everyone knows about ChatGPT and many, many people use it and all its cousins in their personal and business lives to generate text, images and even videos. But did you know that we here at Hornetsecurity have been using AI for many years in our products, and that it's underpinning many of the features keeping bad messages out of your email inbox?

Some notable technologies that we use at Hornetsecurity include artificial neural networks, useful for recognizing patterns, deep learning which uses several layers of those neurons which improve computer vision, speech recognition, natural language processing and image classification.

Machine Learning (ML) is a cornerstone of AI and refers to programs that can improve their performance on a task automatically. Unsupervised Machine Learning simply analyzes streams of data, looking for patterns, whereas Supervised Machine learning is based on tagged data (give the model 50 pictures of bananas labelled as such, and 50 apples and it'll then be able to accurately identify apples and bananas in new pictures it's shown). In reinforcement learning the agent is rewarded for good responses and punished for bad ones.

Here is how we use these technologies at Hornetsecurity to keep spam and threats out of email inboxes.

The Role of AI in Hornetsecurity's Next Generation Email Security

As with many technological advances, LLMs don't just provide defenders with additional tools with which to protect themselves, they're also used by attackers to improve their lures. It's difficult to gather hard data on exactly how criminals are using LLMs to improve their emails but anecdotally we're seeing better grammar, plus translations to different languages where societies may not be as used to email borne threats, as well as AI assistance with target research and malware code generation.

Here are some of the areas where we use AI / ML in Advanced Threat Protection:

- **Fraud attempt analysis:** Checks the authenticity and integrity of metadata and mail content.
- **Identity spoofing recognition:** Detection and blocking of forged sender identities.
- **Intent recognition system:** Alerting to content patterns that suggest malicious intent.
- **Spy-out detection:** Defense against attacks designed to obtain sensitive information.
- **Feign facts identification:** Identity-independent content analysis of news to identify falsified facts.
- **Targeted attack detection:** Detection of targeted attacks on individuals who are particularly at risk.

Another very useful technique is grouping (technically called clustering) emails using ML. There's a spectrum of phishing email campaigns, starting with low value, minimal targeting generic emails ("click here to validate your address for your FedEx delivery") which has to be sent in huge numbers to get a return for the attackers, because only a small percentage of recipients fall for them. Then there's spear phishing, where some research on the targets and effort goes into preparing the emails to make sure they're more likely to successfully trick recipients. Finally, there's highly customized email lures, often directed towards executives in a company, hence they're often termed executive phishing (or "whaling"), sent in low numbers but with well researched approaches.

For the first two categories, when you're looking at an individual email, automatically identifying and classifying it can be challenging, but when you're looking at millions of emails, patterns start to emerge that clearly show the individual campaigns the attackers rely on. Here we rely on unsupervised machine learning techniques to combat modern phishing attacks, clustering emails based on content, context, sender IP address, email layout and many more data points. Then the system identifies outliers in these clusters which can identify potential new phishing campaigns.

This gives us a quick way to spot phishing without relying exclusively on reputation lists (which can be slow to update), heuristics (which can be computationally expensive if you have to analyze every single email) and signatures (too slow to be updated with the speed of modern phishing campaigns).

Here are some examples of clustering:

» A sudden surge in similar emails with slight variations in domain names might signify a new phishing campaign.
» A rapid increase in emails with completely different content but a few similar characteristics, e.g., attachment names
or similar links in the first few lines of an email may also indicate a new phishing campaign.

This method is also useful where attackers harvest messages from infected systems and reuse them to attack new victims with slight alterations. This technique is still in use by various botnets (for example QakBot in 2023).

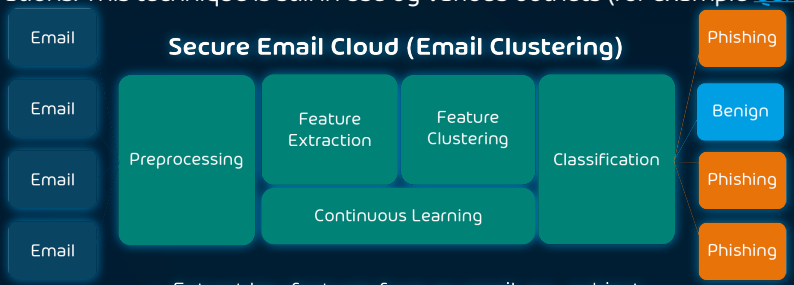
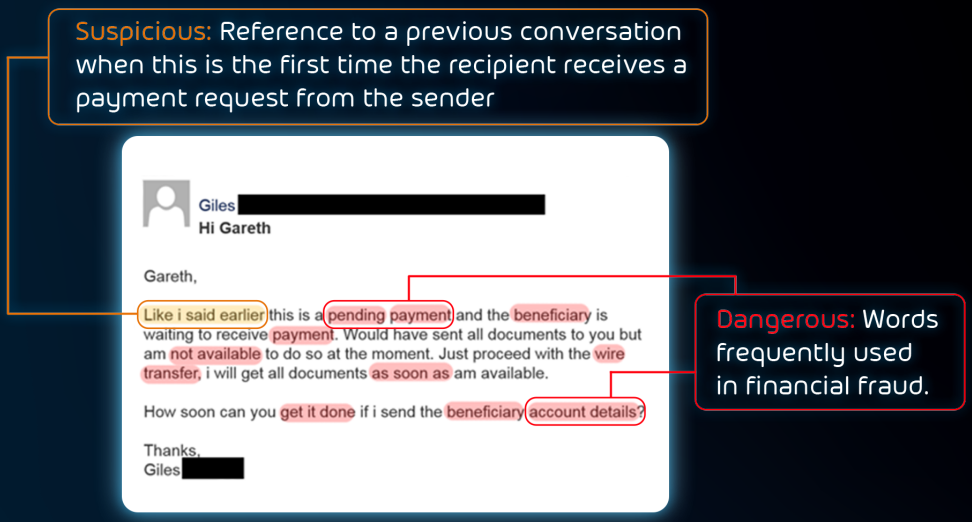


Diagram showing our ML pipeline to identify phishing based on clustering

Another useful AI technique is Natural Language Processing which analyses the text of emails, using approaches such as word embedding and topic modelling, deriving context and semantics. We can then combine this with sequential / time series analysis to detect abnormal communication patterns. A typical example here is an executive asking for a speedy financial transfer which will be flagged if they've never sent such an email in the past.



Example of NLP analyzing text and the signals detected

As previously hinted, this is a continuing game of adaptation by the criminals to bypass our defenses, and us continuously improving our detections to catch new variants. One strength of ML models is that they're good at learning, so we keep them up to date using various sources of signals:

- User feedback is valuable, relying on end users identifying false positives (where we flagged an email as malicious when in fact it wasn't) and false negatives (where we didn't flag an email, but it was actually suspicious).
- Honeypots are simulations of targets or email inboxes which attract both generic and targeted attacks which we use as training data.

This means our models are continually adapting to the rapidly changing threat landscape and this is the best way for a modern email hygiene solution like ours to succeed.

Another approach by criminals is removing the nasty content from the email by hosting the payload externally on a web server and only including a link to it. Malicious link detection is an integral part of our AI powered solution called Secure Links. We replace every link in incoming emails with a version that goes through our Secure Web Gateway (SWG).

This uses ML and deep learning, along with both supervised and unsupervised ML models to analyze 47+ characteristics of the URL links and the web page targets, looking for malicious behavior, URL redirects and obfuscation. It also uses computer vision models to analyze images, including brand logos and QR codes, as well as text content embedded in images. The net result is that we catch malicious content that's linked from emails, even when it's a quick and highly targeted attack. Another popular tactic is compromising a website but not altering the content, then sending out an email campaign and once the emails have been delivered, deploying the malicious payload. This is why Secure Links scan link targets both at time of delivery as well as at time of click.

Attachment scanning

Often attackers will keep their payloads out of the email text itself, and not include links that can be scanned and instead include the malicious content in an attachment. Unlike text-based emails that can be scanned relatively easily, attachments come in many different file formats and can be weaponized in various different ways and can of course also include links to malicious content. Often attackers will keep their payloads out of the email text itself, and not include links that can be scanned and instead include the malicious content in an attachment. Unlike text-based emails that can be scanned relatively easily, attachments come in many different file formats and can be weaponized in various different ways and can of course also include links to malicious content.

Here our Sandbox Engine, again based on ML, will open attached files, identify if they're malicious, and if they are, quarantine the email. This engine will look at the behavior of the file, seeing if it's trying to identify if it's running in a sandbox (a dead giveaway). It also looks at the file system to see if the attachment tries to create new files or alter existing ones. The registry monitor inspects the registry to see if unusual values are created, these are often used to persist malware after PC restarts. And our process monitor spots attempts by malicious PDF and Office files to start child processes. Network traffic by the attachment is also inspected, looking for connections to servers on the internet, again a fairly suspicious tactic by an attached document. Finally, memory in the sandbox is inspected after opening the attachment, unusual types of memory access are another strong signal of malware.

Altogether, the ML engine relies on over 500 indicators in file attachments in the Sandbox Engine and reliably sorts them into benign and malicious files quickly.

Outbound scanning

Another unique solution from Hornetsecurity that's powered by AI is our AI Recipient Validation (AIRV). This analyses each user's email communication patterns, continuously learning and detecting unintended recipients, emails containing Personal Identifiable Information (PII) and inappropriate wording. When issues are found they're flagged for the user, and user responses to these warnings are then incorporated for future emails.

AIRV warns users in the following scenarios:

- Sending an email to a potentially unintended recipient. Honeypots are simulations of targets or email inboxes which attract both generic and targeted attacks which we use as training data.
- An otherwise common recipient from a cohort is missing.
- A user is added or replaced in an existing cohort.
- Sending emails to users from different organizations or personal email addresses for the first time.
- Replying to a large distribution list.
- Sending an email to a recipient whom the user has no previous relationship with.
- Sending emails with sensitive information, such as PII or credit card data.
- Sending an email with inappropriate wording.

Training users using AI

As no email hygiene solution is 100% effective all the time, there are times when the last line of defense is the end user, being cautious when seeing a suspicious email. Hornetsecurity's Security Awareness Training solution has AI at its center, providing the right amount of training for each user. Simulated phishing campaigns are sent out, and users who click links or open attachments are provided with more training, whereas those that don't aren't bothered by training requests. The AI Spear Phishing engine also uses different levels of sophistication in simulations (based on real world attacks that we've captured), helping end users spot even the most advanced attacks. Just like real attacks our links lead to bogus login pages, emails are part of a thread and file attachments come with "malicious" macros.
The real strength of the Security Awareness Training service is that the AI automatically manages it, freeing administrators up to focus on more productive tasks instead of micromanaging phishing simulation campaigns and training assignments.

Conclusion

Hornetsecurity has been on "the AI train" for many years, using various flavors of tools for different challenges, fine tuning our approach to provide effective protection against email borne threats and training users to spot phishing.