Le rôle de l'IA dans la nouvelle génération de sécurité des emails de Hornetsecurity

L'intelligence artificielle (IA) est sous les feux de la rampe depuis quelques années, mais en réalité, elle existe en tant que discipline universitaire depuis 1956. Tout le monde connaît ChatGPT et de très nombreuses personnes s'en servent, ainsi que de tous ses cousins, dans leur vie personnelle et professionnelle pour générer du texte, des images et même des vidéos. Mais saviez-vous qu'ici, chez Hornetsecurity, nous utilisons l'IA depuis de nombreuses années dans nos produits, et qu'elle est à la base de nombreuses fonctionnalités qui empêchent les messages nuisibles d'arriver dans votre boîte aux lettres électronique ?

Parmi les technologies majeures que nous utilisons chez Hornetsecurity, citons les réseaux neuronaux artificiels, qui servent à reconnaître des modèles, le deep learning, qui emploie plusieurs couches de neurones et qui améliore la vision artificielle, la reconnaissance vocale, le traitement du langage naturel et la catégorisation d'images.

Le Machine Learning (ML), est une pierre angulaire de l'IA et se réfère à des programmes qui peuvent améliorer leur performance dans une tâche de manière automatique. L'apprentissage automatique non supervisé analyse simplement des flux de données, à la recherche de modèles, tandis que le Machine Learning supervisé est basé sur des données étiquetées (donnez au modèle 50 photos de bananes étiquetées comme telles, et 50 pommes, et il sera alors capable d'identifier avec précision les pommes et les bananes dans les nouvelles photos qu'on lui montrera). Dans l'apprentissage par renforcement(reinForcement learning), l'agent est récompensé pour ses bonnes réponses et puni pour les mauvaises.

> Voici comment Hornetsecurity utilise ces technologies pour empêcher les spams et les menaces de pénétrer dans les boîtes aux lettres électroniques.

Le rôle de l'IA dans la nouvelle génération de sécurité des emails de Hornetsecurity

Comme c'est le cas pour de nombreuses avancées technologiques, les LLM ne se contentent pas de fournir aux utilisateurs des outils supplémentaires pour se protéger, ils sont également utilisés par les attaquants pour améliorer leurs appâts. Il est difficile de recueillir des données précises sur la manière dont les criminels utilisent les LLM pour améliorer leurs courriels, mais nous constatons que la grammaire pour un fait, est déjà améliorée, que les traductions sont effectuées dans des langues différentes où les sociétés ne sont peut-être pas aussi habituées aux menaces véhiculées par courriel, et que l'IA aide à la recherche de cibles et à la génération de codes de logiciels malveillants.

> Voici quelques-uns des domaines dans lesquels nous utilisons l'IA / ML dans la protection contre les menaces avancées :

- Analyse des tentatives de fraude : **Vérifie l'authenticité et l'intégrité des métadonnées et du contenu des emais.**
- Reconnaissance de l'usurpation d'identité : **Détection et blocage des fausses identités**
- Système de reconnaissance des intentions : Alerte sur les modèles de contenu qui suggèrent une intention malveillante.
- Détection d'espionnage : Identifica patrones de contenido que podrían revelar intenciones maliciosas.
- Detección de espionaje: Protege frente a ataques diseñados para robar información confidencial.
- Identification des faux faits : **Analyse indépendante de l'identité du contenu des nouvelles pour identifier les faux faits.**
- Détection des attaques ciblées : Détection d'attaques ciblées sur des personnes particulièrement exposées.

Une autre technique très utile est le regroupement (techniquement appelé clustering) des courriels à l'aide du ML. Il existe tout un éventail de campagnes d'hameçonnage par courriel, à commencer par les courriels génériques de faible valeur et de ciblage minimal (« cliquez ici pour valider votre adresse pour votre livraison FedEx ») qui doivent être envoyés en très grand nombre pour être rentables pour les attaquants, car seul un petit pourcentage de destinataires tombe dans le panneau. Il y a ensuite le spear phishing, qui nécessite des recherches sur les cibles et des efforts pour préparer les courriels afin de s'assurer qu'ils ont plus de chances de tromper les destinataires. Enfin, il y a les courriels-leurres hautement personnalisés, souvent dirigés vers les cadres d'une entreprise, d'où leur nom de « phishing exécutif » (ou « whaling »), envoyés en petit nombre mais avec des approches bien étudiées.

En ce qui concerne les deux premières catégories, l'identification et la classification automatiques d'un courriel individuel peuvent s'avérer particulièrement difficiles, mais lorsque vous examinez des millions de courriels, des schémas commencent à se dessiner et montrent clairement les campagnes individuelles sur lesquelles s'appuient les auteurs d'attaques. Nous nous appuyons ici sur des techniques d'apprentissage automatique non supervisées pour lutter contre les attaques de phishing modernes, en regroupant les courriels en fonction du contenu, du contexte, de l'adresse IP de l'expéditeur, de la présentation des courriels et de nombreux autres points de données. Le système identifie ensuite les valeurs aberrantes dans ces groupes, ce qui permet d'identifier les nouvelles campagnes de hameçonnage potentielles.

Nous disposons ainsi d'un moyen rapide de repérer le phishing sans avoir recours exclusivement aux listes de réputation (dont la mise à jour peut être lente), à l'heuristique (qui peut être coûteuse en termes de calcul s'il faut analyser chaque courriel) et aux signatures (trop lentes pour être mises à jour à la vitesse des campagnes de hameçonnage modernes).

Voici quelques exemples de regroupement :

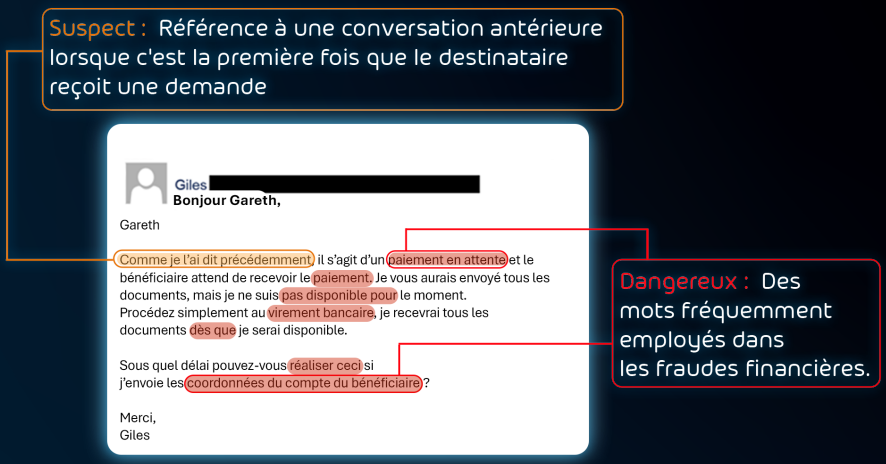
- Une augmentation soudaine du nombre d'e-mails similaires avec de légères variations dans les noms de domaine peut être le signe d'une nouvelle campagne d'hameçonnage.
- Une augmentation rapide du nombre d'e-mails au contenu totalement différent mais présentant quelques caractéristiques similaires, par exemple des noms de pièces jointes ou des liens similaires dans les premières lignes d'un e-mail, peut également indiquer une nouvelle campagne d'hameçonnage.

Cette méthode est également utilisée lorsque les auteurs d'attaques récoltent les messages des systèmes infectés et les réutilisent pour attaquer de nouvelles victimes en y apportant de légères modifications. Cette technique est toujours utilisée par divers botnets (par exemple QakBot en 2023).

- Extraire les caractéristiques clés d'un courriel, par exemple l'objet, l'expéditeur, la pièce jointe, les liens.
- Recherche de groupes d'e-mails présentant des caractéristiques similaires
- Classifier les courriels en fonction des groupes de courriels.
- Mise à jour du modèle avec les caractéristiques extraites.

Graphique montrant notre pipeline de ML pour l'identification du hameçonnage sur la base du clustering

Une autre technique d'IA utile est le traitement du langage naturel (Natural Language Processing) qui analyse le texte des courriels, en utilisant des approches telles que l'intégration des mots (word embedding) et la modélisation des sujets (topic modeling), en déduisant le contexte et la sémantique. Nous pouvons ensuite combiner cette analyse avec une analyse séquentielle / chronologique pour détecter des schémas de communication anormaux. Un exemple typique est celui d'un cadre demandant un transfert financier rapide, qui sera repéré s'il n'a jamais envoyé ce type d'e-mail par le passé.



Exemple de NLP analysant un texte et les signaux détectés

Comme indiqué précédemment, il s'agit d'un jeu continu d'adaptation par les criminels pour contourner nos défenses, et nous améliorons continuellement nos détections pour attraper de nouvelles variantes. L'une des forces des modèles de ML est qu'ils sont capables d'apprendre. Nous les maintenons donc à jour en utilisant diverses sources de signaux :

- Le retour d'information des utilisateurs est précieux, car il repose sur l'identification, par les utilisateurs finaux, des faux positifs (lorsque nous avons signalé un courriel comme étant malveillant alors qu'il ne l'était pas) et des faux négatifs (lorsque nous n'avons pas signalé un courriel, alors qu'il était en fait suspect).
- Les pots de miel sont des simulations de cibles ou de boîtes de réception d'e-mails qui attirent des attaques génériques et ciblées et que nous utilisons comme données d'entraînement.

Cela signifie que nos modèles s'adaptent en permanence à l'évolution rapide du paysage des menaces, ce qui est la meilleure façon pour une solution moderne d'hygiène du courrier électronique comme la nôtre de réussir.

Une autre approche adoptée par les criminels consiste à supprimer le contenu malveillant de l'e-mail en hébergeant la charge utile à l'extérieur sur un serveur web et en n'incluant qu'un lien vers celui-ci. La détection des liens malveillants fait partie intégrante de notre solution basée sur l'IA appelée Secure Links. Nous remplaçons chaque lien dans les courriels entrants par une version qui passe par notre Secure Web Gateway (SWG).

Ce système utilise le ML et le deep learning, ainsi que des modèles ML supervisés et non supervisés pour analyser 47+ caractéristiques des liens URL et des cibles des pages web, à la recherche de comportements malveillants, de redirections d'URL et d'obFuscation. Il utilise également des modèles de reconnaissance visuelle pour analyser les images, y compris les logos de marque et les codes QR, ainsi que le contenu textuel incorporé dans les images. Le résultat net est que nous capturons le contenu malveillant lié aux courriels, même lorsqu'il s'agit d'une attaque rapide et très ciblée. Une autre tactique très répandue consiste à compromettre un site web sans en modifier le contenu, puis à envoyer une campagne de courriels et, une fois les courriels envoyés, à déployer la charge utile malveillante. C'est pourquoi Secure Links analyse les liens cibles à la fois au moment de la distribution et au moment du clic.

Analyse des pièces jointes

Bien souvent, les auteurs d'attaques n'insèrent pas leurs charges utiles dans le texte même de l'e-mail, ni de liens susceptibles d'être analysés, mais le contenu malveillant est inclus dans une pièce jointe. Contrairement aux courriels textuels qui peuvent être analysés relativement facilement, les pièces jointes se présentent sous de nombreux formats de fichiers différents et peuvent être utilisées de différentes manières.

Ici, notre moteur Sandbox, toujours basé sur le ML, ouvrira les fichiers joints, identifiera s'ils sont malveillants et, s'ils le sont, mettra l'e-mail en quarantaine. Ce moteur examine le comportement du fichier, pour voir s'il essaie d'identifier s'il s'exécute dans un sandbox (ce qui est un signe avant-coureur). Il examine également le système de fichiers pour voir si la pièce jointe tente de créer de nouveaux fichiers ou de modifier des fichiers existants. Le moniteur de registre inspecte le registre pour voir si des valeurs inhabituelles sont créées, celles-ci étant souvent utilisées pour maintenir les logiciels malveillants après le redémarrage de l'ordinateur. Enfin, notre moniteur de processus détecte les tentatives des fichiers PDF et Office malveillants de lancer des processus enfants. Le trafic réseau de la pièce jointe est également inspecté, à la recherche de connexions à des serveurs sur Internet, ce qui constitue une tactique assez suspecte pour un document joint. Enfin, la mémoire dans le sandbox est inspectée après l'ouverture de la pièce jointe, les types inhabituels d'accès à la mémoire étant un autre signal fort de logiciel malveillant.

Au total, le moteur de ML s'appuie sur plus de 500 indicateurs dans les pièces jointes des fichiers du moteur Sandbox et les classe rapidement en fichiers bénins et malveillants.

Analyse des documents sortants

Une autre solution unique de Hornetsecurity qui est alimentée par l'IA est notre AI Recipient Validation (AIRV). Elle analyse les habitudes de communication par courriel de chaque utilisateur, apprend continuellement et détecte les destinataires involontaires, les courriels contenant des informations d'identification personnelle (PII) et les formulations inappropriées. Lorsque des problèmes sont détectés, ils sont signalés à l'utilisateur, et les réponses de l'utilisateur à ces avertissements sont ensuite incorporées dans les futurs courriels.

AIRV avertit les utilisateurs dans les scénarios suivants :

- Envoi d'un courrier électronique à un destinataire potentiellement involontaire. Les pots de miel sont des simulations de cibles ou de boîtes de réception de courrier électronique qui attirent des attaques génériques et ciblées et que nous utilisons comme données d'entraînement.
- Un destinataire par ailleurs courant d'une cohorte manque à l'appel..
- Un utilisateur est ajouté ou remplacé dans une cohorte existante.
- Envoi pour la première fois d'e-mails à des utilisateurs d'organisations ou d'adresses e-mail personnelles différentes.
- Réponse à une grande liste de distribution.
- Envoi d'un courrier électronique à un destinataire avec lequel l'utilisateur n'a jamais eu de relation.
- Envoyer des courriels contenant des informations sensibles, telles que des informations confidentielles ou des données de carte de crédit.
- L'envoi d'un courriel dont la formulation est inappropriée.

Former les utilisateurs à l'aide de l'IA

Étant donné qu'aucune solution d'hygiène des emails n'est efficace à 100% en permanence, il arrive que la dernière ligne de défense soit l'utilisateur final, qui doit être prudent lorsqu'il voit un email suspect. La solution de formation à la sensibilisation à la sécurité d'Hornetsecurity repose sur l'IA, qui fournit la bonne quantité de formation à chaque utilisateur. Des campagnes de hameçonnage simulées sont envoyées, et les utilisateurs qui cliquent sur des liens ou ouvrent des pièces jointes reçoivent une formation plus approfondie, tandis que ceux qui ne le font pas ne sont pas importunés par les demandes de formation. Le moteur AI Spear Phishing utilise également différents niveaux de sophistication dans les simulations (basés sur des attaques réelles que nous avons capturées), aidant les utilisateurs finaux à repérer même les attaques les plus avancées. À l'instar des attaques réelles, nos liens mènent à de fausses pages de connexion, les courriels font partie d'un fil de discussion et les pièces jointes sont accompagnées de macros « malveillantes ».
La véritable force du service de formation à la sensibilisation à la sécurité est que l'IA le gère automatiquement, libérant les administrateurs pour qu'ils se concentrent sur des tâches plus productives au lieu de micro-gérer les campagnes de simulation d'hameçonnage et les missions de formation.

Conclusion

Hornetsecurity suit le « train de l'IA » depuis de nombreuses années, en utilisant diverses saveurs d'outils pour différents défis, en affinant notre approche afin de fournir une protection efficace contre les menaces véhiculées par le courrier électronique et en formant les utilisateurs à repérer l'hameçonnage.