Las compañías de seguridad cibernética Sophos y ReversingLabs, lanzaron este lunes conjuntamente el primer conjunto de dato de investigación de malware a escala de producción que se pondrá a disposición del público en general y que tiene como objetivo construir defensas efectivas e impulsar mejoras en toda la industria en detección y respuesta de seguridad.

SoReL-20M (Sophos-ReversingLabs-20Million), como se llama, es un conjunto de datos que contiene metadatos, etiquetas y cuenta con 20 millones de archivos de Windows (portable ejecutable .PE), incluyendo 10 millones de muestras de malware desarmadas, con el objetivo de diseñar enfoques de aprendizaje automático para mejorar las capacidades de detección de malware.

> *«El conocimiento y la comprensión abiertos sobre las amenazas cibernéticas también conducen a una ciberseguridad más predictiva. Los defensores podrían anticipar lo que están haciendo los atacantes y estar mejor preparados para su próximo movimiento»*, dijo el grupo de IA de Sophos.

Junto con el lanzamiento de hoy, existe un conjunto de modelos de aprendizaje automático basados en PyTorch y LightGBM, previamente entrenados con estos datos como líneas de base.

A diferencia de otros campos como el lenguaje natural y el procesamiento de imágenes, que se han beneficiado de vastos conjuntos de datos disponibles públicamente como MNIST, ImageNet, CIFAR-10, IMDB Reviews, Sentiment140 y WordNet, conseguir conjuntos de datos etiquetados estandarizados dedicados a la ciberseguridad ha resultado un desafío debido a la presencia de información de identificación personal, datos confidenciales de infraestructura de red y propiedad intelectual privada, sin mencionar el riesgo de proporcionar software malicioso a terceros desconocidos.

Aunque EMBER (Endgame Malware BEnchmark for Research) se lanzó en 2018 como un clasificador de malware de código abierto, su tamaño de muestra más pequeño (1.1 millones

de muestras) y su función como conjunto de datos de etiqueta única (benigno/malware), significaba que era un *«límite del rango de experimentación que se puede realizar con él».*

SoReL-20M tiene como objetivo solucionar estos problemas con 20 millones de muestras de PE, que también incluyen 10 millones de muestras de malware desarmadas (que no se pueden ejecutar), así como características extraídas y metadatos para 10 millones de muestras benignas adicionales.

Además, el enfoque aprovecha un modelo de etiquetado basado en aprendizaje profundo entrenado para generar descripciones semánticas interpretables por humanos que especifican atributos importantes de las muestras involucradas.

El lanzamiento de SoReL-20M sigue las iniciativas similares de la industria en los últimos meses, incluyendo una coalición liderada por Microsoft, que lanzó Adversarial ML Threat MAtrix en octubre, para ayudar a los analistas de seguridad a redactar, responder y remediar ataques adversarios contra sistemas de aprendizaje automático.

> *«La idea de compartir inteligencia sobre amenazas en seguridad no es nueva, pero es más crítica que nunca dada la innovación que los actores de amenazas han mostrado en los últimos años. El aprendizaje automático y la inteligencia artificial se han vuelto fundamentales para estos esfuerzos, lo que permite a los cazadores de amenazas y los equipos de SOC ir más allá de las firmas y la heurística y ser más proactivos en la detección de malware nuevo o dirigido»*, dijeron los investigadores de ReversingLabs.