Security Data Science - Fase 1

Diana Zaray Corado #191025
Pablo Alejandro Méndez #19195
Orlando Osberto Mejía #19943
José Javier Hurtarte #19707

Security Data Science
Sección 10

Universidad del Valle de Guatemala
Facultad de Ingeniería
Guatemala, martes 21 de febrero del 2023

Índice

Introducción

Android es uno de los sistemas operativos móviles más populares. Según la página BusinessofApps, hay aproximadamente 2.5 mil millones de usuarios activos que utilizan dispositivos móviles con Android en el 2022.

Sin embargo, es importante reconocer que la popularidad de Android también la vuelve vulnerable a ataques maliciosos. De acuerdo con ProofPoint (2022), la cantidad de malware dirigido a Android incrementó en 2022.

Todo esto refleja la importancia de la detección de malware para dispositivos Android. De manera que se desarrollará una herramienta para detectar la presencia de Malware en dispositivos Android.

Motivación

El proyecto tiene como objetivo desarrollar una herramienta que emplee algoritmos de Machine Learning y diversas características relevantes de cada aplicación en dispositivos Android; como permisos, tráfico de red, código ofuscado, llamadas a sistema, llamadas de otras APIs relevantes, elementos visuales de la aplicación y metadata. Todo con el fin de detectar la presencia de Malware sin depender de métodos poco eficientes como firmas digitales. Mediante esto lo que se busca es proteger a los usuarios de Android ya que debido a su fácil accesibilidad permite la creación y publicación de aplicaciones fachada sin pasar por ningún control de seguridad o análisis profundo

Objetivos específicos:

- Identificar y seleccionar características relevantes de aplicaciones android que puedan ser útiles para entrenar un modelo de machine learning.
- Diseñar y desarrollar distintos modelos de machine learning para la detección de malware

Preguntas Clave

- ¿Cuáles son las propiedades o características de una aplicación que permiten identificarla como malware?
- ¿Cuáles son los permisos que una aplicación de malware necesita activar/solicitar para ser ejecutada en Android?

Revisión de la Literatura

En 2016 se realizó una investigación en la Universidad Estatal de San José en donde se examinó aplicaciones de Android utilizando análisis dinámico y estático para entrenar modelos de Machine Learning a reconocer si una aplicación es maligna o benigna. De todos los algoritmos usados, un modelo de Random Forest que utilizó como *features* los permisos utilizados por la aplicación y otros aspectos de análisis estático, obtuvo un valor AUC de 0.972. Al tener un AUC elevado, se puede decir que el modelo mencionado clasifica de una manera bastante precisa cuando un app es malware o no. (Kapratwar, A, 2016).

Similarmente, en 2018 la IEEE creó un modelo que utiliza aprendizaje por *Ensemble Learning*, que permite entrenar a distintos modelos de Machine Learning con los mismos datos y luego se conviene los modelos para proveer mejores resultados. Para *features,* utilizaron análisis estático de código para extraer permisos de apps, características del hardware, android *intents,* llamadas a API restringidas, permisos del app, entre otros. Este modelo unificó Support Vector Machines, k-nearest neighbor y random forest. Luego de varias optimizaciones alcanzó una precisión del 98.4%, una alta mejora con respecto a los modelos que existían previamente (Wang et al., 2018).

Almin y Chatterjee (2015) en *Novel Approach to detect Android Malware* presentan una sistema que permite el análisis de las aplicaciones instaladas en un dispositivo mediante la extracción de las características y los permisos solicitados por las aplicaciones. El sistema cuenta varias fases:

- Identificación de las aplicaciones instaladas
- Extracción de permisos: se recolectó todos los permisos que una aplicación solicitó al ser instalada.
- Clustering o Agrupación de Permisos

- Clasificación de permisos: mediante *Naive Bayes* se realizó un segundo proceso de clasificación para asegurar que aquellas aplicaciones clasificadas en el agrupamiento anterior correspondan al cluster correcto.
- Eliminación del Software si fuera necesario.

Ehsan et al. (2022) en *Detecting Malware by Analyzing App Permissions on Android Platform: A Systematic Literature Review* realiza una comparación de varios métodos utilizados para la detección de malware en Android basados en el análisis de permisos. Ehsan et al. (2022) menciona que en general los análisis estáticos basados en redes neuronales muestra mejores resultados que aquellos que no, y si bien, el análisis estático puede ser efectivo al detectar aplicaciones maliciosas, si se combina una diversa gama de categorías de análisis estático se logran mejores resultados. Uno de los principales retos que enfrenta el análisis estático es la creación de un conjunto óptimo de características para el entrenamiento eso se puede resolver mediante el uso de análisis dinámico. A su vez, comentan que se cuenta con una alta tasa de falsos positivos ya que una gran cantidad de aplicaciones seguras son clasificadas como *malware* debido a que tienen ciertas combinaciones de permisos considerados como peligrosos.

Recolección de datos

El conjunto de datos a utilizar para el análisis de malware de Android se encuentra en la plataforma de Kaggle. De hecho, dentro de está plataforma se pudieron encontrar varios conjuntos de datos que realizan la extracción de características de las aplicaciones y poseen la columna de clasificación dentro.

Un ejemplo de estos datos se puede encontrar en:

- https://www.kaggle.com/datasets/defensedroid/android-malware-detection

A su vez en AndroZoo se pudieron encontrar un conjunto de Android Application Package con los cuales se pretende realizar análisis estático para hacer una extracción de características manuales. Estos se pueden encontrar en:

- https://doi.org/10.1145/2901739.2903508

Bibliografía

Almin, S. B., & Chatterjee, M. (2015b). A Novel Approach to Detect Android Malware. *Procedia Computer Science, 45,* 407-417. https://doi.org/10.1016/j.procs.2015.03.170

Ehsan, A. (s. f.). *Detecting Malware by Analyzing App Permissions on Android Platform: A Systematic Literature Review*. MDPI. https://www.mdpi.com/1424-8220/22/20/7928

GeeksforGeeks. (2019, 6 noviembre). *Distance Vector Routing (DVR) Protocol*. Recuperado 6 de septiembre de 2022, de https://www.geeksforgeeks.org/distance-vector-routing-dvr-protocol/

Kapratwar, A. (2016). Static and dynamic analysis for Android Malware detection. https://doi.org/10.31979/etd.za5p-mqce

Curry.D (02/2023). Android Statistics (2023). https://www.businessofapps.com/data/android-statistics/

ProofPoint (2022). Mobile Malware is Surging in Europe: A Look at the Biggest Threats. https://www.proofpoint.com/us/blog/email-and-cloud-threats/mobile-malware-surging-europe-look-biggest-threats

Wang, W., Gao, Z., Zhao, M., Li, Y., Liu, J., & Zhang, X. (2018). Droidensemble: Detecting Android malicious applications with ensemble of string and structural static features. IEEE Access, 6, 31798–31807. https://doi.org/10.1109/access.2018.2835654