UNIVERSIDAD DE CASTILLA-LA MANCHA

ESCUELA SUPERIOR DE INFORMÁTICA
(departamento del director)[1]

ANTEPROYECTO DEL TRABAJO FIN DE GRADO
GRADO EN INGENIERÍA INFORMÁTICA
TECNOLOGÍA ESPECÍFICA DE COMPUTACIÓN

Detección y Clasificación de Malware usando técnicas inteligentes

Autor: Rubén Donate Serrano
Director: José Luis Martínez Martínez
Director: José Miguel Puerta Callejón

Abril, 2018

[1]DEPARTAMENTO DE SISTEMAS INFORMÁTICOS o DEPARTAMENTO DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA, AUTOMÁTICA Y COMUNICACIONES o DEPARTAMENTO DE MATEMÁTICAS o cualquier otro de la UCLM al que pertenezca el director.

Índice de contenido

1. INTRODUCCIÓN

Este Trabajo Fin de Grado (TFG, de ahora en adelante) se basa en un reto de Kaggle publicado por Microsoft en 2015 [?]. El cual, según se establece en la descripción del mismo, fue llevado a cabo debido a que en los últimos años, la industria del malware se ha convertido en un mercado bien organizado que involucra grandes cantidades de dinero. Los sindicatos multijugador, bien financiados, invierten mucho en tecnologías y desarrollar capacidades para evadir la protección tradicional, lo que exige que los proveedores de antimalware desarrollen contra-medidas para detectarlos y desactivarlos. Mientras tanto, infligen un daño financiero y emocional real a los usuarios de los sistemas informáticos.

Uno de los principales desafíos que enfrenta el antimalware hoy en día es la gran cantidad de datos y archivos que deben evaluarse para detectar posibles intenciones maliciosas. Por ejemplo, los productos antimalware de detección en tiempo real de Microsoft están presentes en más de 160 millones de computadoras en todo el mundo e inspeccionan más de 700 millones de computadoras mensualmente. Esto genera decenas de millones de muestras diarias para ser analizadas como posibles malware. Una de las razones principales de estos grandes volúmenes de archivos diferentes es el hecho de que, para evadir la detección, los autores de malware introducen polimorfismo en los componentes maliciosos. Esto significa que los archivos maliciosos que pertenecen a la misma "familia" de malware, con las mismas formas de comportamiento malicioso, se modifican y/u ofuscan constantemente mediante diversas tácticas, de modo que se parecen a muchos archivos diferentes.

Para ser efectivos al analizar y clasificar cantidades tan grandes de archivos, debemos ser capaces de agruparlos e identificar a la familia que le corresponde. Además, dichos criterios de agrupamiento se pueden aplicar a los archivos nuevos que se encuentran en las computadoras para detectarlos como maliciosos y determinar a que familia corresponde.

Por todo esto, Microsoft decidio proporcionar a la comunidad de cientificos de datos un conjunto de datos de malware sin precedentes [?] y fomento la creación de técnicas efectivas y de código abierto para agrupar variantes de archivos de malware en sus respectivas familias.

En cambio, desde mi punto de vista, el motivo por el cual se eligió este reto como propuesta para este TFG fue aunar en este los dos conjuntos de capacidades como son la Minería de Datos, Big Data, etc. por las cuales escogí la Intensificación de Computación, y la Ciberseguridad la cual me ha hecho escoger optativas orientadas a este ámbito, como puede ser Auditoría de Sistemas de Información, Criptografía y Dispositivos y Redes Inalámbricos, y ademas de otras cuestiones. Por todo esto, a la hora de elegir el tema de mi TFG busque opciones que englobaran estos dos conjuntos de capacidades. Por eso cuando, se me ofreció la posibilidad de realizar mi TFG sobre este reto no lo dude.

2. TECNOLOGÍA ESPECÍFICA / INTENSIFICACIÓN / ITINERARIO CURSADO POR EL ALUMNO

Este TFG se contextualiza dentro de la intensificación de Computación, conforme se puede observar en la tabla 1.

Marcar la Tecnología Cursada
Tecnologías de la Información
Computación
Ingeniería del Software
Ingeniería de Computadores

Tabla 1: Tecnología Específica cursada por el alumno

A continuación, en la tabla 2 se muestran la competencias abordadas junto con su correspondiente justificación.

Competencias	**Justificación**
Competencia 1	[Exponer y argumentar cómo y en qué parte se va a abordar esta competencia en el TFG]
Competencia 2	[Exponer y argumentar cómo y en qué parte se va a abordar esta competencia en el TFG]

Tabla 2: Justificación de las competencias específicas abordadas en el TFG

El Trabajo Fin de Grado (TFG, de ahora en adelante) siempre deberá demostrar la aplicación de las competencias generales de la titulación. Además, el TFG deberá aplicar **algunas** de las competencias específicas asociadas a la **Tecnología Específica o Intensificación** que el alumno ha cursado. Por lo tanto, el alumno incluirá en el anteproyecto **dos tablas**. Una tabla para seleccionar la tecnología cursada y en la que se contextualiza el TFG:

En la segunda tabla, el alumno deberá justificar cómo **algunas** de las competencias específicas de la intensificación se aplicarán o tomarán forma en el TFG, **La relación de competencias por intensificación se encuentran en el Anexo I al final de este documento.**

Competencias	**Justificación**
Competencia 1	[Exponer y argumentar cómo y en qué parte se va a abordar esta competencia en el TFG]
Competencia 2	[Exponer y argumentar cómo y en qué parte se va a abordar esta competencia en el TFG]

Tabla 3: Justificación de las competencias específicas abordadas en el TFG

3. OBJETIVOS

Según el reto que se va a llevar a cabo en este TFG, el objetivo principal es al creación de un sistema experto para la agrupación de las distintas muestras proporcionadas en la base de datos de malware en sus correspondientes familias.

Para realizar mi propuesta de solución, el lenguaje en el que va a ser implementada es Python 3. Los recursos que se van a utilizar son mi portátil personal (Toshiba L150-850) y mi disco duro externo, con las siguientes características:

- Procesador Intel Core i7-3610QM a 2,30 GHz.
- 8 Gb RAM DDR3.
- Tarjeta Gráfica Radeon 7670M.
- Disco Duro SSD Samsung EVO 850 de 250 Gb.
- Disco Duro externo Seagate Expansion de 4 Tb.

También se van a utilizar Docker y MongoDB, las cuales son tecnologías correctamente consolidadas al igual que Python 3.

Debido a que este objetivo es muy genérico, se puede subdividir en sub-objetivos, siendo estos:

- Familiarizarse con las seguridad informática.
- Estudio de las distintas familias de malware utilizadas en el reto. Esto es necesario para su conocer su comportamiento y ser capaz de identificar las posibles características y de ese modo seleccionar las mas apropiadas para estas familias de malware.
- Buscar y analizar las soluciones propuestas por otros participantes en el reto. Esto es posible ya que una de la condiciones del reto es que las soluciones sigan la filosofía de código abierto. Esto me permite obtener ideas de los que van realizando otros equipos y la posibilidad de comparar los resultados con los obtenidos en mi solución.
- Estudiar y analizar distintos métodos para el procesado de fichero de texto. Esto es necesario porque la bases de datos proporcionadas para el reto se encuentran almacenadas dos ficheros comprimidos formados por ficheros de texto donde se almacenan las información de las muestras proporcionadas, por lo cual es necesario manejar estos ficheros para la extracción de características.
- Diseño e implementación de un algoritmo de extracción de las características que van a ser utilizadas para la identificación de las muestras proporcionadas para el reto.
- Estudio y análisis de distintas técnicas para el procesamiento de lenguaje natural. Es necesario ya que uno de los tipos de ficheros proporcionados contiene las instrucciones del malware desensambladas con IDA, lo que supone el manejo de las ocurrencias de palabras utilizadas en ensamblador junto otras como pueden ser las llamadas al sistema. Esto hace que sea necesario la utilización de técnicas que permita la probabilidad de que estas palabras en todos los documentos proporcionados.

- Estudio de distintos sistemas para el almacenaje de grandes cantidades de datos. El motivo por el que es necesario este objetivo es porque tenemos que utilizar una gran cantidad de información en todo momento por lo cual no es posible mantenerlo en todo momento en memoria, por lo cual es obligatoria la utilización de algún sistema de almacenaje adecuado para grandes cantidades de datos.
- Estudio e implementación de técnicas para la reducción de la dimensionalidad. Esto, al igual que el punto anterior es en parte debido a la gran cantidad de información disponible para su análisis, por lo cual es necesario para conseguir hacer la implementación abordable sobre todo con los recursos que van a ser utilizados.
- Implementación de un sistema experto utilizando técnicas de Minería de Datos y Aprendizaje Automático.
- Prueba del sistema experto construido con la bases de datos proporcionadas.
- Comparación de los resultados obtenidos con otras soluciones propuestas por otros equipos.
- Analizar la posibilidad de extender el algoritmo implementado a un escenario mas general de detección de malware. Esto sub-objetivo solo se analizara de manera teórica y se realizar ya que viene establecido en el reto su realización.
- Propuesta de posibles mejoras.

4. MÉTODO Y FASES DE TRABAJO

Para el desarrollo del TFG y la obtención de los objetivos expuestos en el punto anterior, se seguirán las siguientes fases:

1. Lectura y aprendizaje sobre diferentes tipos de Malware y sus características.
2. Estudio y aprendizaje de técnicas de Minería de Datos para el procesado de grandes bases de datos.
3. Realización de una propuesta de clasificación para el reto propuesto.
4. Evaluación del algoritmo propuesto y su comparación con otras propuestas.
5. Posibilidad de extender el algoritmo propuesto a escenarios para la detección de malware.
6. Escritura de la Memoria.

5. MEDIOS QUE SE PRETENDEN UTILIZAR

5.1. Medios Hardware

Para este TFG, es necesario disponer de al menos los siguientes recursos:

- 8 Gb de Ram.
- Disco duro de 2 Tb.
- Acceso a internet, en caso de no disponer sería necesario proporcionar a docker los contenedores utilizados como base para este proyecto.

5.2. Medios Software

El TFG se ha implementado haciendo uso de contenedores docker, por lo cual es necesario disponer tanto de docker como de docker-compose para su correcta ejecución.