UNIVERSIDAD DE CASTILLA-LA MANCHA

ESCUELA SUPERIOR DE INFORMÁTICA
DEPARTAMENTO DE SISTEMAS INFORMÁTICOS

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

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 [1]. 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 [2] 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 tan distintas como son por un lado la Minería de Datos, Big Data, etc., capacidades que me motivaron a escoger 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.

El motivo de esta elección desde el punto de vista de la ciberseguridad, es que se trata un tema muy habitual y muy interesante como es el tema de los malware, tanto que una tan importante a nivel Internancional como es Microsoft dedidio proporcionar este reto como bien indica en su descripción. Esto hace que en este TFG sea necesario trabajar con la información obtenida mediante técnicas de reversing de las muestras de malware. Estas técnicas de reversing nos proporcionan información en lenguaje ensamblador y en lenguaje binario, siendo necesario manejar y familiarizarse con esta información, estos lenguajes y las posibles características que se puedan extraer de ellos y nos sean de utilizad.

En cambio, en lo que respecta a las capacidades relacionadas con la Intensificación de Computación y conforme se puede deducir de la cifras que se indican en la descripción por parte de Microsoft, nos encontramos ante

un problema de Big Data. También, es necesario la utilización de técnicas de Minería de Datos para conseguir la clasificación de las distintas muestras en su correspondiente familia. Pero no solo eso, de las cifras que se observan también se puede llegar a concluir que debido a la gran cantidad de variantes dentro de una misma familia de malware, sea necesario la utilización de una gran cantidad de características para la correcta identificación de la familia a la que pertenece, haciendo así que nos encontremos ante uno de lo mayor problemas a la hora de realizar procesos de Minerías de Datos como es la alta dimensionalidad. Este problema nos hace que sea necesario utilizar técnicas para reducir el número de características a un número que sea abordable y permita la clasificación en la familia de manera correcta. Además, debido a los recursos que voy a utilizar para este TFG, vease el apartado 5, hace que se agrave este problema, haciendo necesario la utilización de otras tecnologías, como pueden ser MongoDB, para conseguir que sea posible abordarlo.

Debido a todas estas consideraciones, cuando se me presento la oportunidad de realizar este TFG me pareció tremendamente interesante, ya que me permitiría aprender muchas capacidades interesantes de estos dos grupo. Por ese motivo, fue que no dude en seleccionar este tema para mi TFG una vez se me propuso.

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.

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.

Competencias	Justificación
[CM1] Capacidad para tener un conocimiento profundo de los principios fundamentales y modelos de la computación y saberlos aplicar para interpretar, seleccionar, valorar, modelar, y crear nuevos conceptos, teorías, usos y desarrollos tecnológicos relacionados con la informática.	En este TFG consiste en el diseño de un modelo para la clasificación de las muestra de malware que se nos proporcionan, esto hace que haya que seleccionar, construir y valorar modelo, e implementar algoritmos proprios, además de estudiar todos estos y llevarlos a la práctica para concluir obteniendo una predicción para el reto.
[CM3] Capacidad para evaluar la complejidad computacional de un problema, conocer estrategias algorítmicas que puedan conducir a su resolución y recomendar, desarrollar e implementar aquella que garantice el mejor rendimiento de acuerdo con los requisitos establecidos.	Este punto es muy importante durante el desarrollo del TFG, ya que conforme se indica en el introducción nos encontramos ante un problema de alta dimensionalidad, el cual implica tener una gran relevancia la eficiencia de los recursos de memoria que se utilizaran, haciendo que sea necesario priorizar este frente a otros aspectos de eficiencia.
[CM4] Capacidad para conocer los fundamentos, paradigmas y técnicas propias de los sistemas inteligentes y analizar, diseñar y construir sistemas, servicios y aplicaciones informáticas que utilicen dichas técnicas en cualquier ámbito de aplicación.	Con respecto a esta competencia va a ser cubierta debido a que el objetivo principal del reto consiste en la construcción de un sistema inteligente mediante el cual se clasifiquen las muestras de malware proporcionado en la familia de malware que le correspondo.
[CM5] Capacidad para adquirir, obtener, formalizar y representar el conocimiento humano en una forma computable para la resolución de problemas mediante un sistema informático en cualquier ámbito de aplicación, particularmente los relacionados con aspectos de computación, percepción y actuación en ambientes o entornos inteligentes.	Otro de los puntos importes en el reto que aborda este TFG es la extracción de las características que se utilizaran posteriormente en el construcción del sistema inteligente y posteriormente realizar la predicción, a partir de la base de datos de muestras, lo cual implica la extracción del conocimiento de las bases de datos que se nos proporciona.
[CM7] Capacidad para conocer y desarrollar técnicas de aprendizaje computacional y diseñar e implementar aplicaciones y sistemas que las utilicen, incluyendo las dedicadas a extracción automática de información y conocimiento a partir de grandes volúmenes de datos.	En cuanto a esta competencia, esta justificada debido a que en todo momento este TFG consiste en extracción de las características de las bases de datos proporcionadas, para posteriormente construir un sistema inteligente para la clasificación de las muestras suministradas. Además, a pesar de que el número de muestras de cada base de datos es de aproximadamente 11.000 unidades la gran cantidad de caracteríastica que se extraen de cada muestra implica el manejo de una gran volumen de datos llegando al punto de encontrarnos ante un problema de alta dimensionalidad, el cual ha de ser analizado, abordado y resulto para por concluir el reto con su correspondiente predicción.

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

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

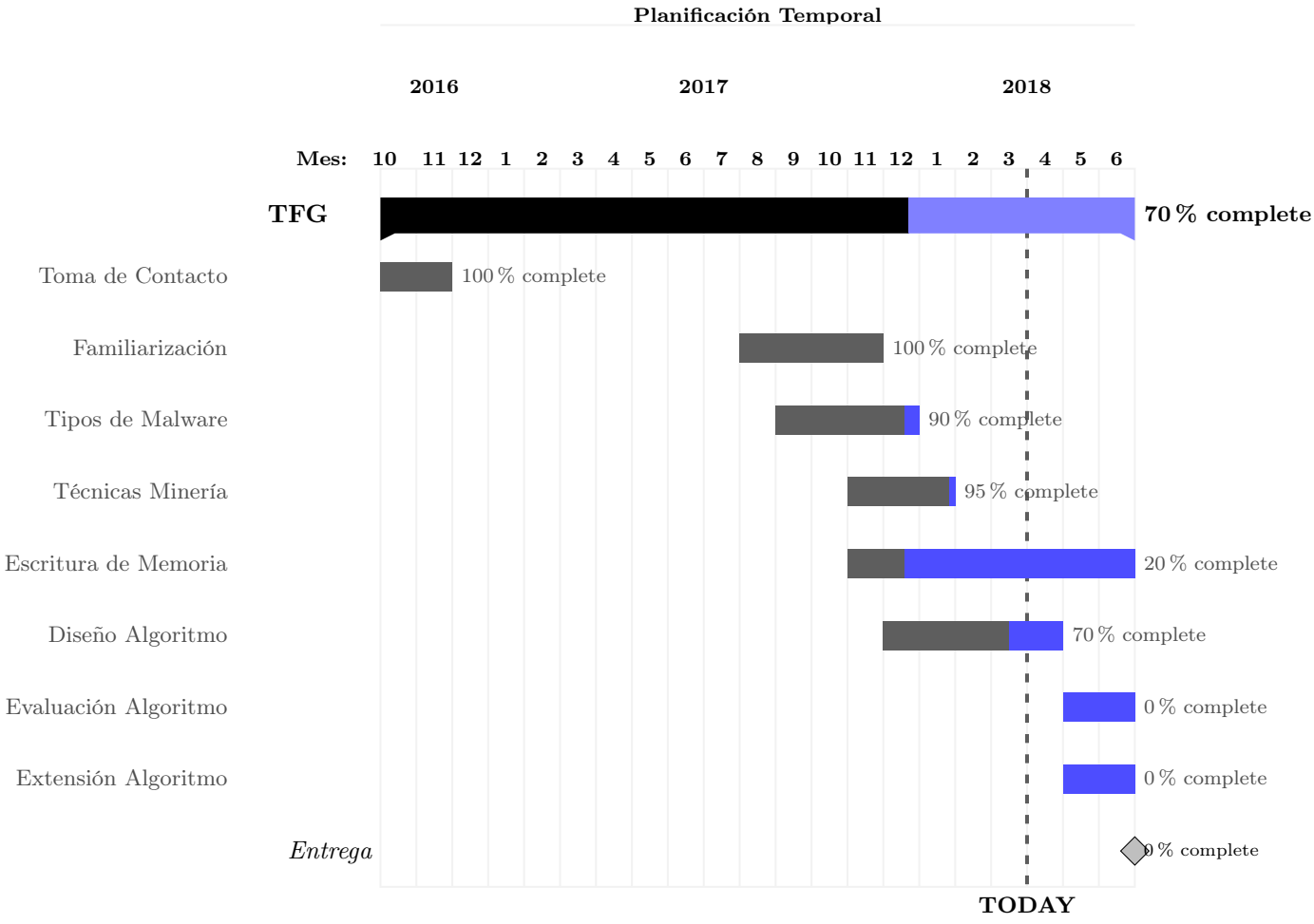
- Familiarizarse con las seguridad informática.
- Estudio de las distintas familias de malware utilizadas en el reto. Esto es necesario para su conocer su comportamiento y ser capaz de identificar las posibles características y de ese modo seleccionar las mas apropiadas para estas familias de malware.
- 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. Familiarizarse con las seguridad informática.
2. Lectura y aprendizaje sobre diferentes tipos de Malware y sus características.
3. Estudio y aprendizaje de técnicas de Minería de Datos para el procesado de grandes bases de datos.
4. Realización de una propuesta de clasificación para el reto propuesto.
5. Evaluación del algoritmo propuesto y su comparación con otras propuestas.
6. Posibilidad de extender el algoritmo propuesto a escenarios para la detección de malware.
7. Escritura de la Memoria.

A continuación, se mostrará una diagrama de gantt en el cual se observa la planificación temporal de las tareas de este TFG.



5. MEDIOS QUE SE PRETENDEN UTILIZAR

5.1. Medios Hardware

Para este TFG, es necesario disponer de al menos los siguientes recursos:
Los recursos que se van a utilizar para llevar a cabo el reto en lque se va a centrar el TFG, 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.
- Acceso a internet, en caso de no disponer sería necesario proporcionar a docker los contenedores utilizados como base para este proyecto.

De los cuales, son obligatorios para la ejecución de la solución propuesta para este TFG los siguientes:

- 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. Además, se implementado la solución propuesto en Python 3 y se ha utilizado MongoDB para almacenar la información generada, por lo cual se han utilizado los contenedores de la última versión oficial disponible para estas tecnologías.

6. REFERENCIAS

[1] (2015) Microsoft Malware Classification Challenge (BIG 2015). Microsoft Malware Protection Center and Microsoft Azure Machine Learning and Microsoft Talent Management. [Online]. Available: https://www.kaggle.com/c/malware-classification

[2] R. Ronen, M. Radu, C. Feuerstein, E. Yom-Tov, and M. Ahmadi, "Microsoft malware classification challenge," *arXiv preprint arXiv:1802.10135*, 2018.