Detección de Spear Phishing basada en métodos de decisión multicriterio

José Mariano Velo, Ángel Jesús Varela-Vaca,y Rafael M. Gasca
IDEA Research Group, Universidad de Sevilla
josvelmor2@alum.us.es, {ajvarela, gasca}@us.es

***Resumen*—Una de las principales amenazas que existen hoy en día a nivel de ciberseguridad, es el phishing y especialmente el denominado como spear phishing, también conocido como phishing elaborado o phishing dirigido. El impacto económico en términos globales de éstos sofisticados ataques es muy grande y crece a diario. Para tratar de mejorar el análisis que se realiza en un sistema de filtrado de correo y tras una primera aproximación en la que hemos usado un sistema para la toma de decisiones basado en modelos de decisión para la fase final, en este artículo abordamos la aplicación de Métodos de Decisión Multicriterio como TOPSIS y VIKOR para determinar con mayor precisión si un correo recibido es spear phishing o no. Para ello vamos a definir un modelo decisión multicriterio usando TOPSIS y VIKOR sobre más veinte características y criterios cognitivos y otros relacionados con elementos intrínsecos de los correos electrónicos. Este modelo será evaluado sobre un conjunto de datos con correos legítimos, phising y spear phising que nos proporcionarán rankings sobre la posibilidad de spear phising o no.**

***Index Terms*—Spear phishing, Métodos de Decisión Multicriterio, TOPSIS, VIKOR**

Tipo de contribución: *Investigación original*

I. Introducción

La principal característica del spear phishing frente al phishing estándar o clone phishing, es la minuciosa elaboración y preparación del mismo por parte de los atacantes. Dicha preparación suele conllevar un largo y detallado estudio de la víctima potencial y le confiere un grado de peligrosidad muy alto a éste tipo de ataques. No hay que decir que éste tipo de phishing tiene un enorme impacto económico no solo a nivel particular sino a nivel global, como lo demuestran los miles de millones de dólares que cuesta reparar los daños producidos por los mismos [1]. La detección de spear phishing lleva asociada una serie de problemas complejos, que en la mayoría de los casos impide que se pueda realizar por parte de los sistemas de seguridad de correo, un análisis certero sobre la posibilidad de que un determinado correo sea o no una amenaza. El principal problema es inherente a los propios correos de spear phishing ya que su elaboración implica, como se ha dicho, un estudio detallado de la víctima y en muchos casos conlleva como parte de ese estudio, los propios sistemas de seguridad que protegen a la víctima y al sistema de correo, además de sus puntos débiles. Llegando en casos extremos a realizarse por parte de los atacantes pequeñas pruebas de penetración con correos muy aislados que suelen pasar desapercibidos, para ir determinando las capacidades de análisis del sistema de seguridad.

Otro problema es la similitud en varios aspectos de los correos de spear phishing con los correos legítimos. A veces son difíciles de distinguir y solo se consigue atendiendo a pequeños detalles. Cuando hablamos de las peculiaridades de éste tipo de correos maliciosos nos referimos, por ejemplo, a características como los dominios de origen y los dominios de los enlaces, que suelen ser dominios “limpios”no catalogados en listas negras o de reputación, o bien dominios legítimos previamente comprometidos para el fin perseguido. O incluso la coincidencia o no entre dichos dominios. Esas características y otras, al ser distintas, suelen provocar que los filtros de correo cataloguen como legítimos dichos correos cuando en realidad no lo son. A priori, se podrían realizar ajustes más agresivos en los filtros de seguridad de modo que se llegaran a rechazar parte de los correos con esta tipología, pero con el riesgo de que se bloqueen correos legítimos, cosa que no suelen hacer los administradores de sistemas, quienes suelen centrarse más en la operativa que en la seguridad.

Respecto a la analítica de correos, probablemente sucederá como ha sucedido con las páginas http y https, ya que se empezará a obligar (como ejemplo, Google [2]) a usar mecanismos como Sender Policy Framework (SPF), Domainkeys Identified Mail (DKIM) y especialmente Domain-based Message Authentication, Reporting, and Conformance (DMARC) para asegurar el envío de correo y evitar todo tipo de correos basura, spam, phishing y otros. Desde ese momento la seguridad se incrementará, pero aún así no podemos hablar de estar libres de riesgos para determinados correos como los de spear phishing, dada la naturaleza de los mismos. Finalmente, podemos citar otro problema, que es la inexistencia de datasets adecuados para entrenar los modelos de ML/DL, problema que ha sido advertido y comentado por otros autores [3] [4], llegando a proponer incluso la creación de datasets sintéticos basados en las características comunes de dichos correos. Los datasets sintéticos pueden ayudar a entrenar modelos de mejor manera pero en nuestra opinión volveríamos a tener un problema con los correos que no se ajusten a las características estudiadas. Añadimos que en un porcentaje muy amplio de correos de spear phishing, por no decir todos, un analista de seguridad entrenado sería capaz de detectarlos.

Por ello y tras la primera aproximación basada en computación cognitiva para abordar el problema [5], proponemos mejorar la toma de decisión sobre si un determinado correo es spear phishing o no, usando Métodos de Decisión Multicriterio como, por ejemplo, TOPSIS [6] y VIKOR [7]. En la Figura

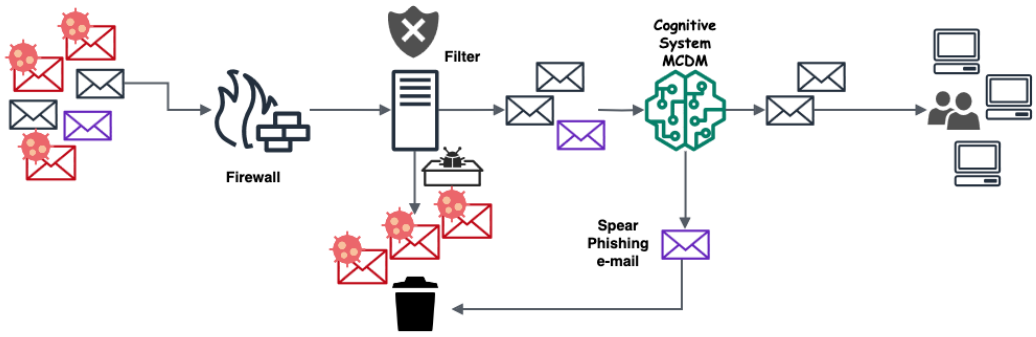

Figura 1. Flujo general en el análisis de correos entrantes.

1 siguiente se puede observar el flujo de correo entrante en un sistema corporativo, donde los correos maliciosos son descartados y nuestro sistema detecta los correos con spear phishing que han conseguido superar el filtro. Aportaremos entonces a la toma de decisiones un modelo de decisión multicriterio, y además una serie de criterios o características de tipo psicológico, de manera adicional a las que se habían propuesto inicialmente.

Como resumen, las contribuciones principales de éste trabajo son:

- Establecer las características o criterios más relevantes o con una correlación superior para ser procesadas por el método de decisión.
- Atribuir los pesos correspondientes a dichas características/criterios, necesarios para obtener los resultados más precisos posibles.
- Establecer los métodos de evaluación de dichas características.
- Probar el método para un dataset real, aunque limitado, de correos que mezcla correos legítimso, phishing y spear phishing.

II. Background: Métodos de Decisión Multicriterio: TOPSIS y VIKOR

El problema de clasificación de un determinado correo electrónico como spear phishing o no, se puede realizar de acuerdo con criterios determinados que capturan las características (factores, propiedades o atributos) de dicho problema de toma de decisiones [8]. Los Métodos de Decisión Multicriterio (MCDM) [9] utilizan diferentes técnicas para elaborar una selección óptima de las posibles alternativas de decisión de acuerdo con dichos criterios. En nuestro caso solamente tendremos tres alternativas posibles: (1) si es spear phishing; (2) si es phishing; o (3) si es un correo legítimo. Y tendremos un problema de selección entre dichas alternativas. Además de la selección de la mejor alternativa posible, debemos tener en cuenta también la eficiencia en la la toma de decisiones agilizando este proceso de selección, dada la urgencia de la misma en la detección. Además, dicha detección temprana permite la alineación estratégica con los objetivos de la organización, relacionados con la mejora continua mediante la gestión de riesgos de forma proactiva de los correos empresariales personalizados con riesgos de seguridad significativos.

Los MCDM se pueden dividir:

1. De acuerdo con la información de partida, si se basa en un conjunto de atributos, toma de decisiones multi-atributo (MADM) o ampliando a múltiples objetivos a satisfacer, que a menudo están en conflicto, la toma de decisiones objetivas (MODM);
2. Según el tipo de información, determinista, estocástica o incierta;
3. De acuerdo con el grupos de responsables de la toma de decisiones, uno o varios grupos.

Dependiendo de la estrategia utilizada para clasificar las alternativas, existen métodos basados en la distancia (TOPSIS, VIKOR, etc.), comparación por pares de alternativas (AHP, ANP, etc.), métodos de puntuación (SAW, Modelo de Suma Ponderada (WSM), Modelo de Producto Ponderado (WPM), etc.) y métodos de clasificación superior; por ejemplo, PROMETHEE o ELECTRE.

En el contexto de decisiones complejas, como en el caso de este trabajo hemos decidido que las mejores opciones podrían ser TOPSIS y VIKOR. Estas metodologías se han destacado por su amplia adopción y uso en la comunidad académica y empresarial, debido a su efectividad en la gestión de problemas discretos que involucran múltiples criterios y alternativas. Su popularidad se debe a su capacidad para abordar problemas de decisión de compleja, su versatilidad y su sólida base teórica resultado de décadas de investigación y aplicaciones exitosas. En las siguientes subsecciones se introducen estos MCDM.

II-A. TOPSIS

TOPSIS (acrónimo de Technique for Order of Preference by Similariry) as un método de análisis de decisiones multicriterio originalmente propuesto por Ching-Lai Hwang y Yoon en 1981 [10] y más tarde desarrollado por Yoon en 1987 [11] y Hwang, Lai y Liu en 1993 [6].

Se basa en que dado un conjunto de alternativas, se generan dos nuevas alternativas una solución ideal positiva (PIS) y otra una solución ideal negativa (NIS), de tal forma que la mejor

alternativa es la que tiene la distancia euclídea más corta con respecto al PIS y la distancia más larga al NIS.

La principal ventaja del método es que puede además generar un ranking de las alternativas en función de su proximidad a la solución ideal y su lejanía a la solución ideal negativa, considerando múltiples criterios y sus correspondientes ponderaciones.

Dada una matriz de decisión de n criterios y m alternativas, los pasos de este método son:

- Paso 1: Calcular la matriz de decisión normalizada teniendo en cuenta criterios positivos y criterios negativos.
- Paso 2: Calcular la matriz de decisión normalizada ponderada de acuerdo con a las ponderaciones de los diferentes criterios.
- Paso 3: Calcular la solución ideal positiva (PIS) y la solución ideal negativa (NIS). El PIS se define como los valores máximos para cada criterio y NIS como los valores mínimos.
- Paso 4: Calcular la distancia desde el PIS y el NIS a las distintas alternativas.
- Paso 5: Calcular la puntuación de las diferentes alternativas teniendo en cuenta las distancias anteriores.

En la última década, TOPSIS se ha utilizado para la toma de decisiones en áreas tales como la gestión de la cadena de suministro, el medio ambiente, la energía, la salud o los negocios, para clasificar o seleccionar diferentes alternativas u optimizar procesos [12][13][14].

II-B. VIKOR

Al igual que el método TOPSIS, se basa en métricas de distancia. VIKOR (acrónimo de VIseKriterijumska Optimizacija I Kompromisno Resenje) es una metodología multicriterio para la toma de decisiones. Fue desarrollado originalmente por Serafim Opricovic [7] a finales de la década de 1970 para resolver problemas de decisión con criterios en conflicto y no conmensurables.

VIKOR ofrece la ventaja de proporcionar una clasificación basada en la solución para el mejor y en el peor de los casos, así como a la media geométrica, permitiendo considerar diferentes escenarios para la clasificación de las alternativas. Su ventaja radica en que se aplica a situaciones en las que se requiere un equilibrio entre la optimización y la robustez, ofreciendo un compromiso que minimice la pérdida potencial en caso de variaciones adversas en la criterios.

Por tanto, dado un conjunto de alternativas, criterios y un valor $v \in [0, 1]$ que es utilizado en el cálculo del índice Q, los pasos que propone este método son:

- Paso 1: Calcular las soluciones en el mejor y en el peor de los casos con los mejores y peores valores
- Paso 2: Calcular los índices S y R utilizando el mejor caso anterior y las soluciones en el peor de los casos. El índice S representa qué tan bien una alternativa satisface los criterios, mientras que R representa la medida en que una alternativa no alcanza las soluciones anteriores para cada criterio de acuerdo con la Lp-métrico.
- Paso 3: Calcular el índice Q, que se basa en los índices anteriores y se determina mediante una función que los combina y el valor v . Esta función produce un valor del índice Q para cada alternativa.
- Paso 4: Las alternativas se clasifican de acuerdo con el índice Q. Aquellos con valores más bajos se consideran preferibles y se clasifican en la parte superior de la lista, indicando su prioridad. La clasificación obtenida determina una solución de compromiso, proporcionando una utilidad máxima de grupo para "la mayoría" y un rechazo mínimo individual para el "oponente".

Desde mediados de la década de 2000, VIKOR se ha convertido en una empresa de toma de decisiones multicriterio, metodología que ha atraído el interés de muchos investigadores. Según un estudio [15] publicado en 2016, en el que se revisaron un total de 176 artículos publicados entre 2004 y 2015 en revistas científicas, VIKOR se ha utilizado en áreas como la investigación operativa, la ciencias de la gestión (management), tomas de decisiones, sostenibilidad o energías renovables. En [16], se realizó un análisis comparativo de VIKOR con otros métodos (TOPSIS y métodos de clasificación), a través de la discusión de sus características distintivas y sus resultados de aplicación.

III. Propuesta de sistema de detección de Spear Phishing basada en MCDM

En esta propuesta que hacemos para el diseño de un modelo de decisión o sistema multicriterio de detección para Spear Phishing, necesitamos como hemos comentado en el apartado anterior, una serie de criterios que integrarán la matriz de decisión correspondiente, puesto que las alternativas serán los correos a determinar si son spear phishing o no.

Supondremos que los criterios que vamos a considerar son independientes entre ellos y los valores que se dan a cada criterio con respecto a cada alternativa se basan en la evaluación que se hace del criterio con respecto a ser un spear phishing. Tras un exhaustivo análisis de los posibles criterios que podrían determinar si un correo es spear phishing o no y acudiendo a una perspectiva holística del problema, proponemos cinco bloques de criterios de decisión, de los cuales los tres primeros son criterios ya expuestos en [5] y adicionalmente hemos añadido a nuestra propuesta dos bloques más. El bloque cuarto de criterios basados en factores psicológicos [17] [18] [19] y el bloque quinto de criterios basados en la persuasión personalizada [20] [21] [3]. En ambos casos se trata de criterios que a nuestro juicio enriquecen el análisis, ya que los correos de spear phishing cuentan con características que inicialmente la máquina no es capaz de reconocer. Suelen ser, además, características muy similares a las de los correos legítimos que impiden clasificarlos como maliciosos, lo cual hace que dichos correos pasen los filtros de seguridad con relativa facilidad. Por ello, analizar cuestiones psicológicas y persuasivas personalizadas tal y como lo haría un analista de seguridad humano, más allá de los criterios puramente técnicos asociados a cada correo que se reflejan en los tres primeros bloques, proporciona a nuestro juicio un

avance importante en la detección de las amenazas basadas en spear phishing.

1. **Criterios basados en la cabecera del correo** [Peso asignado: 10 %]
   a) Subcriterio **SPF: Sender Policy Framework** o Marco de Políticas del Remitente en lo que podría ser una traducción a nuestro idioma. Consultando los registros DNS del dominio origen del correo, los servidores de correo actuales suelen marcar los correos entrantes con la etiqueta "Received-SPF" y a continuación el valor del resultado de la consulta, que suele ser "Pass" si el remitente es permitido y la IP desde la que llega es uno de los registros MX legítimos del dominio o "None" en caso contrario. Generalmente además los correos suelen pasar por distintos servidores y por ello suele encontrarse más de una etiqueta de "Received-SPF" generada por parte del filtro de seguridad en cada correo. Asignaremos un valor entre cero y uno a este campo en función del valor encontrado, siendo cero cuando todos las etiquetas contienen "Pass" y uno cuando todas las etiquetas contienen "None", con valores intermedios si hay etiquetas con distintos valores y proporcionalmente al número de "Pass" encontrados.
   b) Subcriterio **DKIM: Domainkeys Identified Mail** o Correo Identificado por Claves de Dominio si lo traducimos al castellano. Mediante técnicas de criptografía asimétrica, el remitente firma en la cabecera de los correos con una firma digital única, de modo que el receptor puede verificar, usando la clave pública existente en el registro TXT del dominio origen, la autenticidad e integridad de los mismos. En nuestro caso asignaremos un cero en caso de que se verifique correctamente el correo con la citada clave pública y un uno en el caso de que no sea así.
   c) Subcriterio **DMARC:** Traducido sería autenticación, informes y conformidad de mensajes basados en dominio. Previene del uso fraudulento de dominios, evitando el phishing y el spoofing. Y para ello se basa en SPF y DKIM inicialmente, pero lleva a cabo más comprobaciones de seguridad y genera informes de malas configuraciones de los dos anteriores. También en éste caso debe existir un registro TXT en el dominio con la política de autenticación aceptada para los correos entrantes. Asignaremos un cero en el caso de que se haya realizado una comprobación válida y un uno en el caso contrario.
   d) Subcriterio **Reputación de la IP del remitente.** Una comprobación rápida y automática en un portal de inteligencia [22] permite catalogar la reputación de una determinada IP en función del historial de la misma a lo largo del tiempo. Asignaremos un valor entre cero y uno dependiendo de si la reputación es mejor o peor según el portal citado, siendo uno el valor asignado a la peor reputación posible.
   e) Subcriterio **Reputación del dominio del remitente.** Al igual que el criterio anterior, una comprobación rápida y automática en un portal de inteligencia como el comentado, permite catalogar la reputación de una determinado dominio en función del historial. Asignaremos un valor entre cero y uno dependiendo de si la reputación es mejor o peor según el portal citado, siendo uno el valor asignado a la peor reputación posible.
   f) Subcriterio **Datos del routing.** Asignaremos un valor entre cero y uno en función de la información de rutado intermedio de un determinado correo entre el emisor y el receptor, teniendo en cuenta los saltos intermedios y su reputación. Siendo uno el valor asignado en caso de que el correo pase por varios sitios de mala reputación.
   g) Subcriterio **Edad del dominio del enlace.** En éste caso asignaremos un valor entero que indicará el número de meses que tiene el dominio del enlace que contiene el correo. A mayor valor, más antiguo será el dominio, ya que por lo general los dominios de enlaces de phishing suelen ser muy recientes para evitar las listas negras y las asignaciones de reputación. En casos concretos de spear phishing, se suelen tratar de dominios legítimos con una cierta antigüedad que no encajarían a priori con lo comentado, pero se ha detectado que en muchos casos se ha producido una modificación reciente del dominio que llevaría a sospechar. Modificaciones para introducir por ejemplo registros MX de servidores maliciosos de modo que luego puedan pasar los filtros basados en pruebas de SPF y DKIM.
   h) Subcriterio **Número de subdominios del dominio modal.** Asignaremos un valor entero que indicará el número total de subdominios sin contar el dominio raiz, de un dominio de correo determinado.
   i) Subcriterio **País del dominio vs país de la IP.** Tras geolocalizar el dominio modal, su IP y el dominio del enlace, asignaremos cero si son coincidentes y uno si no coinciden para nada. Normalmente los correos legítimos, salvo excepciones, provienen de IPs y dominios del mismo país.
   j) Subcriterio **Coincidencia de dominio modal y de enlace.** Si el dominio del enlace en el cuerpo del correo, a pesar de no tener mala reputación, no coincide con el dominio modal del remitente, es muy probable que estemos ante un correo con spear phishing. Por ello asignaremos un cero en caso de no coincidencia y un valor de uno si coinciden. En el caso de que haya más de un enlace en el cuerpo se realizará la media aritmética de los valores de

los mismos.

2. **Criterios basados en errores en el cuerpo del correo** [Peso asignado: 10 %]
   a) Subcriterio **Errores ortográficos.** Bien debido a la mala ortografía o a otros aspectos como que los atacantes sean de otros países y usen otras páginas de códigos, tradicionalmente ha constituido un claro indicador de correo malicioso. Asignaremos un valor entero con el número de faltas encontradas, bajo la premisa de que un remitente legítimo no suele tenerlas, por reputación e imagen. A mayor valor, mayor probabilidad de spear phishing.
   b) Subcriterio **Frases comunes y palabras clave.** Existen diversas frases y palabras que suelen ser empleadas por los atacantes para inducir a la víctima a pulsar en un determinado enlace y lograr la descarga del malware o de cualquier herramienta previa para abordar el sistema. Asignaremos un valor entero con el número de palabras y/o frases encontradas en el cuerpo del correo.
   c) Subcriterio **Inconsistencias y errores gramaticales.** En muchas ocasiones se recurre a traducciones que no son correctas, apareciendo frases que en nuestro idioma pueden resultar inconsistentes o chocantes, así como errores gramaticales derivados de una escritura de baja calidad. Asignaremos un número entero con el total de detecciones realizadas en el texto.
   d) Subcriterio **Técnicas de ofuscación.** Si el texto del cuerpo o incluso el código html del mismo permanece oculto mediante el uso de ciertas técnicas de codificación, asignaremos un valor de uno si están presentes o de cero si no lo están.
3. **Criterios basados en el reconocimiento de elementos multimedia y otros** [Peso asignado: 10 %]
   a) Subcriterio **Calidad y tipo de imágenes.** Asignamos un valor entre cero y diez en función de las imágenes contenidas y su calidad, siendo cero la menor calidad de las mismas.
   b) Subcriterio **Calidad/Reputación de los enlaces.** Asignaremos un valor comprendido entre cero y diez en función de la reputación obtenida, tras consultar los portales de inteligencia, para el enlace o los enlaces contenidos en el cuerpo del correo. En caso de haber más de un enlace, se hará la media aritmética la reputación de todos ellos. El valor diez corresponde a la peor reputación posible.
   c) Subcriterio **Calidad de los ficheros adjuntos.** En este caso y tras un análisis realizado generalmente por el antivirus y la sandbox del filtro de correo, se obtiene un resultado que viene detallado generalmente en los metadatos del correo, donde se indica si se ha encontrado algún tipo de amenaza en los archivos adjuntos. No es común en los correos de Phishing encontrar archivos adjuntos maliciosos, dado que éstos son generalmente fáciles de interceptar. Asignaremos un valor entre cero y diez, representando cero el peor grado de maliciosidad posible.
   d) Subcriterio **Beacons añadidos.** Al igual que en el criterio anterior, se analizan la existencia de beacons en el correo y se le asigna una puntuación entre cero y uno, siendo cero el valor que representa la inexistencia de los mismos.
   e) Subcriterio **Brand Impersonation.** Es común en los correos de Phishing el tratar de suplantar una identidad de marca, no solo en el propio diseño del correo sino en los dominios de los enlaces, para los que se usan trucos de modo que a simple vista no se pueda distinguir del dominio real, empleando letras similares y demás. En éste caso, si se detecta la suplantación, asignaremos un uno. Siendo cero en el caso contrario.
4. **Criterios basados en factores psicológicos** [Peso asignado: 35 %] Están relacionados con las características generales de persuasión. Mediante análisis NLP podemos valorar si se aprecia persuasión o ingeniería social en el texto del cuerpo, que indique el hecho de querer provocar el comportamiento que requiere el atacante.
   a) Subcriterio **Urgencia.** Si existe algún texto en el cuerpo del correo que refleja el hacer algo con urgencia se valora con 1 y si no existe con 0.
   b) Subcriterio **Inducción al miedo mediante amenazas.** Si existe algún tipo de inducción al miedo mediante amenazas se valora con 1 y si no existe se valora con 0.
   c) Subcriterio **Seducción a hacer algo mediante incentivos.** Si existe algún tipo de seducción de este tipo se valora con 1 y si no existe con 0.
5. **Criterios basados en la persuasión personalizada** [Peso asignado: 35 %]
   a) Subcriterio **Explotación de la información personal publicada.** Se valora si aparece algún tipo de información del receptor del correo publicada en abierto o incluso la simulación de que existe la misma. En caso positivo se valora con 1 y si no aparece 0.
   b) Subcriterio **Autoridad.** Una autoridad gubernamental del país donde reside el receptor del correo, una gran compañía de servicios que utiliza el receptor o la propia organización de la persona a la que va dirigido el correo apremia con potestad a que el receptor haga alguna acción. La valoración será uno si existe y cero si no.
   c) Subcriterio **Familiaridad.** Se valora el parecido de los nombres de las páginas Web o los nombres de ficheros adjuntos con que esta familiarizado el receptor del correo, en caso que exista se valora con 1 y si no existe con 0.
   d) Subcriterio **Social Proof** (se imita en el cuerpo

del correo lo que el receptor piensa o como actúa el autor del correo electrónico) En caso que se presente la prueba social se valora con 1 y si no con cero.

IV. Proceso de Decisión y Resultados de Decisión

Podemos considerar que todos los anteriores criterios nos permitirán, bien de forma automática o bien de forma manual detectar un correo que es spear phishing, dependiendo de si se establecen mecanismos o herramientas para determinar automáticamente o no los valores. Esta detección especializada conlleva aparejado **un proceso de decisión y unos resultados de decisión** y que ya aparecen en la bibliografía.

El Proceso de Decisión consta de las siguientes fases como podemos ver en la figura 2:

1. Recolección de los correos legítimos, de phishing y de spear phishing a clasificar
2. Valoración de los criterios establecidos en la sección anterior para cada uno de los correos a clasificar
3. Aplicación de los métodos de decisión multicriterio TOPSIS y VIKOR
4. Análisis de los rankings obtenidos por ambos métodos

A continuación, detallamos dicho proceso para un conjunto de correos electrónicos.

En primer lugar, en cuanto a la recolección de correos para clasificar y ante la dificultad de obtener, como comentamos anteriormente, un conjunto de correos electrónicos recibidos (dataset) con el que probar exhaustivamente el método, hemos calculado manualmente todos los criterios de un grupo reducido de 10 correos: 2 correos legítimos, 6 correos phishing, y 2 correos de spear phising. Destacando que estos correos han sorteado con éxito el filtro de seguridad corporativo[1] y que los propios usuarios nos han cedido para su análisis. Dichos correos tienen, a priori, características que los diferencian de los correos de phishing habituales y que evitan que sean descartados por el filtro, pero salvo dos o tres de ellos, no podemos considerarlos como verdaderos spear phishings. Son correos que han logrado pasar el filtro por tener ciertas características que impiden su rechazo, como el proceder de dominios legítimos o no incluidos en listas negras, tener enlaces con reputación neutral o buena, etc.

Seguidamente, en cuanto a la valoración de criterios, decir que inicialmente le hemos dado un peso equitativo a cada criterio y a cada subcriterio, sin ponderar ninguno de ellos para probar el funcionamiento de TOPSIS y VIKOR. En primera instancia, se le asignó un peso del 20 por ciento a cada bloque y consecuentemente ese porcentaje se dividió entre los subcriterios que componen cada bloque. Tras las primeras pruebas, no se obtuvieron resultados llamativos, especialmente porque los correos de phishing y spear phishing estaban todos en una horquilla relativamente cerrada. Para contextualizar un poco los resultados y después de comprobar que ambos métodos funcionaban correctamente, para lo que hemos usado dos correos sintéticos que simulaban la PIS y la NIS, decidimos añadir un par de correos legítimos para ver como se distanciaban de los correos de phishing, como así resultó. Ambos correos eran posicionados en el ranking elaborado por TOPSIS y VIKOR en el extremo donde se encontraba la NIS. Por ello, tras realizar diferentes pruebas de validación, tuvimos que sobreponderar los criterios de los bloques cuatro y cinco en conjunto sobre el resto, asignando un 35 % al bloque cuatro y un 35 % al bloque cinco. Aún así no obtuvimos resultados totalmente satisfactorios, por lo que tuvimos que sobreponderar algunos subcriterios de los tres primeros bloques. Subcriterios, a priori, muy relacionados con los correos de spear phishing. Tras varias iteraciones y múltiples pruebas, hemos obtenido los pesos que se enumeran en cada criterio en el apartado anterior, que a nuestro juicio son los más adecuados para determinar el tipo de correo analizado. Con dichos pesos hemos realizado las pruebas definitivas con TOPSIS y VIKOR que aparecen en las tablas que se muestran. Los pesos sobreponderan, principalmente y como hemos comentado, las características psicológicas de los correos (Bloques 4 y 5 de criterios), del mismo modo que un analista de seguridad le prestaría más atención a los mismos que a las propios criterios basados en metadatos o datos internos de los correos (Bloques 1, 2 y 3)

En tercer lugar y tras aplicar un peso a cada criterio en función de la importancia del mismo respecto a que sea un spear phishing o no, obtenemos la matriz de criterios, así como el vector de pesos y el de valoraciones a introducir en los cálculos de TOPSIS y VIKOR. Dichas matrices y vectores se ha introducido en una herramienta MDCM [23] y se obtuvieron los resultados de los rankings para TOPSIS en la Tabla I. En la primera columna aparece el orden inicial en el que se ha procesado cada correo, mientras que en la segunda aparece el tipo de correo de que se trata. En la tercera aparece el valor R calculado para cada correo, usando el método comentado y en la cuarta y última columna el ranking que se ha obtenido tras ejecutar el proceso. Para el mismo conjunto de valores y criterios (mismas matrices y vectores) hemos obtenido los resultados con VIKOR [24] que se muestran en la Tabla II.En este caso en las dos primeras columnas tenemos el orden inicial y el tipo de correo al igual que en el caso de TOPSIS. Mientras que en éste caso, tenemos las columnas tercera, cuarta y quinta con los valores generados por el modelo, es decir, S, R y Q. Y finalmente, en la sexta columna 6 aparece el ranking que se obtenido tras la ejecución.

V. Discusión de los resultados de decisión

Para finalizar el proceso y como análisis de los resultados obtenidos y tras haber afinado los pesos aplicables a los criterios, esperábamos que el máximo representante de correo con spear phishing, concretamente el correo 8 de la lista, obtuviera la mayor puntuación con respecto al resto y se constituyera como la Solución Ideal Positiva (PIS), cosa que no ha ocurrido a la vista de los resultados obtenidos. Por contra, los correos 9 y 10 de la lista que son correos legítimos que hemos introducido de prueba, si que han sido clasificado como la Solución Ideal Negativa (NIS) uno de ellos y el

[1]Estos correos han sido extraídos de una administración/corporación real.

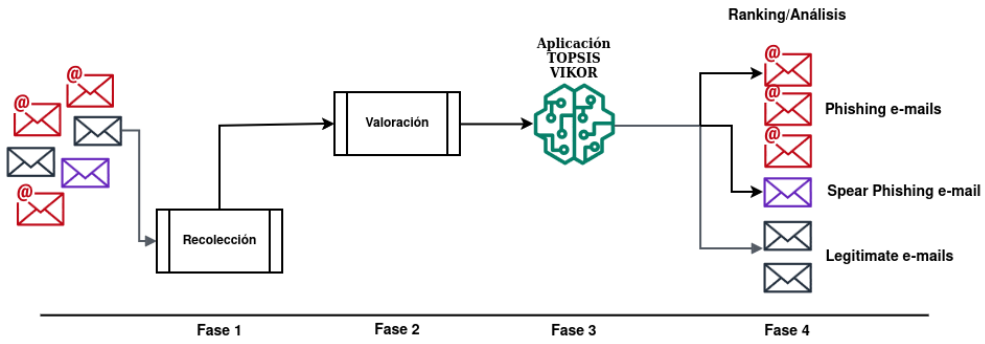

Figura 2. Fases del proceso de toma de decisión.

#	Alternatives	R	Ranking
1	Phishing1	0.65164225	4
2	Phishing2	0.65231107	3
3	Legítimo1	0.05343569	9
4	Phishing3	0.83325898	1
5	Legítimo2	0.05219731	10
6	Spear Phishing1	0.45938017	8
7	Phishing4	0.56159103	6
8	Spear Phishing2	0.53312071	7
9	Phishing5	0.82703226	2
10	Phishing6	0.64166342	5

Tabla I
RESULTADOS DE DECISIÓN SEGÚN TOPSIS.

#	Alternatives	S	R	Q	Ranking
1	Phishing1	0.3265047	0.1050	0.5918238	4
2	Phishing2	0.3120037	0.1050	0.5845733	3
3	Legítimo1	0.9642014	0.1225	0.9821007	9
4	Phishing3	0.1415017	0.0350	0.2136080	1
5	Legítimo2	0.9671017	0.1225	0.9835509	10
6	Spear Phishing1	0.5640002	0.1225	0.7820001	8
7	Phishing4	0.3455041	0.1225	0.6727520	6
8	Spear Phishing2	0.4275125	0.1225	0.7137563	7
9	Phishing5	0.1475003	0.0350	0.2166073	2
10	Phishing6	0.3035054	0.1225	0.6517527	5

Tabla II
RESULTADOS DE DECISIÓN SEGÚN VIKOR.

otro muy próximo, validando el proceso en ese extremo. Tras revisar de nuevo y de manera pormenorizada los pesos y los criterios de los correos, hemos observado que el método (en el caso concreto de TOPSIS) obtiene tres grupos de correos tras ejecutarse y elaborar el ranking:

- Un grupo de correos en el rango más alto, entre 0.64 y 0.84 que se corresponden con claros correos de phishing tradicional o clone phishing.
- Un grupo de correos en el rango más bajo (0.05), que se corresponde con los correos legítimos que hemos comentado.
- Un grupo intermedio de correos en el rango 0.46 a 0.56, donde tenemos los que se podrían considerar correos de spear phishing.

Por tanto, aunque inicialmente pudiera parecer que se está produciendo un error en la catalogación, en la realidad no es así y tiene además, bastante lógica. El modelo clasifica los correos de spear phishing como más próximos a los legítimos, quedando en una zona intermedia, lo cual es correcto por la simple razón de que los correos de spear phishing son bastante más parecidos a los correos legítimos que el resto de correos de phishing. Precisamente ese parecido es lo que los hace tan peligrosos y que consigan sortear los filtros de seguridad y el criterio humano. Y ese parecido es lo que provoca que habitualmente sean catalogados como correos legítimos por parte de los sistemas de seguridad corporativos que analizan los correos entrantes. En nuestro caso y a la vista de los resultados obtenidos con ambos métodos, podemos concluir que el sistema detecta los correos de spear phishing como correos de phishing pero más próximos a los legítimos, aunque guardando bastante distancia con ellos. Y eso es gracias a la ponderación dada a criterios inherentes al propio correo en los tres primeros bloques de los mismos, como complemento efectivo de las valoraciones realizadas de los criterios más “humanos” en los bloques cuatro y cinco.

VI. CONCLUSIONES Y TRABAJOS FUTUROS

Una vez llevada a cabo la aplicación de los sistemas de decisión multicriterio a la detección de correos de spear phishing y phishing podemos concluir que los resultados obtenidos son bastantes prometedores y validan los objetivos planteados. Como trabajo futuro creemos que podría facilitar bastante la valoración de los criterios usando valores fuzzy en vez de discretos y aplicar las correspondientes métodos de Fuzzy TOPSIS y Fuzzy VIKOR para la clasificación de los correos spear phishing. Con los test llevados a cabo con el conjunto de datos concretos que disponemos hasta ahora hemos obtenido un porcentaje del cero por ciento, tanto de falsos positivos como de falsos negativos respecto al spear phishing y continuamos investigando para obtener un conjunto de datos más amplio con la idea de poder validar estos resultados de forma mucho más general. También se trabaja en automatizar las tareas de valoración de los criterios mediante herramientas relacionadas con el procesamiento del lenguaje

natural, análisis de etiquetas, enlaces y análisis dinámico del software adjunto en los correos. Todo ello se añadiría a los filtros actuales para mejorar la detección de los ataques de spear phishing.

Igualmente consideramos que éste estudio puede ser de utilidad en la detección de lo que se denomina BEC (Business Email Compromise), también conocido como lateral o internal spear phishing, que según el último informe del IC3 del FBI [25] es uno de los principales problemas de seguridad detectados y denunciados. Cuando se producen intentos de fraude, phishing y otros, perpetrados desde cuentas de correo internas de la propia organización, los efectos suelen ser demoledores, ya que normalmente el envío de de correos entre buzones propios de una determinada empresa no suele inspeccionarse con el detalle con el que se analizan los correos que entran desde remitentes externos. Pero sin embargo, dichos correos suelen contener características psicológicas y persuasivas como las que hemos analizado previamente, por lo que es probable que podamos neutralizar dichas amenazas mediante el uso de MCDM como hemos propuesto para los correos de spear phishing.

Agradecimientos

Esta publicación es parte de los proyectos AETHER-US PID2020-112540RB-C44 y ALBA-US TED2021-130355B-C32 financiado por MICIU/AEI/10.13039/501100011033 y "European Union NextGenerationEU/PRTR".

Referencias

[1] A. Greenberg. (2024) Ransomware payments hit a record 1.1 billion in 2023. [Online]. Available: https://www.wired.com/story/ransomware-payments-2023-breaks-record/

[2] Google. (2024) Directrices para remitentes de correos. [Online]. Available: https://support.google.com/a/answer/81126?hl=es

[3] T. Xu, K. Singh, and P. Rajivan, "Personalized persuasion: Quantifying susceptibility to information exploitation in spear-phishing attacks," *Applied Ergonomics*, vol. 108, p. 103908, 2023.

[4] G. Ho, A. Cidon, L. Gavish, M. Schweighauser, V. Paxson, S. Savage, G. M. Voelker, and D. Wagner, "Detecting and characterizing lateral phishing at scale," in *28th USENIX security symposium (USENIX security 19)*, 2019, pp. 1273–1290.

[5] J. M. Velo, A. Varela-Vaca, and R. M. Gasca, "Ciberseguridad cognitiva aplicada al phishing," 2023.

[6] C.-L. Hwang, Y.-J. Lai, and T.-Y. Liu, "A new approach for multiple objective decision making," *Computers & operations research*, vol. 20, no. 8, pp. 889–899, 1993.

[7] S. Opricovic, "Vikor method in multi-criteria optimization of civil engineering systems. the university of belgrade, faculty of civil engineering, 175 p," 1998.

[8] S. M. Mueller, J. Schiebener, M. Delazer, and M. Brand, "Risk approximation in decision making: approximative numeric abilities predict advantageous decisions under objective risk," *Cognitive processing*, vol. 19, pp. 297–315, 2018.

[9] G. F. Barberis and M. d. C. E. Ródenas, "La ayuda a la decisión multicriterio: orígenes, evolución y situación actual," in *VI Congreso Internacional de Historia de la Estadística y de la Probabilidad*, 2011.

[10] C.-L. Hwang and K. Yoon, *Multiple attribute decision making: methods and applications a state-of-the-art survey*. Springer Science & Business Media, 2012, vol. 186.

[11] K. Yoon, "A reconciliation among discrete compromise solutions," *Journal of the Operational Research Society*, vol. 38, pp. 277–286, 1987.

[12] K. Palczewski and W. Sałabun, "The fuzzy topsis applications in the last decade," *Procedia Computer Science*, vol. 159, pp. 2294–2303, 2019.

[13] M. do Carmo Silva, C. F. Simões Gomes, and C. L. da Costa Junior, "The use of topsis for ranking wipo's innovation indicators," *Innovar*, vol. 29, no. 73, pp. 133–147, 2019.

[14] H.-S. Shih and D. L. Olson, *TOPSIS and its extensions: A distance-based MCDM approach*. Springer Nature, 2022, vol. 447.

[15] A. Mardani, E. K. Zavadskas, K. Govindan, A. Amat Senin, and A. Jusoh, "Vikor technique: A systematic review of the state of the art literature on methodologies and applications," *Sustainability*, vol. 8, no. 1, p. 37, 2016.

[16] S. Opricovic and G.-H. Tzeng, "Extended vikor method in comparison with outranking methods," *European journal of operational research*, vol. 178, no. 2, pp. 514–529, 2007.

[17] J. Wang, Y. Li, and H. R. Rao, "Overconfidence in phishing email detection," *Journal of the Association for Information Systems*, vol. 17, no. 11, p. 1, 2016.

[18] G. D. Moody, D. F. Galletta, and B. K. Dunn, "Which phish get caught? an exploratory study of individuals susceptibility to phishing," *European Journal of Information Systems*, vol. 26, pp. 564–584, 2017.

[19] D. Hillman, Y. Harel, and E. Toch, "Evaluating organizational phishing awareness training on an enterprise scale," *Computers & Security*, vol. 132, p. 103364, 2023.

[20] T. Xu, K. Singh, and P. Rajivan, "Personalized persuasion: Quantifying susceptibility to information exploitation in spear-phishing attacks," *Applied Ergonomics*, vol. 108, p. 103908, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003687022002319

[21] A. Ferreira, L. Coventry, and G. Lenzini, "Principles of persuasion in social engineering and their use in phishing," in *Human Aspects of Information Security, Privacy, and Trust: Third International Conference, HAS 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015. Proceedings 3*. Springer, 2015, pp. 36–47.

[22] Cisco Systems. (2022) Cisco talos. [Online]. Available: https://www.talosintelligence.com/

[23] I. Howson. (2024) Topsislinear: Implementation of topsis method for multi-criteria decision. [Online]. Available: https://rdrr.io/cran/MCDM/man/TOPSISLinear.html

[24] ——. (2024) Vikor: Implementation of vikor method for multi-criteria decision. [Online]. Available: https://rdrr.io/cran/MCDM/man/VIKOR.html

[25] F. IC3. (2023) Fbi internet crime report. [Online]. Available: https://www.ic3.gov/Media/PDF/AnnualReport/2023_IC3Report.pdf