FARO

22/07/2019

Herramienta de detección de información sensible

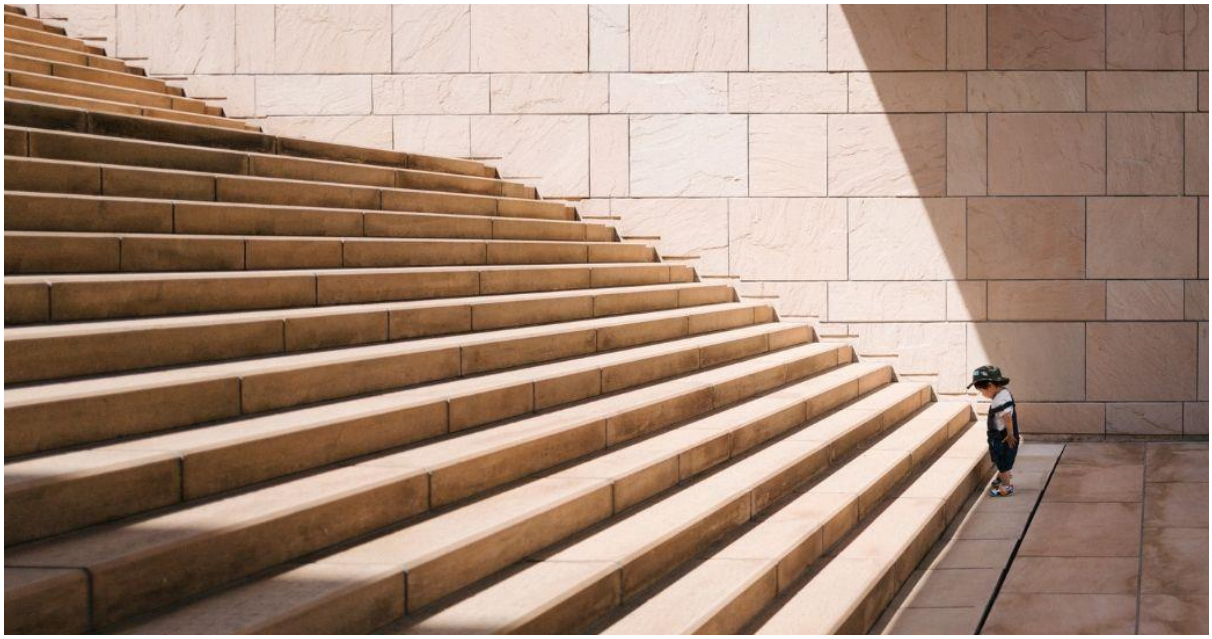
Rúa Fonte das Abelleiras, s/n.
Edificio CITEXVI, local 14
36310 Vigo, Pontevedra, Spain

(+34) 986 120 430
www.tegra.org
tegra@tegra.org

Introducción

Las organizaciones generan y gestionan un gran volumen de documentación acorde con su actividad diaria. Una parte importante puede tener carácter confidencial o estratégico: contratos, acuerdos, facturas, cuentas de resultados, presupuestos, datos personales de los trabajadores, etc. Se trata de datos altamente sensibles que, mal protegidos, pueden suponer un importante problema de reputación/seguridad para la organización.

La RGPD (Reglamento General de Protección de Datos) obliga a las organizaciones a tomar medidas proactivas para asegurar dicha protección. Y, aunque cada vez existe una mayor preocupación en este sentido, la realidad es que no abundan soluciones que permitan la gestión sencilla de documentación sensible, y menos aún soluciones que una pequeña organización pueda permitirse utilizar.



Este informe presenta una aproximación técnica al problema de la detección de información sensible dentro de los documentos. En él se describe una prueba de concepto creada en el seno de TEGRA denominada FARO. Esta herramienta pretende servir de punto de partida para pequeñas organizaciones que, siendo conscientes del riesgo, carecen de alternativas para mejorar un aspecto crítico de su seguridad.

Estado del arte

En el mercado podemos ver distintas aproximaciones a esta problemática. Algunas de ellas están basadas en sistemas de clasificación de Machine Learning (ML). Estos sistemas aprenden una función que les ayuda a discriminar entre datos sensibles y no sensibles a partir de un histórico de datos con las etiquetas buscadas (p.e. sensible/no sensible). Este es el

caso de la solución de Symantec[1] donde se generan modelos específicos para una organización usando como entrada sus propios datos.

Otros sistemas están basados en reglas. Extraen información de los documentos en base a patrones habituales (p.ej. los pasaportes y DNIs siguen la misma estructura para un país concreto). En esta línea está la solución de Varonis[2], donde se emplean patrones y expresiones regulares para extraer un gran número de indicadores para la detección de documentos sensibles tanto en el contenido del documento como en sus metadatos.

La desventaja de usar soluciones de ML para resolver este problema es que necesita grandes volúmenes de entrenamiento y de actualización continua. Esto puede ser una barrera para empresas pequeñas que no pueden dedicar recursos a hacer dicho etiquetado (o que no disponen de suficientes documentos de una determinada clase). Además, las distintas clases deben ser lo suficientemente separables para que le sea fácil aprender una función que discrimine bien entre ellas.

Nuestra aproximación

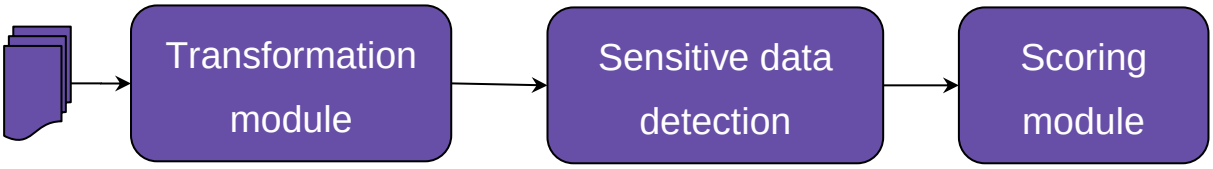


Figura 1. Arquitectura de FARO

Desde TEGRA hemos optado por realizar un sistema basado en la extracción de entidades y sus relaciones (p.ej. personas en una organización con el cargo que ocupan en la misma) y a la extracción de una serie de patrones (p.e. números de teléfono, emails, cuentas bancarias, documentos de identidad, etc) para generar indicadores objetivos que permitan detectar documentos sensibles en una organización. Estos indicadores son comunes a todo tipo de organizaciones, por lo que es más fácil emplearla sin costosas adaptaciones a cada caso concreto.

Estos sistemas que se basan en patrones y reglas generan pocos falsos positivos (documentos que son detectados como sensibles cuando no lo son) aunque es más fácil que se le escape información que no está cubierta por las reglas.

[1] Symantec Data Loss Prevention https://www.symantec.com/products/data-loss-prevention

[2] Varonis Data Classification Engine https://www.varonis.com/products/data-classification-engine/

Detección de información sensible

Dado un documento, el primer objetivo es detectar de forma automatizada las entidades que puedan portar información sensible. Para ello FARO engloba un conjunto de técnicas que nos ayudan a detectar cada categoría de dato potencialmente sensible.

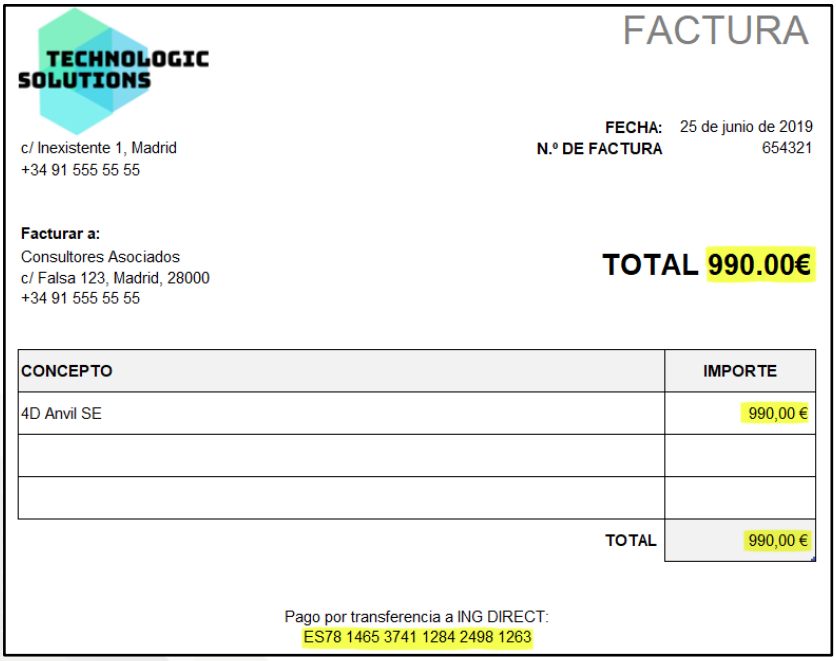
En la tabla que mostramos a continuación detallamos las entidades que cubre FARO junto la técnica empleada para su detección:

Tipo de Entidad	Método para la detección
Personas, cargo y organización	NER con CRFs
Firmas	RegExp
Cantidades Monetarias	RegExp
Números de teléfono móvil	RegExp
Tarjetas de crédito	RegExp + Luhn
Números de cuenta	RegExp
Documentos de identidad (DNI, CIF)	RegExp
Emails personales	RegExp + Clasificador

Tabla 1. Conjunto de entidades detectadas y métodos para la detección.

Ejemplo

El objetivo de FARO es detectar en una factura como la que vemos a continuación las entidades resaltadas:



TECHNOLOGIC SOLUTIONS

c/ Inexistente 1, Madrid
+34 91 555 55 55

FACTURA

FECHA: 25 de junio de 2019
N.º DE FACTURA 654321

Facturar a:
Consultores Asociados
c/ Falsa 123, Madrid, 28000
+34 91 555 55 55

TOTAL 990.00€

CONCEPTO	IMPORTE
4D Anvil SE	990,00 €
TOTAL	990,00 €

Pago por transferencia a ING DIRECT:
ES78 1465 3741 1284 2498 1263

Rúa Fonte das Abelleiras, s/n.
Edificio CITEXVI, local 14
36310 Vigo, Pontevedra, Spain

(+34) 986 120 430
www.tegra.org
tegra@tegra.org

Uso de FARO

Algunas de estas categorías de datos deberían obligar a clasificar al documento que lo contenga como sensible de forma inmediata (p.ej: Tarjeta de crédito) sin embargo, otros solamente deberían significar un aporte bajo a la sensibilidad del documento final (p.ej: cantidades monetarias) por lo que la ocurrencia es una variable que debemos tener en cuenta.

Hemos incorporado a la arquitectura de nuestra prueba de concepto un módulo de scoring que facilita dicha distinción. Éste módulo de scoring se encarga de ponderar la relevancia de cada tipo de dato detectado y su ocurrencia de cara a poder clasificar un determinado documento.

Algoritmo de scoring

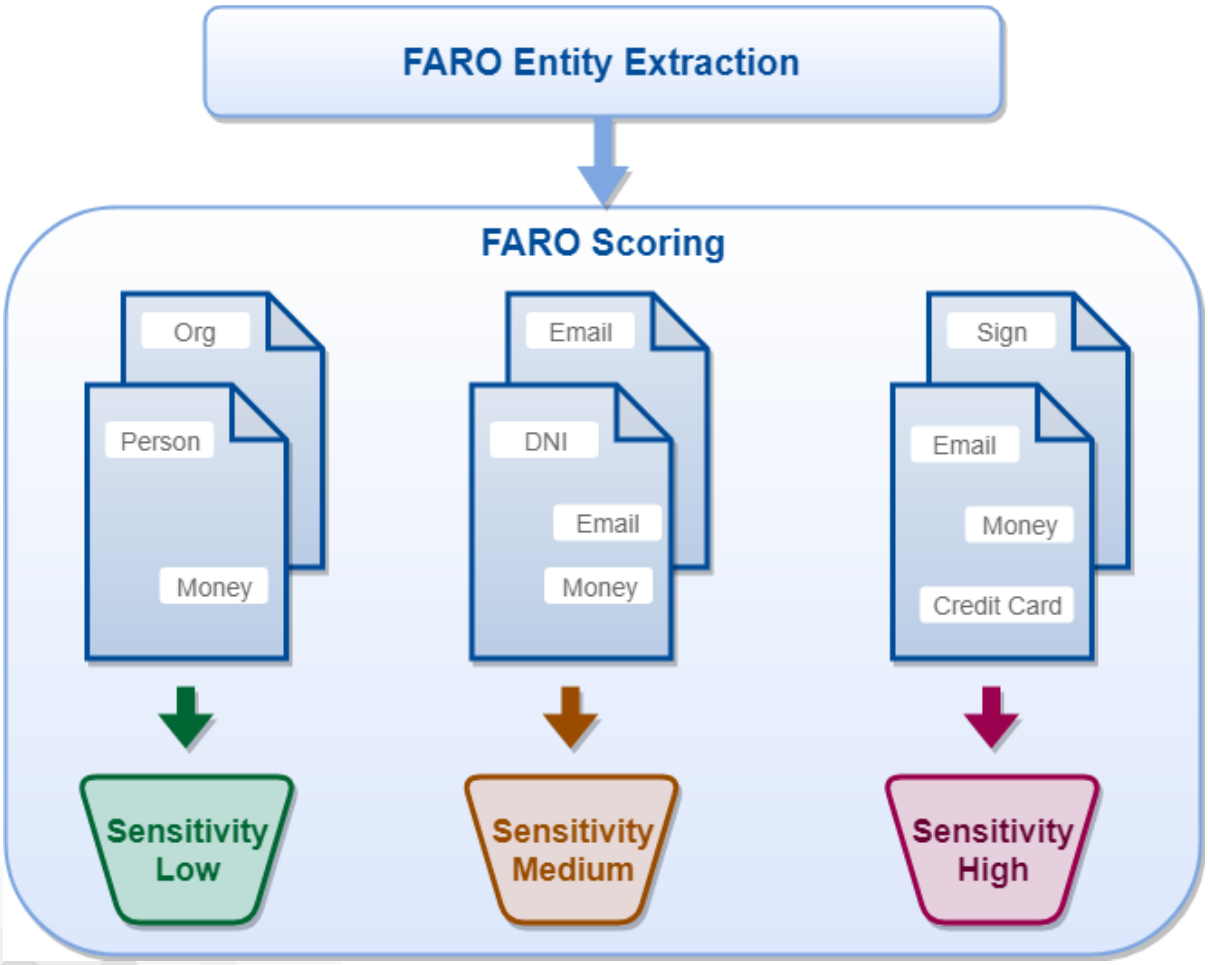


Figura 2. Clasificación de los documentos en niveles de confidencialidad con FARO

Quizá más importante que el diseño concreto de nuestro **algoritmo de *scoring*** (siempre opinable), éste está diseñado como **fácilmente configurable** para adaptarse al criterio de cada organización particular.

Detecciones realizadas	Baja	Media	Alta
Personas con cargo y organización	≤ 5	> 5	
Cantidades monetarias	≤ 5	> 5	
Emails Personales	≤ 1	1 < n ≤ 3	> 3
Números de móvil	≤ 1	1 < n ≤ 3	> 3
Firmas	0	≤ 1	> 1
Datos financieros	0	0	≥ 1
Documentos de identidad	0	0	≥ 1

Tabla 2. Número de detecciones por tipo de entidad[3].

Cuando observamos en un mismo documento la presencia de múltiples criterios simultáneos dentro de un nivel, existe la capacidad de elevar al siguiente nivel si así lo indicamos en la configuración.

Ejemplo de ejecución individual de FARO

Si utilizamos FARO con un ejemplo que representa un acuerdo de confidencialidad:

ACUERDO DE CONFIDENCIALIDAD

ENTRE

(Parte A)

Y (Parte B)

En Vigo, a 30 de Febrero de 3000

REUNIDOS

De una parte, Dª María González, con DNI 55555555-K, representando en este acto a la empresa Consultores Asociados, con sede en c/Falsa 123, y NIF B55555551 inscrita en el Registro Mercantil de Vigo con número 12345, y de la que es director adjunto de RRHH que le faculta para representar a la empresa y de la cual tiene concedido poder suficiente (en adelante "Parte A").

De otra parte, D. Alberto Pérez, con DNI 33333333P, representando en este acto a la empresa Technologic Solutions, con sede en c/Inexistente 1, y NIF A33.333.337 inscrita en el Registro Mercantil de Madrid con número 1 y de la cual tiene concedido poder suficiente, y de la que es presidente que le faculta para representar a la empresa (en adelante Parte B).

[3] Si se cumplen 3 o más criterios individuales la sensibilidad se escalará al siguiente nivel

La salida del sistema es la siguiente indicando que se han encontrado cuatro documentos de identidad y dos personas con su organización y cargo de la empresa. Siguiendo los umbrales de scoring especificados le corresponde un nivel alto de confidencialidad.


```
{
  "summary": {
    "person_jobposition_organization": 2,
    "id_document": 4
  },
  "score": "high"
}
```


Modo recursivo

FARO incluye un modo de ejecución mediante el que es posible analizar todos los documentos de una determinada carpeta o volumen de forma recursiva.

Además, en este modo recursivo las tareas de análisis se paralelizan aprovechando la existencia de múltiples núcleos de procesamiento dentro de la máquina en la que se lanza FARO, con el fin de conseguir una mayor velocidad. FARO utiliza GNU Parallel para realizar dicha paralelización, podéis obtener más información sobre GNU Parallel aquí[4].

Los resultados totales se entregan en un único fichero, en formato CSV, incluyendo la ruta del fichero, las puntuaciones en cada categoría y el score final para cada documento analizado, como se muestra en el siguiente ejemplo.


```
filepath,score,person_position_organization,monetary_quantity,signature,personal
_email,mobile_phone_number,financial_data,document_id
boe/dias/2019/01/08/pdfs/BOE-B-2019-422.pdf,high,0,0,0,1,0,0,3
boe/dias/2019/01/08/pdfs/BOE-B-2019-392.pdf,low,0,0,0,0,0,0,0
boe/dias/2019/01/08/pdfs/BOE-B-2019-427.pdf,low,0,0,0,0,0,0,0
```


También se genera un fichero de detalle con las entidades detectadas para cada documento analizado en formato JSON.


```
[
  {
    "filepath": "boe/dias/2019/01/08/pdfs/BOE-B-2019-422.pdf",
    "datetime": "2019-06-28 13:22:37",
    "entities": {
      "document_id": {
        " X4813918H ": 2,
        " X4658630A ": 1,
        "  X4658630A,": 1
      },
      "email": {
```


[4] GNU Parallel

```
        "merpastor@icav.es": 1
      }
    }
  },
  {
    "filepath": "boe/dias/2019/01/08/pdfs/BOE-B-2019-392.pdf",
    "entities": {},
    "datetime": "2019-06-28 13:22:48"
  },
  {
    "entities": {
      "email": {
        "soia@csic.es": 1
      },
      "document_id": {
        " B07411598.": 1
      },
      "monetary_quantity": {
        "61.305,12 euros": 1,
        "66.500,47 euros": 1,
        "0,00 euros": 1
      }
    },
    "filepath": "boe/dias/2019/01/08/pdfs/BOE-B-2019-483.pdf",
    "datetime": "2019-06-28 13:22:51"
  }
]
```


Limitaciones

FARO es una herramienta desarrollada como prueba de concepto y, por tanto, sujeta a una serie de limitaciones:

- La herramienta está **orientada al idioma español** ya que los modelos de detección de entidades (detección de Personas, Organizaciones y Cargos) se han entrenado sólo para ese idioma. Entrenar los modelos para múltiples idiomas es un proceso costoso que escapa al alcance de esta prueba de concepto.
- Los modelos entrenados fallan ocasionalmente (p.e. no todas las entidades son extraídas correctamente). Esto puede hacer que no todos los indicadores de sensibilidad sean reales, afectando al ranking final del documento. Si bien los modelos pueden mejorar con más datos de entrenamiento, siempre puede haber errores, ya que trabajan con un contexto limitado. En todo caso, documentos altamente sensibles, como facturas o contratos, contienen suficientes indicadores para que su detección como información sensible sea positiva independientemente de que se cometan errores puntuales.
- Nos hemos limitado a un **conjunto reducido de indicadores** para la prueba de concepto. Sin embargo, el sistema es fácilmente extensible a nuevos patrones y tipos de indicadores a detectar en un futuro.

Evolución

Bajo nuestro punto de vista algunas de las mejoras de FARO podrían seguir las siguientes líneas:

- Integración con herramientas de correo u ofimáticas a modo de plugins para poder revisar los documentos que se van a compartir de forma instantánea.
- Incorporar capacidades de descubrimiento de documentos: NAS, crawling, etc.
- Mejorar el rendimiento del modo recursivo.
- Inventario de documentos analizados para evitar análisis redundantes.

Hemos liberado el código fuente de FARO, bajo licencia MIT, con el afán de que cualquier desarrollador interesado en esta prueba de concepto pueda contribuir a su evolución a través de la extracción de nuevas entidades, mejoras en la precisión de detección de las actuales o contribuir en cualquiera de las líneas propuestas anteriormente.

Autores

- Juan Elosua Tomé – Director por parte de ElevenPaths del centro I+D en Ciberseguridad TEGRA de Galicia.
- Rafael P. Martínez Álvarez – Investigador de ciberseguridad del centro tecnológico Gradiant, partner de ElevenPaths en TEGRA.
- Héctor Cerezo Costas – Investigador de ciberseguridad del centro tecnológico Gradiant, partner de ElevenPaths en TEGRA.

TEGRA cybersecurity center se enmarca en la unidad mixta de investigación en ciberseguridad IRMAS (Information Rights Management Advanced Systems), que está cofinanciada por la Unión Europea, en el marco del Programa Operativo FEDER Galicia 2014-2020, para promover el desarrollo tecnológico, la innovación y una investigación de calidad.