PLIEGO DE CLÁUSULAS TÉCNICAS PARTICULARES DE LA FUNDACIÓN CENTRO TECNOLÓXICO DE TELECOMUNICACIÓNS DE GALICIA (GRADIANT) PARA LA CONTRATACIÓN DE SERVICIOS POR PROCEDIMIENTO NEGOCIADO CON PUBLICIDAD

En el marco de la cuarta convocatoria de la Compra Pública Precomercial lanzada por el Instituto Nacional de Ciberseguridad (INCIBE), Gradiant ha sido adjudicataria del proyecto fAIr: "Fight Fire with "fAIr": Suite de herramientas basadas en IA multimodal para combatir las amenazas derivadas del uso malicioso de la IA.

El objetivo final de este proyecto es el desarrollo de una plataforma de IA multimodal confiable, robusta y centrada en los datos que permita desarrollar modelos específicos para enfrentar los peligros asociados a un uso malintencionado de la IA. El proyecto incluye la creación de varias herramientas o módulos de detección de Deepfakes de audio, video, imagen, texto, así como demostradores y versiones de pruebas, detección de contenidos sintéticos, protección de contenidos y el análisis de trazabilidad de estos contenidos.

El objeto de esta licitación es la contratación de servicios de desarrollo de frontend y backend para la creación de una plataforma web, en la que se deben integrar los demostradores y versiones de pruebas, incluyendo gestión de usuarios, así como un panel de control en el que se pueda contar con estadísticas de uso y resultados de las tecnologías, desde el que se pueda monitorizar también el uso de las herramientas, todo ello unificado.

1. Funcionalidades de la plataforma

El diseño y desarrollo de la plataforma tiene que dar cabida a las funcionalidades descritas a continuación.

1.1. Funcionalidades

- Detección de contenido sintético creado por redes generativas de vídeo, imagen, audio y texto, lo que permitirá alertar sobre su origen y autenticidad. Se recibe como entrada un contenido (texto/documento, audio, video y/o imagen) y se recibirá información acerca de si es auténtico e información adicional como puntuaciones, o imágenes indicando información visual de relavancia).
- Trazabilidad de contenido sintético (imágenes, vídeos y audio) mediante el uso de técnicas forenses a través de la detección de trazas asociadas a aplicaciones de mensajería y de RRSS, lo que permitirá determinar procedencia, distribución y publicación de los contenidos. La entrada serán contenidos de audio, vídeo y/o imagen, y la respuesta contendrá información textual indicando puntuaciones u otra información de relevancia.
- Protección de contenido frente a la IA utilizando técnicas como “data poisoning”, que añaden cambios sutiles e invisibles a aquel contenido multimedia que queramos subir a RRSS, con el objetivo de confundir a los algoritmos de IA y hacer más difícil la generación de contenido malicioso (p.ej. contenido sexual falso). Recibirá como entrada audio, imagen y/o vídeo y se obtendrá una versión ligeramente diferente de ese contenido junto con información textual adicional a visualizar en la pantalla. Probablemente se

requiera de un selector indicando la herramienta contra la que se quiere proteger el contenido.

1.2. Formato de los contenidos de entrada

Los formatos de entrada serán texto/documentos, audio, vídeo o imagen. No todas las funcionalidades permiten todos los contenidos.

Los textos podrían bien escribirse o ser añadidos en un documento/imagen.

1.3. Salidas y visualización

Todas las funcionalidades obtendrán como salida información textual indicando puntuaciones o probabilidades. Algunos de ellos podrán mostrar imágenes (cómo mapas de calor) o vídeos indicando las regiones sospechosas o relevantes para tomar la decisión.

2. Integración de la plataforma

Todas las funcionalidades estarán desplegadas en servicios REST que recibirán el contenido a procesas y devolverán un JSON con la información necesaria para que la plataforma pueda mostrar el resultado.

3. Gestión de usuarios y permisos

La gestión de usuario se realizará a través de un token que será creado fuera de la plataforma. Existirá un servicio que proveerá Gradiant para validar si el token permite el acceso o no. Será también necesario que el usuario acepte las condiciones de uso de la plataforma e indique si permite el almacenamiento de los datos.

Cada token tiene asociado permisos de uso. Por lo tanto, la plataforma debe contemplar limitar el acceso a:

- Uso de contenidos de entrada. El usuario puede tener solo acceso a un tipo de contenidos limitados (texto/documentos, audio, imagen, vídeo)
- Funcionalidades. El usuario puede tener acceso a limitado a funcionalidades (trazabilidad, protección, detección).

Estos permisos están asociados al token, por lo que en el momento de solicitar la petición al servicio REST, este devolverá un error indicando la falta de permisos.

La plataforma tendrá que gestionar correctamente estos errores/alertas para informar al usuario la falta de permisos.

4. Procesos en lotes

La plataforma tendrá en cuenta que los usuarios podrán hacer uso del procesado en lotes para facilitar la ingesta de múltiples ficheros y evitar al usuario tener que procesos los contenidos de uno en uno.

5. Modo simulado

La plataforma permitirá mostrar las funcionalidades usando contenidos precargados para los usuarios que accedan sin un token, de manera que puedan entender el funcionamiento de la web.

Será un entorno completamente idéntico al modo de acceso con token, salvo que sólo podrán escoger entre un conjunto limitados de contenidos a procesar.

6. Compatibilidad con dispositivos móviles

La plataforma debe ser usable desde dispositivos móviles.

7. Dashboard / panel de control

Se desarrollará un panel de control asociado a cada token para visualizar estadísticas de uso.

Se permitirá acceso a los resultados de pruebas realizadas con anterioridad, siempre y cuando el usuario haya permitido el almacenamiento de los datos utilizados.

Gradiant podrá acceder con acceso de administrador y visualizar todos los usos y datos relacionados con todos los tokens.