**NOMBRE DEL PROYECTO**

| Desarrollo de una herramienta de análisis automático de texto para usuarios inexpertos del ámbito judicial. |
| --- |

**SÍNTESIS DEL PROYECTO**

| El proyecto consistirá en el desarrollo de un sistema informático que utilice técnicas de inteligencia artificial, en particular procesamiento de lenguaje natural o NLP, para analizar texto en el ámbito judicial. Este sistema está pensado como una herramienta para el análisis de evidencia textual en investigaciones judiciales. Servirá como una interfaz orientada a usuarios no expertos, que pueda incorporar modelos de análisis de texto con diferentes objetivos.  Se implementarán modelos para la identificación de posible violencia de género y para la detección de entidades. Para esto se crearán y etiquetarán sets de datos para el entrenamiento de los mismos.  Finalmente, estos módulos se pondrán a disposición del usuario no experto a través de una interfaz amigable en un sistema web. |
| --- |

**EQUIPO DE PROYECTO**

Integrantes estudiantes

| **Nombre** | **DNI** | **Carrera** |
| --- | --- | --- |
| Fernández, Valentina | 41854445 | Ingeniería Informática |
| Buendia, Pablo Agustín | 40254945 | Ingeniería Informática |

Director

| Ing. Bruno Constanzo |
| --- |

Co-Director

| Ing. Martín Castellote |
| --- |

Referente Funcional

| Lic. Lucía Belén Algieri |
| --- |

Demandante

| InFo-Lab Laboratorio de Investigación y Desarrollo de Tecnología en Informática Forense. |
| --- |

**PROBLEMA A RESOLVER**

| Análisis de texto para usuarios del ámbito judicial que no tienen acceso o conocimiento sobre el manejo de herramientas de inteligencia artificial, que les permitan realizar estos análisis de altos volúmenes de datos en cortos periodos de tiempo, para tratar con casos reales. |
| --- |

**OBJETIVO DEL PROYECTO**

| El proyecto tiene como objetivo el desarrollo de un sistema informático que utilice técnicas de Inteligencia Artificial para analizar textos en el ámbito judicial a través de una interfaz web. En particular, se pondrá el foco en los casos de violencia de género.  Se busca ofrecer a usuarios inexpertos una interfaz de usuario que les permita usar de manera simple tecnología de procesamiento de lenguaje natural. Y que estos usuarios puedan disponer de una herramienta que logre analizar, en un corto periodo de tiempo, altos volúmenes de datos. Se espera crear un sistema web donde se carguen lotes de documentos, y sea posible decidir con qué modelos de análisis de lenguaje se quieren procesar los datos para obtener resultados. Además, se busca poner énfasis en la usabilidad del producto, brindando diferentes interfaces para la visualización de los datos, haciendo a los resultados fáciles de analizar. |
| --- |

**PRODUCTO ESPERADO**

| El producto esperado consiste en un sistema web a disposición de usuarios judiciales no expertos en el uso de herramientas forenses de inteligencia artificial. Este sistema presentará al usuario una interfaz gráfica amigable a donde cargar lotes de documentos y le permitirá elegir entre distintos modelos de análisis de lenguaje para procesarlos. Estos documentos pueden ser presentados de diversos formatos, si por ejemplo estos estuvieran comprimidos, el sistema realizará la descompresión y el consiguiente análisis de cada uno de ellos. Una vez efectuado el análisis, el sistema ofrecerá distintas visualizaciones de los resultados obtenidos para que los usuarios puedan hacer inferencias de manera sencilla.  En el desarrollo de este proyecto se implementarán en principio dos modelos de análisis:  En primer lugar, un modelo de detección de entidades que buscará detectar, clasificar y mostrar entidades relevantes presentes en un texto (organizaciones, personas, lugares, fechas, entre otros). Posteriormente, brindará una visualización sencilla de las entidades detectadas junto con estadísticas de las mismas.  En segundo lugar, se creará un modelo de detección de violencia de género. Este modelo buscará detectar comportamientos violentos, presencia de amenazas, y clasificar distintos tipos de violencia encontrada (psicológica, simbólica, física, sexual, etc.). Con el objetivo de armar un informe preliminar sobre la posible existencia de violencia de género. Siendo de utilidad en investigaciones judiciales que involucran la violencia por motivos de género.  Eventualmente, pueden ser agregados nuevos modelos de análisis dependiendo del relevamiento de requerimientos realizado una vez iniciado el proyecto y los tiempos de desarrollo empleados. |
| --- |

**CARACTERÍSTICAS DE INNOVACIÓN LOCAL DEL PROYECTO**

| En el ámbito judicial local, el análisis de textos se realiza principalmente de dos maneras: sin el uso de herramientas de automatización, en cuyo caso depende de los usuarios incorporar, analizar y clasificar los datos; o, por otro lado, mediante el uso de herramientas complejas como Autopsy, que requieren del manejo de usuarios expertos con mayores conocimientos informáticos.  Mediante la incorporación de esta herramienta de diagnóstico a disposición directa del usuario no experto, se permitirá proporcionar información objetiva y rápida sin necesitar conocimiento previo de IA o del manejo de herramientas complejas. Esto proporcionará un mayor grado de automatización del proceso de análisis y diagnóstico. |
| --- |

**PLAN DE TRABAJO**

**Fecha de Inicio prevista: 24/10/2022**

**Fecha de Finalización prevista: 31/07/2023**

Se prevé una carga horaria de 10 horas semanales por integrante, distribuidas en 5 días. Se tiene en cuenta que cada semana estas horas serán distribuidas a criterio de cada integrante.

| **Tareas Previstas**  *Etapa 1:*   * ***Análisis e Investigación de herramientas y tecnologías a utilizar:*** Esta tarea implicará un análisis y familiarización de las tecnologías, lenguajes de programación y herramientas que serán utilizadas a lo largo del proyecto. Principalmente, esto involucrará al lenguaje de programación Python, la librería SpaCy, librerías de visualización de datos, la herramienta de etiquetado Label Studio, el framework Django, entre otros. * ***Relevamiento de requerimientos:*** Esta tarea consistirá en realizar un relevamiento de las necesidades de los usuarios, a partir de una o varias reuniones con el referente funcional, para la posterior creación del documento de especificación de requisitos de software (SRS). * ***Diseño de arquitectura****:* Esta tarea implicará el diseño de las interfaces web, de la estructura del servidor web y dónde estará alojado. Además, incluirá un diseño de las API de los modelos de análisis y de las estructuras de datos que estas devolverán.   *Etapa 2:*   * ***Recopilación y etiquetado de datos para el modelo de detección de entidades:*** Esta tarea implicará la creación de un dataset con datos etiquetados para poder llevar a cabo el entrenamiento y la prueba de un modelo de NLP capaz de detectar entidades en un texto. * ***Implementación de modelo de detección de entidades****:* Se utilizará la librería SpaCy de Python como herramienta para crear el modelo que será entrenado con un subconjunto del dataset previo. Este modelo recibirá un conjunto de textos y devolverá las entidades encontradas en los mismos. * ***Testeo y Entrenamiento de modelo de entidades****:* Se entrenará el modelo y se compararán los resultados obtenidos con los esperados. Si es necesario se reentrenará el modelo hasta llegar a un nivel de precisión aceptable. * ***Validación del modelo***: Se validará el modelo avanzado con el referente funcional, y se harán las correcciones finales de acuerdo a las observaciones recibidas.   *Etapa 3:*   * ***Recopilación* y etiquetado de datos para el modelo de detección de violencia de género**: Esta tarea implicará la creación de un dataset con datos etiquetados para poder llevar a cabo el entrenamiento y la prueba de un modelo de NLP con el objetivo de detectar este tipo de violencia. Se buscará clasificar partes de un texto como violentos, indicando el tipo de violencia, o no violentos. * ***Implementación de modelo de detección de violencia de género:*** Implementación, usando SpaCy, de un modelo capaz de detectar comportamientos violentos en un texto. Encontrar y clasificar partes de un texto como violentos o no violentos. Se buscará que luego del entrenamiento muestre como resultado un subconjunto de interés que se marcó con mayor posibilidad de contener evidencia relevante de una situación de violencia de género. * ***Testeo y entrenamiento de modelo de detección de violencia de género:*** Se entrenará el modelo y se compararán los resultados obtenidos con los esperados. Si es necesario se lo reentrenará hasta llegar a un nivel de precisión aceptable. * ***Validación del modelo***: Se validará el modelo avanzado con el referente funcional, y se harán las correcciones finales de acuerdo a las observaciones recibidas.   *Etapa 4:*   * ***Implementación de servidor backend:*** Esta tarea implicará la creación del servidor en Django e implementación de toda la estructura de controladores, servicios y modelos que interactuarán con el módulo de SpaCy. El objetivo es ofrecer una serie de API que pongan a disposición los datos que después serán formateados por la interfaz web. * ***Implementación de interfaz web; Carga de datos y uso de modelos:*** Esta tarea implicará crear el maquetado de las interfaces y estilos de las páginas web que servirán como interfaz para los usuarios. Además, se diseñarán las funcionalidades para cargar y llamar al servidor de backend para el procesado de datos. * ***Implementación de interfaz web; Visualización de resultados:*** Esta tarea implicará crear el maquetado de las interfaces, estilos de las páginas web que visualizarán los resultados. Procesarán los datos devueltos por el servidor backend y mostrarán los resultados mediante gráficos, tablas y otras herramientas de visualización.     Debido a la retroalimentación entre tareas dada en cada etapa, se indica en el Gantt qué tarea tiene el foco en cada periodo de tiempo, destacando la misma con un color más intenso. |
| --- |

**FODA DEL PROYECTO**

| | **Fortalezas**   * Poca inversión necesaria en recursos, herramientas y materiales * Desarrollo propio y con herramientas de código abierto. * Experiencia previa del equipo con Python y SpaCy. * Contacto con actores del dominio, para la provisión de datos y asistencia técnica y social. | **Debilidades**   * El equipo no dispone de experiencia en el desarrollo de aplicaciones que integren inteligencia artificial. * El equipo tiene un número de personas reducido y debe etiquetar una gran cantidad de datos para poder entrenar los modelos. | | --- | --- | | **Oportunidades**   * Existencia de necesidad concreta de análisis de texto automático en el proceso legal. * Expandir la capacidad del producto para analizar otros tipos de texto o incorporar nuevos módulos de análisis. * Solución con impacto social. | **Amenazas**   * Cambios en las tecnologías que impacten en el proyecto. | |
| --- | --- | --- | --- | --- |