Asignatura	Datos del alumno	Fecha
Investigación y gestión de proyectos en inteligencia artificial	Apellidos: Ruiz Vallecillo	19/01/2025
	Nombre: Araceli	

Actividad: Propuesta de plataforma de despliegue de un proyecto de inteligencia artificial basado en el desarrollo experimental de un clasificador de mensajes de odio

1. Análisis del problema e identificación del alcance

La empresa ***SureTech Innovations*** se enfrenta a un problema relacionado con el *aumento de mensajes de odio en redes sociales*, este no solo afecta a la dinámica de sus empleados, sino que también compromete *la diversidad, la inclusión y la seguridad en el entorno laboral*. Este contexto ha llevado a la empresa a buscar una solución tecnológica que permita detectar y mitigar estos mensajes, *protegiendo a sus usuarios internos y cumpliendo con las normativas legales aplicables* al discurso de odio.

El objetivo principal es desarrollar un sistema automatizado que, mediante técnicas avanzadas de inteligencia artificial, sea capaz de identificar y clasificar mensajes de odio. Este clasificador debe, además, *prevenir el ciberacoso* identificando patrones comunes en mensajes ofensivos y contribuir a mantener un *entorno inclusivo*. Para asegurar su efectividad, el sistema debe lograr una alta precisión en la clasificación, como, por ejemplo, un *F1-score superior al 90%,* y ser *escalable* para integrarse fácilmente en las plataformas internas de SureTech.

La base del proyecto será el uso del ***dataset Hatemedia***, que proporciona una colección exhaustiva de lemas simples y compuestos en castellano, clasificados como potencialmente odiosos. Este recurso permitirá entrenar y validar un modelo robusto de procesamiento del lenguaje natural (NLP), optimizado para identificar mensajes que fomenten el acoso, la exclusión o la discriminación hacia individuos o grupos. Se ***define un mensaje de odio*** como aquel que contiene lenguaje que degrada, amenaza o discrimina en función de características como la raza, género, orientación sexual, religión, entre otros.

El alcance de este proyecto incluye la creación de un sistema de clasificación accesible a través de una API, que podrá integrarse en las plataformas digitales de SureTech para su uso interno. Este enfoque permitirá detectar, analizar y clasificar mensajes de forma continua y eficiente.

2. Propuesta de plataforma de despliegue para la solución tecnológica

La solución planteada para este problema se basa en un ***sistema de clasificación*** de texto diseñado para identificar mensajes de odio en redes sociales mediante técnicas avanzadas de procesamiento del lenguaje natural (NLP) y aprendizaje automático. Para

ello, se desarrollará un modelo basado en arquitecturas modernas como *Transformers (BERT),* entrenado con el *dataset Hatemedia*. Este enfoque permite capturar el contexto semántico y lingüístico de los mensajes en castellano, maximizando la precisión en la detección de contenido ofensivo o discriminatorio.

El sistema estará respaldado por una ***infraestructura tecnológica en la nube***, utilizando plataformas como *AWS SageMaker* o *Google Cloud AI,* aprovechando su hardware especializado como *GPU o TPU*, para procesar el dataset Hatemedia. Estas tecnologías son ideales para el entrenamiento y despliegue del modelo, sobre todo para arquitecturas basadas en *Transformers (BERT)*.

La ***interacción con el modelo*** será gestionada a través de una *API RESTful*, accesible desde las plataformas internas de SureTech, garantizando facilidad de integración y actualización.

Adicionalmente, se integrará un ***sistema de procesamiento en tiempo real*** basado en tecnologías como *Apache Kafka*, que posibilitará el análisis continuo de mensajes desde las plataformas internas de SureTech, permitiendo que los datos sean procesados y clasificados con mínima latencia. Este enfoque modular y escalable asegura que el sistema pueda manejar incrementos en la demanda sin comprometer su rendimiento.

El ***almacenamiento*** a largo plazo de los datos, incluyendo logs operativos y copias del dataset preprocesado, se llevará a cabo en *Amazon S3*. Este servicio proporciona escalabilidad y fiabilidad, lo que permite gestionar de manera efectiva grandes volúmenes de datos históricos que podrían ser necesarios para reentrenar el modelo o realizar análisis adicionales.

El desarrollo se estima en un plazo de *seis meses*, tiempo en el que se completará el entrenamiento, validación y despliegue inicial del modelo. Durante las iteraciones posteriores se realizarán ajustes y mejoras continuas para garantizar la máxima efectividad y adaptabilidad de la solución.

3. Metodología de desarrollo y ciclo de vida del proyecto

El desarrollo del sistema de detección de mensajes de odio seguirá una ***metodología ágil mixta*** basada en ***Scrum y Kanban***. Este enfoque permite una división del trabajo en *sprints* cortos y manejables de duración de *2 semanas*. Dentro de cada sprint, se mantendrá una reunión semanal de seguimiento en la que se discutirán el actual estado de cada tarea y los posibles problemas que, al definir la tarea, no se tuvieron en cuenta. Esta metodología garantiza una estructura iterativa y colaborativa, donde, a través de las reuniones de seguimiento se pueden detectar problemas de manera temprana y ajustar las tareas según las necesidades del proyecto. Además, este enfoque asegura que el sistema final sea preciso, escalable y adaptable, alineado con los objetivos de SureTech Innovations.

Asignatura	Datos del alumno	Fecha
Investigación y gestión de proyectos en inteligencia artificial	Apellidos: Ruiz Vallecillo	19/01/2025
	Nombre: Araceli	

El desarrollo del sistema de detección de mensajes de odio seguirá un **ciclo de vida** estructurado en seis fases principales:

Fase de Planificación

En esta fase se definirán los objetivos del proyecto, los requisitos funcionales y las métricas clave de evaluación, como *precisión, recall y F1-score*. También se identificarán las herramientas necesarias, incluyendo *AWS SageMaker o Google Cloud AI* para la infraestructura, y el *dataset Hatemedia* como base para el entrenamiento del modelo. El cronograma del proyecto se dividirá en *sprints de dos semanas* para asegurar una gestión ágil y flexible.

Fase de Adquisición de Datos

Se procesará el dataset Hatemedia, limpiando datos duplicados y mensajes irrelevantes. Además, se aplicarán *técnicas de preprocesamiento como tokenización, normalización y eliminación de stop-words*. Se analizará el desequilibrio de clases y, si es necesario, se emplearán técnicas de imputación y normalización para poder evitar sesgos en los datos.

Fase de Modelado y Evaluación

En esta fase se utilizarán modelos basados en arquitecturas como *Transformers (BERT)*, entrenadas para identificar patrones lingüísticos en el texto. Durante esta fase, se *ajustarán los hiperparámetros* y se evaluará el modelo utilizando *métricas como el F1-score*. Se seleccionará el modelo más eficaz y se validará con conjuntos de datos no utilizados en el entrenamiento *(fase de test)*.

Fase de Implementación

El modelo seleccionado en la fase anterior se integrará en una *API RESTful* para hacerlo accesible desde las plataformas internas de SureTech. Adicionalmente, se configurará un sistema de procesamiento en tiempo real basado en *Apache Kafka* para garantizar que el análisis de mensajes sea continuo y eficiente. También utilizará la herramienta de almacenamiento *Amazon S3*, que permite gestionar grandes cantidades de volumen de datos, que se podría necesitar para reentrenar el modelo o realizar análisis adicionales. Se realizarán pruebas exhaustivas para garantizar que la interacción entre los componentes se realiza de forma correcta.

Fase de Monitorización y Mantenimiento

El sistema será monitorizado constantemente para asegurar su rendimiento continuo y adaptabilidad. Se analizarán los falsos positivos y negativos para reentrenar el modelo con nuevos datos, permitiendo que evolucione junto con los patrones cambiantes de mensajes de odio en redes sociales.

Fase de Despliegue

El modelo será alojado en plataformas cloud como *AWS o Google Cloud AI*, aprovechando su capacidad de escalabilidad para manejar grandes volúmenes de datos. Se realizarán comprobaciones adicionales para confirmar que el sistema cumpla con los objetivos planteados en la fase de planificación.

4. Gestión de recursos materiales y humanos

El éxito del proyecto de detección de mensajes de odio en redes sociales depende de una buena planificación y asignación de recursos materiales y humanos. Para este proyecto, se han seleccionado de manera cuidadosa los recursos necesarios para garantizar que el sistema *sea preciso, eficiente y escalable*, cumpliendo con los objetivos de SureTech Innovations.

En cuanto a los **recursos materiales**, se empleará una *infraestructura tecnológica* basada en la nube, utilizando plataformas como *AWS SageMaker o Google Cloud AI*, que ofrecen servicios especializados para el entrenamiento, evaluación y despliegue de modelos de machine learning. Estas plataformas permiten escalar los recursos según las necesidades del proyecto y, además, ofrecen acceso a hardware especializado, como *GPU y TPU*, indispensables para procesar grandes volúmenes de datos textuales. A su vez, se emplearán *bibliotecas de aprendizaje automático* como *TensorFlow o PyTorch* y *herramientas de procesamiento de lenguaje natural* (NLP), como *SpaCy o NLTK,* para el diseño y entrenamiento del modelo. El *almacenamiento y gestión de datos* se realizará en bases de datos escalables en la nube, como *Amazon S3*, que faciliten la organización y consulta de manera eficiente del dataset Hatemedia. Además, también se utilizará un sistema de procesamiento en tiempo real basado en *Apache Kafka.*

Respecto a los **recursos humanos**, el proyecto requerirá un equipo multidisciplinario con experiencia en áreas clave de la inteligencia artificial y el procesamiento de datos. Este equipo se compondrá por los siguientes perfiles:

- **Jefe de proyecto**: Coordinará los esfuerzos del equipo, supervisará la ejecución de las tareas y garantizará el cumplimiento de los plazos establecidos en cada sprint.
- **Ingeniero de datos**: Encargado de la preparación y gestión del dataset Hatemedia, asegurando su calidad y disponibilidad durante todas las etapas del proyecto.
- **Científico de datos**: Responsable del análisis exploratorio de los datos, la limpieza y la selección de características relevantes, así como del diseño y entrenamiento del modelo de clasificación.

- **Especialista en NLP**: Encargado de aplicar técnicas avanzadas de procesamiento del lenguaje natural, como tokenización y lematización, para optimizar el modelo en el análisis semántico y lingüístico en la lengua del castellano.
- **Ingeniero de aprendizaje automático**: Supervisará la implementación del modelo en la API y trabajará en su optimización para maximizar métricas como precisión, recall y F1-score.
- **Arquitecto de sistemas**: Diseñará la infraestructura tecnológica necesaria para el despliegue del modelo, garantizando la escalabilidad y robustez de la solución en la nube.
- **Experto en ética y regulación IA**: Se encargará de identificar y evaluar los riesgos éticos y legales posibles, así como desarrollar estrategias de mitigación para que la solución sea transparente, responsable y no discriminatoria, garantizando así que el proyecto IA cumpla con las regulaciones aplicables y siga los principios éticos.
- **Especialista en ciberseguridad y privacidad de datos**: Se encargará de garantizar la protección de los datos mediante técnicas de anonimización y cumplimiento normativo, de esta forma se aseguraría la seguridad del sistema y su alineación con las diferentes regulaciones legales, como el *GDPR*.

Este enfoque, asegura que cada aspecto del proyecto sea abordado por profesionales con experiencia en su área, reduciendo así los riesgos y maximizando la eficiencia. La combinación de los recursos materiales escogidos y un equipo humano capacitado permitirá que el sistema final cumpla con las expectativas de SureTech Innovations y sea una herramienta efectiva para la detección de mensajes de odio.

5. Entorno de seguridad

La seguridad y privacidad de los datos son aspectos fundamentales en el desarrollo y despliegue del sistema de detección de mensajes de odio. Dado que el proyecto utiliza datos sensibles provenientes del dataset Hatemedia, es crucial implementar *medidas estrictas de securización y anonimización* que garanticen el cumplimiento de normativas como el *Reglamento General de Protección de Datos (GDPR)* y otras regulaciones aplicables.

Para **proteger los datos** utilizados en este proyecto, se aplicarán *técnicas de anonimización y pseudonimización*. Esto incluirá:

- La eliminación de identificadores personales, como nombres o direcciones IP, así como, la sustitución de estos por identificadores genéricos o encriptados que permitan analizar el contenido sin comprometer la identidad de los individuos involucrados.

- Los datos serán cifrados tanto en reposo como en tránsito utilizando protocolos avanzados de seguridad, como AES-256 en el caso del almacenamiento y TLS para las comunicaciones. De esta forma, se asegura que no puedan ser interceptados ni manipulados.
- Se implementará un proceso de revisión continua a cargo del responsable de ciberseguridad y privacidad de datos para garantizar que los datos recolectados cumplan con los requisitos legales y éticos.

En el caso de la **infraestructura**, para implementar la ***seguridad en la red***, se utilizarán configuraciones avanzadas que incluyan *firewalls* y *sistemas de prevención de intrusiones (IPS)*. Los servidores en los que se alojará el sistema serán actualizados regularmente para garantizar que cuenten con los parches de seguridad más recientes.

Para la **seguridad en la nube y en entornos híbridos**, el modelo y los datos serán alojados en *plataformas cloud como AWS o Google Cloud AI*, que cumplen con altos estándares de seguridad y ofrecen certificaciones como *ISO 27001 y SOC 2*. Estas plataformas proporcionan servicios de encriptación, auditorías automatizadas y herramientas para gestionar incidentes. Además, dado que el proyecto podría operar en un *entorno híbrido*, se implementarán túneles VPN para conectar de forma segura las operaciones locales y en la nube, evitando la exposición de la información en redes públicas.

Para garantizar **la gestión de identidad y el control de acceso**, se implementarán controles de identidad robustos mediante *autenticación multifactor (MFA)* para que solo personal autorizado acceda al sistema. El *acceso estará basado en roles*, de modo que los privilegios sean otorgados según las responsabilidades de cada miembro del equipo. Este enfoque minimiza los riesgos al limitar el alcance de las acciones de cada usuario, evitando accesos indebidos o modificaciones no autorizadas.

Por último, en cuanto a la **monitorización y respuestas ante incidentes de seguridad**, se configurarán sistemas de monitorización continua para detectar posibles anomalías en tiempo real. Herramientas como *AWS CloudTrail o Google Cloud Operations Suite* permitirán registrar y auditar todas las actividades relacionadas con el sistema, facilitando la identificación de comportamientos inusuales. Además, se diseñará un plan de respuesta ante incidentes que contemple la detección, análisis, contención y resolución de problemas, asegurando que el impacto de cualquier incidente sea mínimo y que las operaciones puedan restablecerse rápidamente.

Las medidas de seguridad adoptadas aseguran que el sistema cumpla con los estándares más altos de protección de datos, garantizando tanto la privacidad de los individuos como la integridad y fiabilidad del sistema desarrollado por SureTech Innovations.

Asignatura	Datos del alumno	Fecha
Investigación y gestión de proyectos en inteligencia artificial	Apellidos: Ruiz Vallecillo	19/01/2025
	Nombre: Araceli	

6. Planificación y estimación de costes

En cuanto a la **planificación del proyecto**, este se desarrollará en un *período de seis meses*, dividido en *sprints de dos semanas*, cada uno con entregables específicos y una serie de metas. Durante este tiempo, se completarán las fases del ciclo de vida del proyecto:

1. **Fase de planificación (0-1 mes):** Definición de objetivos, requisitos funcionales, herramientas necesarias y del cronograma.
2. **Fase de adquisición de datos (1-2 meses):** Recopilación, limpieza y preprocesamiento del dataset Hatemedia, incluyendo la identificación y resolución de desequilibrios o sesgos en las clases.
3. **Fase de modelado y evaluación (2-4 meses):** Diseño, entrenamiento y validación del modelo de clasificación.
4. **Fase de implementación y pruebas (4-5 meses):** Integración del modelo en una API RESTful y configuración del sistema de procesamiento en tiempo real.
5. **Fase de despliegue y monitoreo inicial (5-6 meses):** Implementación en la nube y monitorización de métricas operativas para ajustar el sistema.

Cada fase estará gestionada mediante reuniones regulares de seguimiento y revisiones al final de cada sprint, asegurando que las tareas se completen a tiempo y dentro del presupuesto.

Para poder realizar una **estimación del coste del proyecto**, será necesario tener en cuenta los recursos necesarios para llevar a cabo el mismo:

- **Infraestructura tecnológica:**
  - *Servicios en la nube:* AWS SageMaker o Google Cloud AI para el entrenamiento y despliegue del modelo.
  - *Hardware especializado:* Acceso a GPU y TPU para procesar grandes volúmenes de datos de manera eficiente.
  - *Herramientas de software:* Bibliotecas de aprendizaje automático (TensorFlow, PyTorch) y procesamiento de lenguaje natural (SpaCy, NLTK).
  - *Almacenamiento:* Bases de datos escalables como Amazon S3 para gestionar el dataset Hatemedia.
  - *Procesamiento en tiempo real*: Apache Kafka para gestionar el flujo continuo de mensajes y garantizar que el sistema pueda analizar y clasificar datos de redes sociales con mínima latencia.

- **Equipo humano:**
  - *Jefe de proyecto:* para coordinar el equipo y gestionar el progreso de las tareas.
  - *Ingeniero de datos:* para la preparación del dataset y su integración con las herramientas de procesamiento.

- *Científico de datos y especialista en NLP:* para el diseño y entrenamiento del modelo.
- *Ingeniero de aprendizaje automático:* para implementar y optimizar el modelo.
- *Arquitecto de sistemas:* para diseñar la infraestructura tecnológica y supervisar su escalabilidad.
- *Experto en ética y regulación IA:* para garantizar que se cumplan las regulaciones y principios éticos.
- *Especialista en ciberseguridad y privacidad de datos:* para garantizar la privacidad de los datos.

Una vez determinados los recursos y el personal necesario para llevar a cabo el proyecto, se puede realizar una estimación del coste, los cuales se dividirán en tres grupos:

Costes iniciales

- Licencias y suscripciones para plataformas cloud: 4,000 € aproximadamente.
- Configuración inicial del entorno y recursos computacionales: 2,000 €.
- Adquisición y limpieza del dataset Hatemedia: 2,000 €.

Costes operativos

- Uso mensual de servicios en la nube (GPU, almacenamiento): 2,000 €/mes.
- Mantenimiento del sistema y monitorización: 1,000 €/mes.

Costes de personal

- Equipo de desarrollo y especialistas (6 meses): 50,000 €, considerando un equipo multidisciplinario con perfiles de alto nivel.

Para mantener los costes bajo control, se utilizará la escalabilidad de los servicios cloud, ajustando los recursos según las necesidades de cada fase. Además, el uso de herramientas de código abierto reducirá los gastos en licencias de software.