



Becas colaboración curso 2023/2024

Fecha: 29 Mayo 2023

Vicerrectorado de Investigación

Subcomisión de I+D+i

Propuesta del departamento *MATEMÁTICA APLICADA*

Núm Proyecto: 2023/26/00004

Responsable

Conejero Casares, José Alberto

E-mail

aconejero@upv.es

Ext.

19664

Título proyecto

Uso de grandes modelos de lenguaje para detectar información dañina

Valoración proyecto

3

Descripción proyecto

El proyecto tiene como objetivo principal utilizar grandes modelos de lenguaje para detectar información dañina en diversos contextos, como redes sociales, sitios web y plataformas de mensajería. La detección de información dañina es de vital importancia para salvaguardar la seguridad, el bienestar y la integridad de los usuarios en línea. Al aprovechar los avances en inteligencia artificial y aprendizaje automático, este proyecto busca desarrollar un sistema eficiente y preciso que pueda identificar contenido perjudicial, incluyendo discursos de odio, acoso, desinformación y contenido violento.

Objetivos:

Entrenar grandes modelos de lenguaje: Utilizando técnicas de aprendizaje profundo y redes neuronales, se entrenarán grandes modelos de lenguaje en los conjuntos de datos recopilados y etiquetados. Estos modelos aprenderán a reconocer patrones y características de contenido dañino.

Validación y ajuste de los modelos: Se realizarán pruebas exhaustivas para validar la efectividad de los modelos entrenados en la detección de información dañina. Durante esta etapa, se ajustarán los modelos según sea necesario para mejorar su precisión y capacidad de detección.

Desarrollo de una API o una aplicación: Se creará una interfaz de programación de aplicaciones (API) o una aplicación que permita a los usuarios acceder y utilizar el sistema de detección de información dañina. Esto facilitará la integración del sistema en diferentes plataformas y entornos en línea.

Evaluación y mejora continua: Se llevarán a cabo evaluaciones periódicas para medir la eficacia del sistema de detección y se realizarán mejoras continuas para mantenerse al día con las nuevas formas de información dañina que puedan surgir. Esto implica la actualización regular de los conjuntos de datos de entrenamiento y la adaptación de los modelos para abordar nuevos desafíos.

Actividades a realizar por el alumno

- Entrenamiento de modelos de lenguaje: Utilizando bibliotecas y herramientas de aprendizaje automático, el alumno deberá entrenar grandes modelos de lenguaje en los conjuntos de datos recopilados. Esto implicará ajustar los hiperparámetros, seleccionar arquitecturas de modelos adecuadas y ejecutar el proceso de entrenamiento para que los modelos aprendan a detectar información dañina.
- Evaluación de la precisión del modelo: Una vez entrenados los modelos, el alumno deberá evaluar su precisión y rendimiento mediante pruebas exhaustivas. Esto implica utilizar conjuntos de datos de prueba y métricas de evaluación adecuadas para medir la capacidad de detección de información dañina de los modelos.
- Ajuste y mejora de los modelos: En base a los resultados de la evaluación, el alumno deberá realizar ajustes en los modelos para mejorar su precisión y capacidad de detección. Esto puede incluir la exploración de



Becas colaboración curso 2023/2024

Fecha: 29 Mayo 2023

diferentes arquitecturas de modelos, el uso de técnicas de transferencia de aprendizaje o el ajuste de los hiperparámetros del modelo.

Localización de la actividad (Campus)

Vera

Horario

(A convenir).