PROFUNDIZACION CLASE 6

Recursos de profundización
Ética y Sesgos en el PLN

Introducción
El Procesamiento de Lenguaje Natural (PLN) ha transformado la interacción entre humanos y máquinas, permitiendo a los sistemas comprender, generar y manipular el lenguaje humano. Aplicaciones como asistentes virtuales, traductores automáticos y chatbots educativos son ejemplos tangibles de cómo el PLN está integrado en la vida cotidiana. Sin embargo, este avance tecnológico plantea cuestiones éticas cruciales, especialmente en relación con los sesgos que los modelos de PLN pueden heredar de los datos de entrenamiento. Estos sesgos pueden reforzar estereotipos, discriminar a grupos específicos y generar información engañosa. En esta sección se abordan las implicaciones éticas del PLN, el origen de los sesgos y las estrategias para mitigarlos.

Ética y Sesgos
Los sesgos en PLN surgen principalmente de los datos utilizados para el entrenamiento. Los modelos como GPT-4 o BERT se entrenan con grandes volúmenes de texto recopilado de internet, libros y redes sociales, que reflejan los prejuicios culturales, sociales y de género presentes en la sociedad. Por ejemplo, si un modelo se entrena mayoritariamente con textos que asocian a las mujeres con tareas domésticas y a los hombres con profesiones técnicas, es probable que reproduzca estos estereotipos en sus respuestas.

Las implicaciones éticas son significativas. En el ámbito educativo, un modelo sesgado podría influir negativamente en la percepción de los roles de género o reforzar desigualdades culturales. En sistemas de reclutamiento automático, los sesgos podrían conducir a la discriminación, favoreciendo a ciertos perfiles sin fundamentos objetivos. Estos escenarios subrayan la necesidad de abordar los sesgos desde una perspectiva ética, garantizando la equidad y la inclusión en el uso del PLN.

Existen diversas estrategias para mitigar los sesgos en PLN. Una de ellas es la curación de los datos de entrenamiento, eliminando o equilibrando la información que pueda inducir prejuicios. Otra es el ajuste fino (*fine-tuning*) del modelo, que consiste en reentrenarlo con conjuntos de datos más diversificados y balanceados. Además, el uso de técnicas de interpretabilidad permite a los desarrolladores analizar cómo los modelos toman decisiones, facilitando la identificación de posibles sesgos.

La regulación también juega un papel fundamental. Instituciones internacionales están desarrollando marcos éticos para la inteligencia artificial que incluyen directrices

específicas sobre PLN. Estas regulaciones buscan garantizar la transparencia, la responsabilidad y la protección de datos, promoviendo un desarrollo tecnológico que respete los derechos humanos.

- El PLN ofrece oportunidades sin precedentes para mejorar la comunicación y el acceso al conocimiento. Sin embargo, los sesgos inherentes a estos modelos plantean desafíos éticos que no pueden ser ignorados. Es esencial implementar prácticas de desarrollo responsables que incluyan la curación de datos, el ajuste fino de modelos y la adopción de marcos regulatorios éticos. Solo así se podrá garantizar que las aplicaciones de PLN sean equitativas, inclusivas y beneficiosas para toda la sociedad.