Por: **Wilson Enrique Ramírez, Wilmar Martínez Luna, Edwin Peñarredonda Novoa**

Propuesta para la creación del Modelo de Limpieza de Datos en la plataforma Hugging fase.

Introducción

La limpieza de datos es un proceso crucial en cualquier pipeline de ciencia de datos, ya que asegura que los datos utilizados para el entrenamiento y evaluación de los modelos sean consistentes, libres de errores y de alta calidad. El presente proyecto tiene como objetivo desarrollar un modelo en la plataforma Hugging Face que facilite la limpieza automática de datasets de texto, ayudando a mejorar la calidad de los datos en múltiples aplicaciones de procesamiento de lenguaje natural (NLP).

2. Objetivo General

Desarrollar un modelo de limpieza de datos que se enfoque en mejorar la calidad de los datasets textuales, eliminando errores, duplicados, incoherencias y otras imperfecciones comunes en los datos crudos.

3. Objetivos Específicos

Implementar funciones de preprocesamiento que aborden las tareas de eliminación de duplicados, corrección de errores ortográficos, normalización de texto y eliminación de ruido (caracteres no deseados, espacios innecesarios, etc.).

Diseñar un modelo que sea capaz de identificar y corregir inconsistencias en el etiquetado de datos.

Desarrollar una API que permita la integración del modelo con otros proyectos de NLP en Hugging Face.

Realizar pruebas exhaustivas con diferentes datasets de NLP disponibles en Hugging Face para asegurar la efectividad del modelo.

4. Justificación

El volumen creciente de datos no estructurados disponibles en la web y otras fuentes hace indispensable contar con herramientas eficientes de limpieza de datos. La disponibilidad de un modelo de limpieza automatizado no solo ahorrará tiempo en el preprocesamiento, sino que

también mejorará la calidad de los modelos de machine learning y Deep learning que utilizan estos datos.

5. Metodología

Análisis del problema: Identificar las principales imperfecciones que suelen encontrarse en los datasets textuales (errores ortográficos, duplicados, formato inconsistente).

Recolección de datos: Se utilizarán datasets disponibles en la plataforma Hugging Face para pruebas y desarrollo del modelo.

Diseño del modelo: Se desarrollará un pipeline de procesamiento de datos utilizando librerías como transformes y datasets, que incluirá módulos para:

Tokenización y normalización de texto.

Eliminación de duplicados.

Corrección de errores ortográficos (usando herramientas como SymSpell o Hunspell).

Identificación y limpieza de texto irrelevante o ruidoso (como enlaces web o símbolos no deseados).

Implementación: Se utilizará PyTorch o TensorFlow como backend del modelo, y se diseñará un API amigable para la comunidad.

Pruebas y evaluación: Validación del modelo utilizando diferentes datasets y métricas como precisión, recall, F1-score para medir la efectividad del proceso de limpieza.

6. Tecnologías a Utilizar

Plataforma Hugging Face (Transformers, datasets).

Python (bibliotecas: pandas, nltk, spacy).

PyTorch/TensorFlow para el desarrollo del modelo.

Herramientas de corrección ortográfica (SymSpell, Hunspell).

Integración con Hugging Face Hub para compartir el modelo con la comunidad.

7. Resultados Esperados

- Un modelo robusto de limpieza de datos disponible en Hugging Face.
- Documentación detallada que permita a otros desarrolladores integrar fácilmente el modelo en sus proyectos de NLP.
- Mejora en la calidad de los datasets utilizados por otros modelos de la comunidad.