Taller de Tesis I – Entrega I

Año 2025 Data Mining - UBA

Alumno: Santiago Tedoldi

Contexto y Objetivo

El tema elegido para cumplir con las consignas del Taller de Tesis I se relaciona con la clasificación arancelaria/tarifario de mercaderías, partiendo de descripciones comerciales y según la nomenclatura del Sistema Armonizado (SA) internacional. La finalidad es evaluar las características de las descripciones comerciales que logran una clasificación más acertada, considerando los desafíos propios de un dataset variado y desbalanceado.

Descripción del Conjunto de Datos

Contenido y Origen:

El dataset está formado por tuplas con información de mercaderías sometidas al comercio internacional (descripción_comercial, código_tarifario) en idioma inglés, en un total de 500 mil ítems.

El origen de estos datos no es preciso. Fue material de un curso suministrado en una capacitación internacional en Corea del Sur, en el marco del Programa BACUDA de la Organización Mundial de Aduanas.

Características:

- **Muestra Variada:** El conjunto de datos presenta una diversidad considerable en cuanto a tipos de productos y mercaderías.
- **Desbalanceo:** Predominan los productos relacionados con máquinas y material eléctrico/electrónico, como también vehículos automotores.

Preguntas y Retos del Proyecto

Con estos datos y la problemática planteada, surgen preguntas que podrías abordarse en el siguiente orden:

1. **Cantidad de Muestras Necesarias:** Cuantas muestras son necesarias, teniendo en cuenta los miles de códigos de la nomenclatura internacional.
2. **Calidad de las Descripciones Comerciales:** Cómo se puede evaluar la calidad de las descripciones comerciales, para que sean bien clasificadas por modelos de DeepLearning.

Técnicas y Enfoques Propuestos

Para enfrentar los desafíos planteados, las potenciales técnicas identificadas son:

- **Encoders de texto:** Se propone el uso de redes neuronales con arquitectura transformer que ya han sido pre-entrenadas con grandes volúmenes de texto.
- **PCA y Clústering:** Cómo método analítico y exploratorioPara reducir la dimensionalidad del conjunto de datos, facilitar el manejo de las características extraídas e identificar patrones de los textos en función de similitudes inherentes sin necesidad de etiquetas predefinidas.
- **Clasificación:** Se contempla la utilización de modelos basados en árboles y redes neuronales para la clasificación de texto, funcionando como “cabezas” que procesan los embeddings producidos por los encoders.