Resumen de Pipeline: Reviews de Pueblos Mágicos a Predicciones

Este pipeline está inspirado en el paper "REST-MEX: Sentiment Analysis of Mexican Restaurants" (https://ceur-ws.org/Vol-3496/restmex-paper14.pdf).

1. Descripción de alto nivel

Este código implementa un pipeline de análisis de sentimiento para reseñas de Pueblos Mágicos. Parte de la carga y preprocesamiento de datos, realiza un balanceo de clases mediante back■translation, adapta un modelo de lenguaje al dominio específico, afina el modelo con validación cruzada y genera predicciones finales para el conjunto de prueba.

2. Pasos principales

Paso	Descripción
1. Instalación de dependencias	Pandas, Transformers, Torch, Datasets, etc.
2. Carga de datos	Lectura de archivos Excel de entrenamiento y prueba.
3. Preprocesamiento	Limpieza de texto y normalización.
4. Balanceo de datos	Back■translation usando MarianMT.
5. Adaptación al dominio	Masked LM pretraining sobre el corpus de reseñas.
6. Fine■tuning con CV	Entrenamiento de clasificación con validación cruzada.
7. Entrenamiento final y predicciones	Entrena con todo el conjunto balanceado y guarda predicciones en .txt.

3. Modelos y Parámetros

3.1 Back■Translation (MarianMT)

Parámetro	Valor
Modelos	Helsinki-NLP/opus-mt-es-en, Helsinki-NLP/opus-mt-en-es
batch_size	128
max_length (es→en y en→es)	128

3.2 Adaptación al Dominio (Masked LM)

Parámetro	Valor
base_model	PlanTL-GOB-ES/roberta-base-bne
checkpoint_dir	domain_adapted_model
epochs	3
batch_size	32
learning_rate	2e-5
weight_decay	0.01
logging_steps	100
save_steps	500

3.3 Fine■tuning con Validación Cruzada

Parámetro	Valor
base_model	domain_adapted_model
num_labels	5
epochs	4
batch_size	8
learning_rate	2e-5
weight_decay	0.01
logging_steps	200
save_strategy	no

3.4 Entrenamiento Final y Predicción

Parámetro	Valor
output_dir	final_train
epochs	3
batch_size	8
save_strategy	no