Documentación Challenge 2: Análisis de comentarios Glassdoor.

Data: https://www.kaggle.com/datasets/davidgauthier/glassdoor-job-reviews

Repositorio: https://github.com/Jorgesr11/Programaci-n-II/tree/main/Challenge%202_JSR

Desarrollar un pipeline integral para analizar comentarios de empleados en Glassdoor, incluyendo:

- Clasificación automática de reseñas
- Análisis de sentimientos en inglés y español
- Implementación de prácticas MLOps para seguimiento y reproducibilidad

1. **Carga y Validación de Datos**
   - **Fuente**: Archivo local CSV (glassdoor_reviews.csv)
   - **Validaciones**:
     - Existencia del archivo en la ruta especificada
     - Presencia de columnas requeridas: review, pros, cons, language, rating
     - Filtrado de idiomas: solo inglés (en) y español (es)
2. **Preprocesamiento de Texto**
   - **Proceso Bilingüe**:
     - Limpieza de caracteres especiales y normalización a minúsculas
     - Tokenización específica por idioma
     - Eliminación de stopwords
     - Lematización (inglés) y stemming (español)
   - **Salida**: Texto procesado listo para modelado
3. **Modelado Predictivo**
   - **Arquitectura**:
     - Vectorización TF-IDF con n-gramas (1,2)
     - Modelo de Regresión Logística para clasificación de ratings
   - **Métricas**:
     - Precisión (accuracy)

    - Reporte de clasificación completo

4. **Análisis de Sentimientos**
    - **Herramientas**:
        - PySentimiento para textos en español
        - VADER para textos en inglés
    - **Clasificación**:
        - Positivo (pos), Neutral (neu), Negativo (neg)

5. **Pipeline MLOps**
    - **Tecnologías**:
        - MLflow para tracking de experimentos
        - Registro de:
            - Parámetros del modelo
            - Métricas de desempeño
            - Artefactos (gráficos, datos crudos)
    - **Visualización**:
        - Distribución de sentimientos (gráfico de barras)

Categoría	Herramientas
Procesamiento de Lenguaje	NLTK, PySentimiento, VADER
Machine Learning	Scikit-learn, Logistic Regression
MLOps	MLflow
Visualización	Matplotlib
Gestión de Datos	Pandas, NumPy