Predicción de riesgo de empeoramiento en pacientes con IBD usando NLP en Reddit y Twitter

Justificación:
Los pacientes comparten síntomas y señales tempranas de empeoramiento en foros y redes. Un modelo capaz de **reconocer lenguaje asociado a empeoramiento** puede servir como herramienta de vigilancia poblacional y generar hipótesis clínicas.

Pregunta de investigación
¿Se puede predecir, con precisión útil, el riesgo de empeoramiento o recaída en pacientes con IBD a partir del lenguaje que usan en Reddit y Twitter?

Objetivo general
Desarrollar un pipeline NLP que identifique publicaciones indicativas de empeoramiento/recaída en usuarios que hablan sobre IBD en Reddit y Twitter, y evaluar su desempeño y principales señales lingüísticas.

Objetivos específicos

1. Recolectar y construir un corpus de posts/tweets en inglés y/o español relacionados con IBD.
2. Anotar un dataset semilla con etiquetas: empeoramiento vs no-empeoramiento.
3. Entrenar modelos de clasificación (baseline: TF-IDF + logistic/RF; avanzado: fine-tuned BERT/BioBERT) para detectar lenguaje de empeoramiento.
4. Evaluar modelos (AUC, F1, precision/recall) y explicar predicciones (SHAP / LIME / tokens importantes).
5. Comparar diferencias CD vs UC (subreddit / keywords) y generar visualizaciones interpretables.

---

2) Metodología

Fase 0 — Definición de señales

- Definir qué frases indican *empeoramiento*: “flare”, “hospitalized”, “started steroids”, “stool with blood”, “worsening pain”, “couldn’t eat” etc.

- Crear lista de keywords iniciales para scrapping (Crowdsource + revisión bibliográfica).

Fase 1 — Recolección de datos

- **Reddit**: usar `praw` o `psaw` para extraer posts + comentarios de subreddits:
  - r/IBD, r/CrohnsDisease, r/UlcerativeColitis, r/ChronicIllness
- **Twitter/X**: usar `snscrape` para buscar tweets con keywords: “Crohn”, “colitis”, “flare”, “stool blood”, y los alimentos clave (provenientes de UKB)
- Guardar: id, fecha, author, subreddit/tweet, text, metadata (score, retweets).

Fase 2 — Preprocesamiento

- Normalización (lowercase etc), quitar URLs, emojis opcionalmente, expandir contracciones, limpieza de signos innecesarios.
- Tokenización, lematización con `spaCy` (modelo en inglés) o `stanza`.
- Detección explícita de idioma.

Fase 3 — Anotación

- Crear guía de anotación corta.
- Etiquetas mínimas: `flare` (1) vs `no-flare` (0) vs `ambiguous` (2).
- Cada post anotado por 2 personas; resolver desacuerdos por adjudicación.
- Tamaño objetivo mínimo para prototipo: **500–1000** ejemplos (más mejora performance).

Fase 4 — Modelado

Baselines

- TF-IDF + Logistic Regression (regularized)
- TF-IDF + Random Forest / XGBoost

Avanzado

- Fine-tune `distilbert-base-uncased` o `bert-base-uncased` (o `BioBERT` / `PubMedBert` vocabulario biomédico) para clasificación ternaria/binaria.
- Entrenamiento con `transformers` (HuggingFace) + `Trainer` (early stopping, class weights).

Entradas

- Texto completo, opcionalmente con features extra: contador de síntomas, presencia de keywords, metadatos (upvotes, length).

Fase 5 — Validación y evaluación

- Train/val/test split estratificado (por usuario?).

- Métricas: ROC-AUC, Precision-Recall AUC (útil si clases desbalanceadas), F1 (macro/weighted), precision@ si interesa detectar top-suspects.
- Curva de calibración si quieres probabilidades útiles.

Fase 6 — Interpretación

- SHAP sobre TF-IDF o LIME/Integrated Gradients para BERT: identificar tokens/patrones que más contribuyen a predicción de flare.
- Extraer frases representativas por clase.

Fase 7 — Visualización y reporte

- Word clouds, heatmaps de tokens, timeline de tweets/posts con flares, comparativa CD vs UC.
- Preparar notebook + README + small web demo (opcional: Streamlit).

---

3) Roles

- **A — Data Engineer - Scraper**
  - Implementa scrapers, almacena datos, mantiene CSV, anonimiza.
- **B — NLP Engineer - Annotation Lead**
  - Define esquema de anotación, coordina anotadores, preprocesa texto, entrena baselines.
- **C — Modeler - Visualizer**
  - Fine-tune BERT, evalúa y explica modelos, visuales y el informe final.

---

4) Entregables

- Dataset anonimizado (CSV) con texto, etiqueta, source, fecha.
- Scripts reproducibles: scraping, preprocess, train, evaluate.
- Notebook con análisis exploratorio y resultados.
- Modelo entrenado (weights) + explicación (SHAP plots).
- Paper
- README + documentación del proyecto.
- Presentación con hallazgos y limitaciones.

---

5) Ética y limitaciones

- **Privacidad**: solo usar contenido público. Anonimizar IDs; no publicar usernames.
- **Consentimiento**: Reddit/Twitter son públicos, pero trata los datos con respeto; considera quitar citas directas o acortarlas.
- **Bias / representatividad**: usuarios de redes no representan a todas las poblaciones - limitacion.
- **Validación clínica**: el modelo sugiere señales, no reemplaza diagnóstico; aclarar.

---

7) Métricas de éxito

- Baseline TF-IDF + LR: F1 ≥ 0.65 en test
- BERT fine-tune: mejora de +0.05–0.10 F1 sobre baseline
- Interpretabilidad: listado de tokens/frases relevantes y ejemplo de 10 posts correctamente identificados

---

8) Librerías

- Scraping: `praw`, `psaw`, `snscrape`
- Preprocesamiento: `spaCy`, `nltk`, `emoji`, `ftfy`
- Modelado: `scikit-learn`, `xgboost`, `transformers` (HuggingFace), `torch`
- Interpretación: `shap`, `lime`, `captum` (para pytorch)
- Visualización: `matplotlib`, `seaborn`, `plotly`, `wordcloud`
- Repositorio & reproducibilidad: `requirements.txt`, `Makefile` o `poetry`, notebooks