本阶段已完成两件事：**数据清洗与准备**——对 Kaggle IMDB 原始 CSV 统一列名、去缺失与去重，positive/negative 映射为 1/0，文本做 HTML/URL/邮箱/用户名替换、数字归一与小写化，并导出 imdb\_clean.csv 与用于快速调试的分层平衡子集 imdb\_balanced\_10k.csv；**文本→向量**——基于清洗后的 text,label 生成三套特征：词与字符 **TF‑IDF n‑grams**（imdb\_tfidf.npz 与词典 JSON）、**Word2Vec 平均句向量**（imdb\_w2v\_avg.npy 与 w2v.kv），以及 **ELECTRA 句向量**（imdb\_electra\_cls.npy/mean 与 electra\_meta.json），均为后续 LR/LSTM/ELECTRA 微调的输入，不涉及监督训练。