Dataset Description

1. Dataset & Task Deep-Dive

We are dealing with a Multi-Class Text Classification problem on a dataset of ~100k news articles.

The Data Schema

1. ***article (Text):*** The primary feature. Expect extreme variation in length (short breaking news vs. long editorials).
2. ***title (Text):*** High-density feature. In news, the title often contains the class label explicitly (e.g., "Apple acquires AI startup" -> Business/Tech).
3. ***Strategy:*** Concatenate title + " [SEP] " + article to force the model to see the title as part of the context.
4. ***source (Categorical)***: A strong predictor (e.g., "TechCrunch" is almost certainly Technology).
5. ***page_rank (Numerical):*** Likely correlates with "International" or "Business" (high authority sources) vs. "General News" (lower authority).
6. ***timestamp (Temporal):*** Critical Risk.
    1. News topics drift over time.
    2. Validation Strategy: Do not use a random train_test_split. Use a Time-Series Split (train on Jan-Mar, validate on April). This mimics the real-world scenario of predicting future news.

2. The Metric: Macro-F1

1. **Formula:** $\text{Macro-F1} = \frac{1}{N} \sum_{i=0}^{N} F1_i$
2. **Implication:** Minority classes matter just as much as majority classes. If "Health" has only 50 samples and "Sports" has 5,000, a 0% score on Health penalizes you heavily.
3. **Pitfall:** High accuracy (e.g., 90%) can hide a terrible Macro-F1 (e.g., 50%) if you ignore minority classes.