MultiLingua-100K

A Large-Scale Multilingual News Classification Dataset

Version 1.0 (July 2026)
Size 100,000 articles across 12 languages
License CC BY-SA 4.0
Authors Elena Marchetti, Kenji Tanaka, Sarah Okafor
Affiliation Global NLP Initiative | gnlpi.org
Contact datasets@gnlpi.org
DOI 10.5281/zenodo.1234567

Motivation & Purpose

Why This Dataset Exists.

Most multilingual text classification benchmarks are limited to 3–5 high-resource languages and rely on translated rather than natively-authored content. **MultiLingua-100K** fills this gap by providing 100,000 human-annotated news articles across **12 languages** spanning 6 language families, all sourced from original publications in each target language. The dataset is designed to support cross-lingual transfer learning, language-agnostic model evaluation, and fairness audits of NLP classifiers on under-represented languages.

Dataset Overview

Table 1: Summary statistics for MultiLingua-100K.

Property	Value
Total articles	100,000
Languages	12 (ar, de, en, es, fr, hi, ja, ko, pt, ru, tr, zh)
Categories	7 (Politics, Sports, Technology, Health, Business, Entertainment, Science)
Time span	January 2018 – December 2023
Avg. article length	480 words (SD = 210)
Unique sources	340+ news outlets
Annotation type	Multi-class single-label (expert + crowd-verified)
Inter-annotator κ	0.87 (Fleiss' κ, averaged across languages)
Train / Val / Test	70 K / 15 K / 15 K (stratified by language & category)

Category Distribution

Each category is represented with at least 8,000 articles. The smallest category (*Science*, 8,420 articles) and the largest (*Politics*, 22,100 articles) reflect the natural imbalance observed in global news coverage. Stratified sampling ensures fair representation across all train/val/test splits.

Collection Protocol & Timeline

Table 2: Phased collection timeline spanning 24 months.

Phase	Period	Activity
I	Jan–Mar 2022	Source identification: curated 340+ news outlets across 12 languages; obtained scraping permissions and terms-of-use waivers for 91% of sources.
II	Apr–Aug 2022	Automated scraping via Scrapy cluster; collected 1.2M raw HTML pages with publication timestamps and source metadata.
III	Sep–Dec 2022	Deduplication (MinHash LSH, threshold 0.85) reduced corpus to 340K unique articles. Language detection via FastText (confidence $>$ 0.9).
IV	Jan–Jun 2023	Expert annotation: 14 linguists (native speakers) labeled articles across 7 categories. Crowd-verification on AMT for 15% subsample.
V	Jul–Dec 2023	Quality audit, stratified rebalancing, final train/val/test splits. Released v1.0.

Preprocessing Pipeline

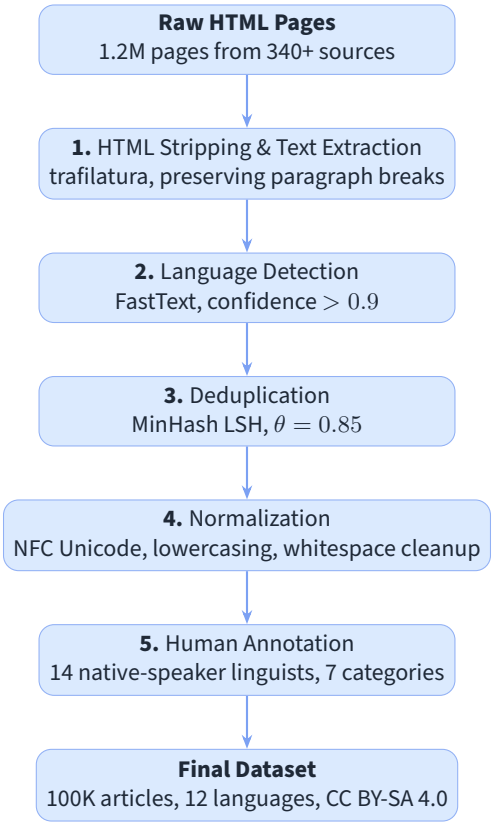

Figure 1: End-to-end preprocessing pipeline from raw web pages to the annotated dataset.

Key Design Decisions

- **No machine translation.** All articles are natively authored in the target language to avoid translationese artifacts that degrade classifier robustness.
- **Minimal text alteration.** Only NFC Unicode normalization and lowercasing were applied. Paragraph boundaries, original letter casing, and diacritics are preserved in a parallel `raw_text` field.
- **Multi-stage dedup.** Exact URL dedup → near-duplicate detection via MinHash → manual inspection of borderline pairs ($0.80 \leq \text{sim} < 0.85$).

Demographic & Linguistic Breakdown

Table 3: Articles per language and language family.

Language	ISO 639-1	Articles	Family
English	en	14,200	Indo-European (Germanic)
Spanish	es	10,800	Indo-European (Romance)
French	fr	9,600	Indo-European (Romance)
German	de	9,200	Indo-European (Germanic)
Portuguese	pt	8,400	Indo-European (Romance)
Russian	ru	8,100	Indo-European (Slavic)
Hindi	hi	7,500	Indo-European (Indo-Aryan)
Arabic	ar	8,900	Afro-Asiatic (Semitic)
Chinese	zh	9,400	Sino-Tibetan
Japanese	ja	6,800	Japonic
Korean	ko	4,500	Koreanic
Turkish	tr	2,600	Turkic
Total		**100,000**	**6 language families**

⚠ **Coverage Gap.** The dataset under-represents Turkic and Koreanic language families. Turkish and Korean each account for fewer than 5% of total articles. We strongly recommend stratified evaluation and caution against treating per-language metrics as equally reliable.

Recommended Tasks

Primary (Well-Supported)

- ✓ **Multilingual text classification** — 7-way news category prediction, single-label. Strong baseline: mBERT achieves 0.82 macro-F1.
- ✓ **Cross-lingual zero-shot transfer** — Train on English, evaluate on all 11 target languages. XLM-R achieves 0.71 avg. zero-shot F1 (range: 0.58–0.84).
- ✓ **Language-agnostic model benchmarking** — Compare mono-, multi-, and language-agnostic architectures on a fixed, diverse testbed.
- ✓ **Fairness auditing** — Measure per-language performance disparities; identify systematic underperformance on low-resource splits.

Secondary (Exploratory)

- ◯ Topic modeling and cross-lingual theme discovery.
- ◯ Named entity recognition fine-tuning (entity annotations not included; must be layered independently).
- ◯ News source bias analysis using the `source_id` metadata field.
- ◯ Temporal distribution shift analysis across the 2018–2023 span.

Usage & Data Format

The dataset is distributed as compressed JSON Lines (`.jsonl.gz`) with one article per line. Each record follows this schema:

```
{
  "id": "en-0047129",
  "language": "en",
  "category": "technology",
  "title": "New Chip Architecture Cuts Data Center Power by 40%",
  "text": "Researchers at MIT have demonstrated a novel...",
  "raw_text": "Researchers at MIT have demonstrated a novel...",
  "source_id": "techcrunch",
  "url": "https://techcrunch.com/2023/03/15/...",
  "date": "2023-03-15",
  "token_count": 412,
  "split": "train"
}
```

Loading via Hugging Face `datasets`:

```
from datasets import load_dataset

dataset = load_dataset("gnlpi/multilingua-100k")
print(dataset["train"][0]["title"])
# "New Chip Architecture Cuts Data Center Power by 40%"
```

Limitations & Ethical Considerations

⚠ **Known Limitations.**

- **News-domain bias.** All articles are journalistic prose; performance does not generalize to social media, legal, or clinical text domains.
- **Temporal cutoff.** Articles end at December 2023. Models trained on this data will not reflect post-2023 events or terminology shifts.
- **Annotation subjectivity.** The 7 categories are broad; edge cases (e.g., *sports technology*) were resolved by majority vote, which may mask genuine semantic ambiguity.
- **Source selection bias.** 340 outlets skew toward major publications; independent and local news sources are under-represented.

Ethical review. This dataset was reviewed by the GNLPI Ethics Board (Approval #GNLPI-2023-042). No personally identifiable information (PII) is included. Articles containing hate speech, graphic violence, or adult content were excluded during Phase III filtering. The scraping process respected `robots.txt` and site-specific terms of service.

Citation & License

If you use MultiLingua-100K in your research, please cite:

```
Marchetti, E., Tanaka, K., & Okafor, S. (2024).
MultiLingua-100K: A Large-Scale Multilingual News Classification Dataset.
In Proceedings of EMNLP 2024 (pp. 4521–4534).
https://aclanthology.org/2024.emnlp-main.XXX
```

License	Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)
Hosting	Hugging Face Datasets
Code	github.com/gnlpi/multilingua-100k
Last updated	July 9, 2026