

Marathi News Classification using Fine-Tuned IndicBERT

A fine-tuned IndicBERT classifier that assigns Marathi news headlines and articles to topic categories, trained on the L3Cube MahaNews corpus with Devanagari-native tokenization and class-weighted training. Delivered built-to-order with training code, inference demo, report, PPT and viva kit.

01. Introduction / Background

Indian-language NLP remains a low-resource domain: high-quality labelled corpora are rare, and multilingual models trained mostly on English underperform on Devanagari text. IndicBERT (AI4Bharat), an ALBERT-based model pretrained on 12 Indian languages, provides a strong starting point. The L3Cube MahaNews corpus — Marathi news headlines and articles labelled into topic classes — is the standard public benchmark for Marathi news classification. This project fine-tunes IndicBERT on MahaNews to classify Marathi news into six topic classes (sports, politics, entertainment, business, technology, health).

02. Problem Statement

Newsrooms, aggregators and researchers dealing with Marathi content need automatic topic tagging, but class imbalance (health articles are rare) and Devanagari tokenization quirks make naive classifiers unreliable.

03. Objectives

- Fine-tune IndicBERT for 6-class Marathi news classification on the MahaNews corpus.
- Handle class imbalance with class-weighted loss and stratified splits.
- Report macro-F1 and per-class F1 honestly on the MahaNews test split.
- Ship an inference demo: paste a Marathi headline, get the class with full probability distribution.

04. Proposed Methodology

Devanagari normalization and IndicNLP tokenization; headline plus first paragraph joined as input. IndicBERT is fine-tuned with a classification head and class-weighted cross-entropy; an 80/10/10 stratified split is used with 5 epochs and early stopping on validation macro-F1. Inference returns a softmax over the six classes.

05. System Architecture / Working

- Input: Marathi headline/article text in Devanagari.
- Preprocessing: normalization, tokenization, headline + first paragraph join.
- Model: IndicBERT fine-tuned with a classification head; class-weighted loss.
- Inference: softmax over 6 classes; top class plus full distribution returned.
- Demo: paste or pick a sample headline and classify instantly.

06. Hardware / Software Requirements

- Python 3, PyTorch, Hugging Face Transformers.
- L3Cube MahaNews corpus; IndicNLP for Devanagari preprocessing.
- GPU recommended for fine-tuning; inference runs on CPU (~12 ms per headline, expected).
- Single-file HTML/CSS/JS demo.

07. Expected Outcomes

- Fine-tuned IndicBERT classifier for 6 Marathi news topics.
- Macro-F1 design target 89.4% and accuracy design target 91.2% on the MahaNews test split — design targets, measured after the run.
- Per-class F1 table identifying the weakest class (health, smallest class).
- Demo, report, PPT and viva Q&A; kit.

08. Applications

- Automatic topic tagging for Marathi news portals and aggregators.
- News recommendation and trending-topic dashboards.
- Content moderation and archival pipelines for regional publishers.
- A template for other low-resource Indic-language classification tasks.

09. Future Scope

- Expand to the full 12 MahaNews topic classes.
- Add Marathi-specific augmentation and domain-adaptive pretraining.
- Multilingual classifier covering Hindi and Marathi together.
- Deploy as a REST API for newsroom ingestion pipelines.