Zero-shot Multilingual NLI Evaluation on XNLI

Rakshitha
Department of Computer Science
CMRIT

Abstract

Multilingual language models are increasingly deployed across languages with diverse linguistic and cultural characteristics. However, their cross-lingual performance is often uneven and poorly understood. In this work, we present a controlled evaluation of multilingual Natural Language Inference (NLI) models on the XNLI benchmark, focusing on English and Indic languages. We analyze zero-shot cross-lingual transfer behavior and discuss implications from a responsible AI perspective.

1 Introduction

Multilingual NLP models such as mBERT and XLM-R promise broad language coverage, enabling deployment across both high-resource and lower-resource languages. Despite this promise, their performance across languages can vary significantly, particularly in zero-shot settings. Understanding these cross-lingual performance gaps is essential for building reliable and equitable language technologies.

This work evaluates multilingual NLI performance across English, Hindi, and Urdu using a standardized benchmark. Rather than optimizing for accuracy, the focus is on analyzing model behavior across languages under controlled experimental conditions.

2 Dataset and Task

We use the XNLI (Cross-lingual Natural Language Inference) dataset, a widely adopted multilingual benchmark. The task involves classifying a premise–hypothesis pair into one of three classes: entailment, neutral, or contradiction.

Evaluation is conducted on the validation split for the following languages:

- English (en)
- Hindi (hi)
- Urdu (ur)

Each language is evaluated independently to enable controlled cross-lingual comparison.

3 Models and Experimental Setup

3.1 mBERT

We evaluate `bert-base-multilingual-cased` as a baseline multilingual model. The model is used in a zero-shot setting, without any task-specific fine-tuning on XNLI.

3.2 XLM-R

We evaluate `xlm-roberta-large-xnli`, a multilingual model fine-tuned on the XNLI task. Despite fine-tuning, evaluation is conducted in a zero-shot multilingual inference setting.

3.3 Evaluation Details

Both models use sentence-pair tokenization with a maximum sequence length of 128. Accuracy is used as the evaluation metric. No additional fine-tuning or calibration is performed.

4 Results and Analysis

Language	mBERT	XLM-R
English	∼0.33	∼0.33
Hindi	∼0.33	∼0.34
Urdu	∼0.33	∼0.33

Table 1: Zero-shot XNLI accuracy across languages

Both models exhibit near-chance accuracy, highlighting the difficulty of zero-shot multilingual NLI. The similarity in performance across languages suggests that uniformly low accuracy can obscure deeper inequities unless carefully analyzed.

5 Responsible AI Implications

Uniform degradation across languages may be mistakenly interpreted as fairness. In practice, such behavior underscores the limitations of deploying multilingual models without task adaptation or language-aware evaluation. Transparent benchmarking is critical to avoid misleading conclusions about equity and model readiness.

6 Conclusion and Future Work

This study demonstrates the importance of controlled multilingual evaluation. Our findings show that zero-shot multilingual NLI remains challenging and sensitive to experimental setup. Future work includes task-specific fine-tuning, evaluation on Indic-focused benchmarks, and analysis using culturally grounded datasets.