

Pre-training Corpus Size and Zero-Shot Cross-Lingual Retrieval Performance in Dense Retrievers

Assignee Research

July 23, 2026

Abstract

Using task-specific pre-training and leveraging cross-lingual transfer are two of the most popular ways to handle code-switched data. In this paper, we aim to compare the effects of both for the task of sentiment analysis. We work with two Dravidian Code-Switched languages - Tamil-English and Malayalam-English and four different BERT based models. We compare the effects of task-specific pre-training and cross-lingual transfer and find that task-specific pre-training results in superior zero-shot and supervised performance when compared to performance achieved by leveraging cross-lingual transfe

1 Introduction

This paper examines: Task-Specific Pre-Training and Cross Lingual Transfer for Code-Switched Data. Research question: How does the choice of pre-training corpus size affect the zero-shot cross-lingual performance of dense retrievers on XNLI, when comparing models pre-trained on small (e.g., 100M tokens) vs. large (e.g., 1B tokens) code-switched datasets, as measured by accuracy differences?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.5/10.

3 Results

13 papers retrieved. 10 claims extracted; 7 independently verified. Quality review score: 7.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The best performing systems for Tamil-English and Malayalam-English sentiment analysis tasks were built on top of BERT v	✓	0.22
The TweetEval sentiment classifier was trained on a dataset of 60M English Tweets.	✓	0.16
The underlying RoBERTa model in TweetEval was trained on English data and has 160M parameters.	✓	0.19
The Tamil-English dataset contains 10,559 positive, 2,037 negative, and 850 neutral samples.	×	0.11
The Malayalam-English dataset contains 2,811 positive, 738 negative, and 1,903 neutral samples.	×	0.09
The Hinglish dataset contains 6,616 positive, 5,892 negative, and 7,492 neutral samples.	×	0.09
Previous works have shown that mBERT and XLM-RoBERTa based models achieve state of the art performance when dealing with	✓	0.24
The uncased-base XLM-RoBERTa model was trained on 2.5TB of webcrawled data.	✓	0.28
The evaluation metrics used are weighted average scores of precision, recall, and F1.	✓	0.27
The weighted average scores are calculated based on the number of samples in each class.	✓	0.25

References

- <http://arxiv.org/abs/2301.06527v1>
- <http://arxiv.org/abs/2102.12407v1>
- <http://arxiv.org/abs/2403.01364v2>