Cross-lingual Euphemism Detection Performance of XLM-R-Large vs. Larger Models on XTREME-R Benchmark

Assignee Research

July 24, 2026

Abstract

Euphemisms are culturally variable and often ambiguous, posing challenges for language models, especially in low-resource settings. This paper investigates how cross-lingual transfer via sequential fine-tuning affects euphemism detection across five languages: English, Spanish, Chinese, Turkish, and Yoruba. We compare sequential fine-tuning with monolingual and simultaneous fine-tuning using XLM-R and mBERT, analyzing how performance is shaped by language pairings, typological features, and pretraining coverage. Results show that sequential fine-tuning with a high-resource L1 improves L2 perfo

1 Introduction

This paper examines: When Does Language Transfer Help? Sequential Fine-Tuning for Cross-Lingual Euphemism Detection. Research question: How does the performance of XLM-R-Large in cross-lingual euphemism detection compare to larger models like BLOOM or PaLM when evaluated on the XTREME-R benchmark using sequential vs. simultaneous fine-tuning?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.8/10.

3 Results

9 papers retrieved. 6 claims extracted; 5 independently verified. Quality review score: 7.8/10.

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 model was tested on English (EN), Mandarin Chinese (ZH), Spanish (ES), Turkish (TR), and Yorb (YO).	✓	0.18
The sequential fine-tuning approach involves learning the same task first on one language (L1) and then on a second lang	×	0.13
The number of examples for euphemism detection in the 2025 PETs Datasets are: ZH (2213 euph, 998 non-euph), EN (1841 eup	✓	0.20
The performance of XLM-R and mBERT on different languages are: EN (XLM-R: 0.821, mBERT: 0.791), ES (XLM-R: 0.768, mBERT:	✓	0.17
The performance of XLM-R and mBERT on different language pairs are: EN & ES (XLM-R: 0.821, mBERT: 0.801), EN & ZH (XLM-R	✓	0.18
The performance of XLM-R and mBERT on sequential fine-tuning for different language pairs are: TR $\rightarrow$ EN (XLM-R: 0.835), E	✓	0.21

References

- http://arxiv.org/abs/2508.11831v1
- http://arxiv.org/abs/2508.11281v3
- http://arxiv.org/abs/2506.15415v1