Scaling XLM-R Model Size and F1 Score Stability in Euphemism Detection Across Languages

Assignee Research

July 22, 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 scaling of XLM-R model size (base vs. large) affect the F1 score stability of euphemism detection when using direct cross-lingual transfer versus sequential fine-tuning across high-resource and low-resource languages?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.9/10.

3 Results

10 papers retrieved. 5 claims extracted; 5 independently verified. Quality review score: 7.9/10.

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 number of examples for euphemism (Euph) and non-euphemism (Non-Euph) in the PETs datasets are as follows: ZH (2213,	✓	0.20
The performance of XLM-R and mBERT models on different languages are as follows: EN (XLM-R: 0.821, mBERT: 0.791), ES (XL	✓	0.17
The performance of XLM-R and mBERT models on different language pairs are as follows: EN & ES (XLM-R: 0.821, mBERT: 0.80	✓	0.18
The performance of XLM-R and mBERT models on sequential fine-tuning for different language pairs are as follows: TR $\rightarrow$ EN	✓	0.21

References

- http://arxiv.org/abs/2506.15415v1
- http://arxiv.org/abs/2602.05599v1
- http://arxiv.org/abs/2508.11831v1