Cross-lingual euphemism detection robustness under adversarial perturbations

Assignee Research

June 20, 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 do different fine-tuning orderings (e.g., English$\rightarrow$Spanish$\rightarrow$Yoruba vs. Spanish$\rightarrow$English$\rightarrow$Yoruba) affect cross-lingual euphemism detection robustness, measured by performance degradation under adversarial perturbations in Yoruba text?.

2 Methodology

Systematic literature search across multiple databases yielded 11 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

11 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 is tested on English (EN), Mandarin Chinese (ZH), Spanish (ES), Turkish (TR), and Yorb (YO) after sequential	✓	0.20
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 (Euph) and non-euphemism (Non-Euph) in the 2025 PETs Datasets are as follows: ZH (2	✓	0.21
The performance of XLM-R and mBERT on different languages are as follows: EN (XLM-R: 0.821, mBERT: 0.791), ES (XLM-R: 0.	✓	0.17
The F1 scores for sequential fine-tuning (Seq.) and simultaneous fine-tuning (Sim.) for XLM-R are as follows: TR $\rightarrow$EN (0.	✓	0.25
The F1 scores for sequential fine-tuning (Seq.) for mBERT are as follows: ZH $\rightarrow$EN (0.812), EN $\rightarrow$ES (0.738), ES $\rightarrow$ZH (0.885)	✓	0.21

References

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