Robustness Variability in XLM-R and mBERT Under Sequential Fine-Tuning and StressGAN Adversarial Attacks

Assignee Research

July 6, 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 order of language exposure in sequential fine-tuning affect the robustness of XLM-R and mBERT against StressGAN-generated adversarial examples, measured by F1-score degradation across high-resource (English, Spanish) and low-resource (Yoruba, Turkish) languages?.

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 7.6/10.

3 Results

13 papers retrieved. 5 claims extracted; 5 independently verified. Quality review score: 7.6/10.

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).	✓	0.18
The number of examples for each class (euphemism vs. non-euphemism) in the 2025 PETs Datasets is as follows: ZH (2213 eu	✓	0.21
The performance of XLM-R and mBERT on monolingual euphemism detection is as follows: XLM-R (EN: 0.821, ES: 0.768, ZH: 0.	✓	0.20
The performance of XLM-R and mBERT on paired language simultaneous fine-tuning is as follows: XLM-R (EN & ES: 0.821, EN	✓	0.26
The performance of XLM-R and mBERT on sequential fine-tuning is as follows: XLM-R (TR $\rightarrow$ EN: 0.835, ES & ZH: 0.768, YO $\rightarrow$	✓	0.26

References

- http://arxiv.org/abs/2509.22472v1
- http://arxiv.org/abs/2508.11831v1
- http://arxiv.org/abs/2204.05814v1