DPO-Aligned Reward Scores and Zero-Shot Cross-Lingual Hate Speech Detection Accuracy

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: What is the correlation between DPO-aligned reward scores and zero-shot cross-lingual transfer accuracy for hate speech detection across the HASOC 2021 Indic language subset. 9 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: One to rule them all: Towards Joint Indic Language Hate Speech Detection. Research question: What is the correlation between DPO-aligned reward scores and zero-shot cross-lingual transfer accuracy for hate speech detection across the HASOC 2021 Indic language subset?.

2 Methodology

Systematic literature search across multiple databases yielded 7 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

7 papers retrieved. 9 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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
Fine-tuned multilingual models beat the baselines by at least % in hate speech detection tasks.	×	0.12
XLM-RoBERTa outperformed mBERT and distilmBERT on hate speech detection tasks for English, Hindi, and Marathi.	×	0.13
XLM-RoBERTa secured the 1st position among 24 participants and the 5th position among 34 participants on the HASOC 2021	×	0.06
Using SOUP (Similarity-based Oversampling and Undersampling processing) resulted in a drop of 5% in accuracy compared to	×	0.05
Data augmentation using back-translation did not result in performance gains and observed a reduction of toxicity.	×	0.03
Applying random forest and LightGBM algorithms resulted in an average drop of 5.3% in performance.	×	0.01
The tweet-preprocessor and ekphrasis libraries were used for preprocessing tweet data and hashtags.	×	0.02
NeuralSpace's transliteration tool and langdetect library were used to extract pure Hindi and Marathi text within tweets	×	0.03
Ekphrasis segmenter was used to segment hashtag text into constituent and meaningful tokens for feature extraction.	×	0.05

References

- http://arxiv.org/abs/2112.09986v1
- http://arxiv.org/abs/2109.10255v4

- http://arxiv.org/abs/2109.13711v1