Multilingual Auxiliary Task Scaling for Zero-Shot Hate Speech Detection in Low-Resource Languages

Assignee Research

June 20, 2026

Abstract

The goal of hate speech detection is to filter negative online content aiming at certain groups of people. Due to the easy accessibility and multilinguality of social media platforms, it is crucial to protect everyone which requires building hate speech detection systems for a wide range of languages. However, the available labeled hate speech datasets are limited, making it difficult to build systems for many languages. In this paper we focus on cross-lingual transfer learning to support hate speech detection in low-resource languages, while highlighting label issues across application scenar

1 Introduction

This paper examines: Label modification and bootstrapping for zero-shot cross-lingual hate speech detection. Research question: How does the number of intermediate multilingual auxiliary tasks impact the zero-shot cross-lingual transfer performance of hate speech detection models on low-resource languages in the XTREME-R benchmark, measured by F1-score improvements across domains?.

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 8.7/10.

3 Results

10 papers retrieved. 9 claims extracted; 9 independently verified. Quality review score: 8.7/10.

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
The goal of hate speech detection is to filter negative online content aiming at certain groups of people.	✓	0.33
Due to the easy accessibility and multilinguality of social media platforms, it is crucial to protect everyone which req	✓	0.39
The available labeled hate speech datasets are limited, making it difficult to build systems for many languages.	✓	0.32
Cross-lingual transfer learning can support hate speech detection in low-resource languages.	✓	0.33
Label issues across application scenarios, such as inconsistent label sets of corpora or differing hate speech definitio	✓	0.38
Cross-lingual word embeddings can be used to train neural network systems on the source language and apply them to the t	✓	0.35
Unlabeled target language data can be incorporated for further model improvements by bootstrapping labels using an ensem	✓	0.29
Label imbalance in hate speech datasets, with a high ratio of non-hate examples compared to hate examples, often leads t	✓	0.39
Simple data undersampling and oversampling techniques can be effective in addressing label imbalance in hate speech data	✓	0.24

References

- https://doi.org/10.1162/tacl_a_00633
- https://doi.org/10.18653/v1/2022.emnlp-main.383

- https://doi.org/10.1007/s10579-023-09637-4