Correlation between Reward Model Calibration and Adversarial Robustness in Multilingual Hate Speech Classifiers

Assignee Research

June 11, 2026

Abstract

Detecting and classifying instances of hate in social media text has been a problem of interest in Natural Language Processing in the recent years. Our work leverages state of the art Transformer language models to identify hate speech in a multilingual setting. Capturing the intent of a post or a comment on social media involves careful evaluation of the language style, semantic content and additional pointers such as hashtags and emojis. In this paper, we look at the problem of identifying whether a Twitter post is hateful and offensive or not. We further discriminate the detected toxic cont

1 Introduction

This paper examines: Leveraging Multilingual Transformers for Hate Speech Detection. Research question: What is the correlation between reward model calibration scores and robustness against adversarial perturbations in multilingual hate speech classifiers?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.7/10.

3 Results

8 papers retrieved. 10 claims extracted; 9 independently verified. Quality review score: 7.7/10.

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 evaluation metric used throughout the study is the macro F1-score.	×	0.15
Perspective API features with a multi-layer perceptron classifier provide respectable results on hate and offensive cont	✓	0.33
In the monolingual mode, the use of identity activation for English and tanh activation for German are the most effectiv	✓	0.31
German Task 2 benefits from the multilingual mode due to additional data from English training examples, allowing the mo	✓	0.25
A drop in English results is witnessed in the multilingual mode, possibly due to the reduction in the number of availabl	✓	0.21
The python libraries tweet-preprocessor and ekphrasis are used for tweet tokenization and hashtag segmentation, respecti	✓	0.24
For Hindi tweets, tokenization is done on whitespaces and symbols including colons, commas, and semicolons, followed by	✓	0.31
Hashtag text is segmented into meaningful tokens using the ekphrasis segmenter for the Twitter corpus.	✓	0.23
Information such as URLs, name mentions, quantitative values, and smileys are saved as features for classifiers.	✓	0.16
Emojis are processed using emoji2vec to obtain a semantic vector representing the particular emoji.	✓	0.21

References

- http://arxiv.org/abs/2101.03207v1
- http://arxiv.org/abs/2112.09986v1
- http://arxiv.org/abs/2109.13711v1