

## Enhancing Sinhala Hate Speech and Offensive Language Detection through XAI

Pilapitiya H.M.H.N.<sup>1\*</sup>, Ishanka U.A.P.<sup>2</sup>

<sup>1</sup>Department of Computing and Information Systems,  
Faculty of Computing, Sabaragamuwa University of Sri Lanka, Sri Lanka

<sup>2</sup>Department of Data Science,  
Faculty of Computing, Sabaragamuwa University of Sri Lanka, Sri Lanka  
\*hmhnpilapitiya@std.appsc.sab.ac.lk

The rapid growth of social media has sped up the spread of hatred, abusive and offensive content, creating an urgent need for automated detection systems especially for low-resource languages such as Sinhala. This study develops, and evaluates standard deep learning (DL) models, transformer-based architectures, and a newly proposed hybrid model using DL models with Sinhala Offensive Language Dataset (SOLD) for detecting hate speech in Sinhala. Among Conventional DL models, Bi-LSTM demonstrated the strongest performance with 82% accuracy and a ROC score of 0.88, despite the challenge of informal expressions, and rich morphology in Sinhala language. Among transformer-based models, XLM-R large achieved the best results with 84% accuracy and an ROC of 0.92, demonstrating their effectiveness in modeling nuanced semantic and syntactic structures in Sinhala online discourse. Moreover, a hybrid model integrating multiple DL models was developed and evaluated, achieving superior performance with an accuracy of 86% with a ROC of 0.93 on Sinhala hate speech detection task, outperforming all baseline deep learning and transformer-based models. Beyond predictive performance, this study also contributes important interpretability of model predictions with Explainable AI (XAI) techniques - SHAP and LIME - which detail the local and global contributions of tokens. These explanations provide clear pathways to make decisions about models. Overall, this study presents a comprehensive evaluation about model's performance and interpretability analysis of model's predictions to detect hate speech in Sinhala with exciting brevity for future multilingual and explainable NLP applications.

**Keywords:** *Deep learning; Hybrid model; Sinhala hate speech; Transformerbased models; XAI*