ABSTRACT

Cyberbullying is an act of violence commonly committed on online platforms such as social media X, often causing psychological effects for victims. Despite prevention efforts, traditional methods for detecting cyberbullying show limited effectiveness due to the complexity of language and diversity of expressions, leading to suboptimal performance. This study aims to enhance detection accuracy by applying Convolutional Neural Networks (CNN) and Gated Recurrent Unit (GRU) with an attention mechanism to analyze textual data from tweets. The model uses Term Frequency-Inverse Document Frequency (TF-IDF) for extracting important words and Word2Vec for expanding text representation. A total of 30,084 labeled datasets from tweets on social media X were utilized. Results indicate the hybrid CNN-GRU model with attention achieved the highest accuracy of 80.96%, outperforming stand-alone CNN and GRU models. Additionally, TF-IDF and Word2Vec significantly improved model performance, with the CNN-GRU combination proving most effective for detecting cyberbullying. This study contributes to computer science by proposing a novel approach that integrates CNN, GRU, and attention mechanisms with advanced feature extraction techniques, providing a more reliable detection system for online platforms. It also highlights the potential for integrating multimodal data to further enhance future performance.

Keywords: *Attention Mechanism, Convolutional Neural Network (CNN), Cyberbullying Detection, Gated Recurrent Unit (GRU), Word2Vec*