ABSTRACT

One of the unintended effects of the rapid proliferation of social media platforms is the spread of cyberbullying, which has become a major problem that severely affects the psychological health of users. Conventional detection methods are often incapable of understanding the contextual subtleties of online abuse, and typical intervention strategies are not sufficiently adaptable. In order to overcome these obstacles, this article presents a sophisticated cyberbullying identification and alleviation system based on a Bidirectional Long Short-Term Memory (Bi-LSTM) network.Compared to a unidirectional model, the Bi-LSTM framework reads the text data not only in the forward but also in the backward direction, thus it can keep a larger semantic context is very important for the identification of the less obvious forms of abuse. The proposed method, which uses a comprehensive dataset from Kaggle, categorizes the offensive content into three different degrees of toxicity: Low, Medium, and Intensive. Besides that, the platform features an innovative user-intervention tool that is founded on a fluctuating reputation score; if the reputation score of a user drops below the crucial limit of 10.0, then that user is automatically blocked so that no more damage can be done.The main point of the experimental results is to show that this method is very effective in terms of classification and performance and that it is able to significantly outperform the traditional baselines. Hence, it provides a reliable and scalable tool to make the digital world safer not only through accurate detection but also by proactive user management.

Keywords: Cyberbullying Detection, Bidirectional LSTM, Multi-level Toxicity, Reputation Score, User Blocking, Social Media Safety.