ICCP/SL/OP/016

AI-Powered web filter for detecting and preventing unsafe and inappropriate content to ensure children's online safety

Piyumal KM*, Perera MVV

Department of Computer Systems Engineering, University of Kelaniya, Sri Lanka
*madhushkak@kln.ac.lk

Background: The rapid increase of children's engagement of online platforms for learning and entertainment, ensuring their digital safety become more crucial than ever. Children are highly vulnerable to expose unsafe and inappropriate content-based websites. Adversaries send unsafe URLs (Universal Resource Locator) using various methods, including email, SMS (Short Message Send), and social media to trick the kids and lure children into accessing such inappropriate content. Unsafe URLs were traditionally identified using blacklist and whitelist techniques. The inadequacy of detecting zero-day attacks using these traditional methods, this study focuses on artificial intelligence (AI)-based approaches for unsafe URL detection.

Method: The study uses 11,430 size labeled benchmark dataset including 52 distinct features and RFE (Recursive Feature Elimination) is used to eliminate the unwanted features. The dataset is used to train five machine learning algorithms, including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), XGBoost (XGB), Voting Classifier (VC) (using LR, RF, XGB) algorithms, and one of the deep learning algorithms, the Multilayer Perceptron (MLP). Hyperparameter tuning is performed using the grid search algorithm to enhance model performance. Among the evaluated models, the best-performing one is integrated into a web-based platform to evaluate its performance on unseen data.

Results: The models Logistic Regression, Decision Tree, Random Forest, XGBoost, Voting Classifier and Multilayer Perceptron achieved accuracy scores of 94.88%, 94.31%, 96.98%, 97.28%, 96.98%, and 82.15% respectively, using feature counts of 80, 46, 72, 52, 80, 87.

Conclusions: The results analysis supports the recommendation of the XGBoost algorithm to ingrate with the web-based platform among the algorithms due to its highest accuracy (97.28%) and minimal feature requirement (52), which contribute to enhanced platform performance.

Keywords: Social engineering, child protection, machine learning, deep learning.