EmailSpamML: Optimized ML Algorithms for Spam Email Detection, Performance Insights, and Future Research Directions

Sandeep Reddy KALLURI, Deeksha Reddy BILLA, Akshitha Reddy CHINTHALA, and MD Masud RANA
Department of Computer Science
Lamar University, Beaumont, Texas USA

Corresponding Author: mrana15@lamar.edu

Abstract— In the age of digital communication and AI revolution, the spam email detection and classification has become a fundamental task to guarantee the efficiency and security of email methods. This framework explores the adaptation of Decision Tree, KNN, and XGBoost ML algorithms, to classify emails as spam or ham. The work has used preprocessing and feature extraction techniques, such as tokenization, stemming, and stopword removal. After gathering the email data from Kaggle, it has been splitted into the training (70%) and testing (30%) sets. After training and testing, the simulation results demonstrate that well-implemented preprocessing and feature engineering significantly enhance the performance of spam classification models. Specifically, XGBoost outperformed the other two models with the highest accuracy (97%) and AUC (0.97), showing strong precision (98%) and recall (98%) for both classes. KNN confirmed a high accuracy (95%) and strong precision (94%) and recall (99%) for class 0, while Decision Tree had an accuracy of 92%, but displayed a drop in performance for class 1, making XGBoost the most robust procedure for spam email classification, while KNN could be favorite for its
easiness. Future work incorporates investigating the application of advanced techniques such as deep learning and reinforcement learning to future optimize spam detection and tackle emerging email threats.

Keyword— Spam Email Detection, Decision Tree, K-Nearest Neighbors, XGBoost.

Sandeep Reddy KALLURI has been completed the Msters Degree from Lamar University, Texas. The research interest include Spam Email Detection, Decision Tree, K-Nearest Neighbors, XGBoost.

Deeksha Reddy BILLA has been completed the Msters Degree from Lamar University, Texas. The research interest include Spam Email Detection, Decision Tree, K-Nearest Neighbors, XGBoost.

Akshitha Reddy CHINTHALA has been completed the Msters Degree from Lamar University, Texas. The research interest include Spam Email Detection, Decision Tree, K-Nearest Neighbors, XGBoost.

Dr Rana is an Assistant Professor at Lamar University, Texas. Dr. Rana leads research at the intersection of generative AI and cybersecurity, with a particular emphasis on offensive and defensive security strategies and penetration testing.