World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:18, No:05, 2024

# Local Interpretable Model-agnostic Explanations (LIME) Approach to Email Spam Detection

**Authors :** Rohini Hariharan, Yazhini R., Blessy Maria Mathew

**Abstract :** The task of detecting email spam is a very important one in the era of digital technology that needs effective ways of curbing unwanted messages. This paper presents an approach aimed at making email spam categorization algorithms transparent, reliable and more trustworthy by incorporating Local Interpretable Model-agnostic Explanations (LIME). Our technique assists in providing interpretable explanations for specific classifications of emails to help users understand the decision-making process by the model. In this study, we developed a complete pipeline that incorporates LIME into the spam classification framework and allows creating simplified, interpretable models tailored to individual emails. LIME identifies influential terms, pointing out key elements that drive classification results, thus reducing opacity inherent in conventional machine learning models. Additionally, we suggest a visualization scheme for displaying keywords that will improve understanding of categorization decisions by users. We test our method on a diverse email dataset and compare its performance with various baseline models, such as Gaussian Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes, Support Vector Classifier, K-Nearest Neighbors, Decision Tree, and Logistic Regression. Our testing results show that our model surpasses all other models, achieving an accuracy of 96.59% and a precision of 99.12%.

**Keywords :** text classification, LIME (local interpretable model-agnostic explanations), stemming, tokenization, logistic regression.
**Conference Title :** ICCSCIT 2024 : International Conference on Computer Science, Cybersecurity and Information Technology
**Conference Location :** Honolulu, United States
**Conference Dates :** May 02-03, 2024