A Comparative Analysis of Classical Machine Learning Algorithms for Email Spam Detection

Dhanna Singh
Department of Computer Science
Desh Bhagat College
Sangrur,Punjab,India
Singh.dhann@gmail.com

Abstract

Email spam continues to be a significant challenge due to the increasing volume of unsolicited and malicious messages. This paper presents a comparative analysis of classical machine learning algorithms for email spam detection using the UCI Spambase dataset. Several widely used classifiers, including Naïve Bayes, Support Vector Machine, Logistic Regression, Decision Tree, and Random Forest, are evaluated under identical experimental conditions. Performance is measured by precision, recall, and F1-score ,accuracy. Experimental results demonstrate that the Random Forest classifier outperforms other models, achieving an accuracy of 94.57% with a high precision and balanced recall. In addition to performance evaluation, model interpretability is enhanced using SHAP (SHapley Additive exPlanations) to analyse feature contributions influencing spam classification decisions. The findings indicate that classical machine learning models, when combined with explainability techniques, can provide reliable and interpretable solutions for email spam filtering.

Keywords

Email Spam Detection, Machine Learning, Random Forest, SHAP, Text Classification.

Introduction

Email remains one of the most widely used communication tools for personal and professional purposes. However, the rapid growth of unsolicited and malicious emails, commonly referred to as spam, has significantly reduced the reliability and efficiency of email systems. Spam emails not only consume network resources but also pose serious security risks, including phishing attacks, malware distribution, and identity theft.

Traditional rule-based spam filtering techniques are increasingly ineffective due to the dynamic and adaptive nature of spam content. Consequently, machine learning-based approaches have gained prominence for their ability to learn patterns from data and adapt to evolving spam strategies. Numerous classifiers have been proposed for spam detection, ranging from probabilistic models to ensemble-based approaches. Many studies focus on individual models or use inconsistent experimental settings, making direct comparison difficult.

This study aims to address this issue by conducting a systematic comparison of five widely used classical machine learning algorithms for email spam detection. The primary contributions of this work are:

- Evaluation of multiple classifiers under identical pre-processing and experimental conditions
- Comparative analysis using standard performance metrics
- Discussion of model-specific strengths and limitations for practical deployment

Related Work

Email spam detection has been an active research area for several decades. Early approaches relied on rule-based systems and keyword matching techniques, which required extensive manual effort and were vulnerable to obfuscation strategies used by spammers.

Probabilistic models such as Naive Bayes became popular due to their simplicity and effectiveness in text classification tasks. Several studies have demonstrated the robustness of Naive Bayes classifiers for spam detection, particularly in scenarios where recall is prioritized.

Logistic Regression and Support Vector Machines have also been widely adopted due to their strong theoretical foundations and ability to handle high-dimensional feature spaces. SVMs, in particular, have shown competitive performance in text classification tasks by maximizing the margin between classes.

Ensemble methods such as Random Forest have gained attention for their ability to combine multiple decision trees and capture complex feature interactions. Prior research indicates that ensemble-based classifiers often outperform individual models in terms of accuracy and robustness.

While existing studies demonstrate the effectiveness of various algorithms, differences in datasets, feature extraction techniques, and evaluation protocols limit direct comparison. This work addresses these limitations by employing a unified experimental framework.

Dataset Description

The UCI Spambase dataset was used in this study to evaluate classifier performance. The dataset consists of 4,601 email instances, each represented by 57 continuous features derived from email content, such as word frequencies and character statistics. Each instance is described as either not-spam or a spam.

The dataset is publicly available and widely used in spam detection research, making it suitable for benchmarking and comparative analysis.

Methodology

The dataset was divided into training and testing subsets using an 80:20 stratified split to preserve class distribution. Feature standardization was applied for Logistic Regression and SVM, while tree-based models were trained on raw features.

Five machine learning classifiers were evaluated:

- Naive Bayes
- Logistic Regression
- Support Vector Machine
- Decision Tree
- Random Forest

All models were trained and tested using the same dataset split and pre-processing pipeline. The experiments were implemented in Python using the scikit-learn library.

Experimental Results and Discussion

Classifier	Accuracy (%)	Precision	Recall	F1-score
Naive Bayes	83.27	0.71	0.86	0.78
Logistic Regression	92.41	0.92	0.89	0.90
Support Vector Machine	93.38	0.93	0.90	0.91
Decision Tree	90.12	0.89	0.88	0.88
Random Forest	**94.57**	**0.95**	**0.91**	**0.93**

Table 1: Performance Comparison of Machine Learning Models

Naive Bayes achieved the highest recall, indicating its effectiveness in identifying spam emails; however, its lower precision suggests a higher false-positive rate. Logistic Regression and SVM exhibited balanced performance across all metrics, making them suitable for scenarios requiring reliable classification and interpretability. Random Forest achieved the best overall performance, demonstrating superior accuracy and F1-score due to its ensemble-based learning approach. These results indicate that no single model is universally optimal, and classifier selection should depend on application-specific priorities.

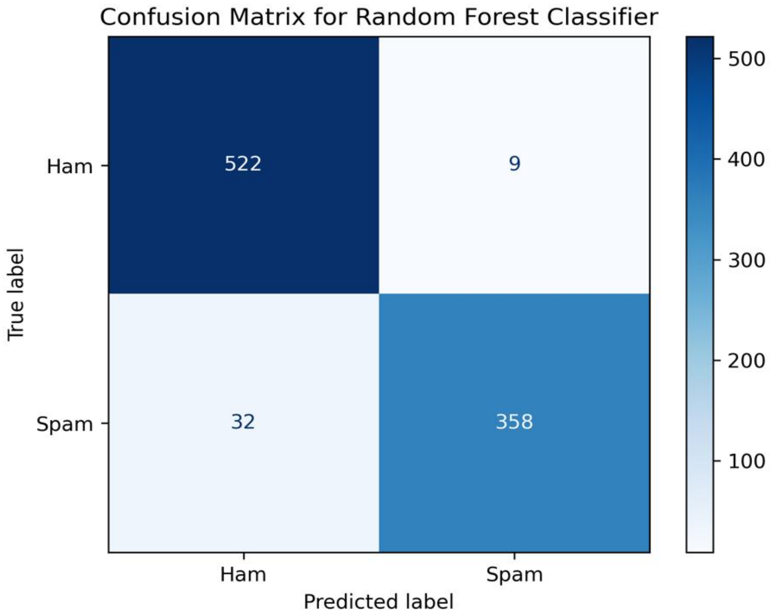

Fig. 1. Confusion matrix of the Random Forest classifier on the UCI Spambase dataset.

Confusion Matrix Analysis:
Figure 1 presents the confusion matrix obtained using the Random Forest classifier. The model correctly classifies the majority of legitimate and spam emails. Only nine legitimate emails are misclassified as spam, indicating a very low false positive rate, which is critical in practical email filtering systems. Although a small number of spam emails are misclassified as legitimate, the overall results demonstrate a balanced and reliable classification performance.

Model Interpretability and Explainability

To enhance model transparency, SHAP (SHapley Additive exPlanations) was employed to interpret the predictions of the Random Forest classifier. The SHAP summary plot illustrates the relative importance of features and their contribution to spam classification. Features with higher SHAP values exert greater influence on the model's decision, where positive values indicate increased likelihood of spam classification. This analysis confirms that the model relies on meaningful attributes rather than random correlations.

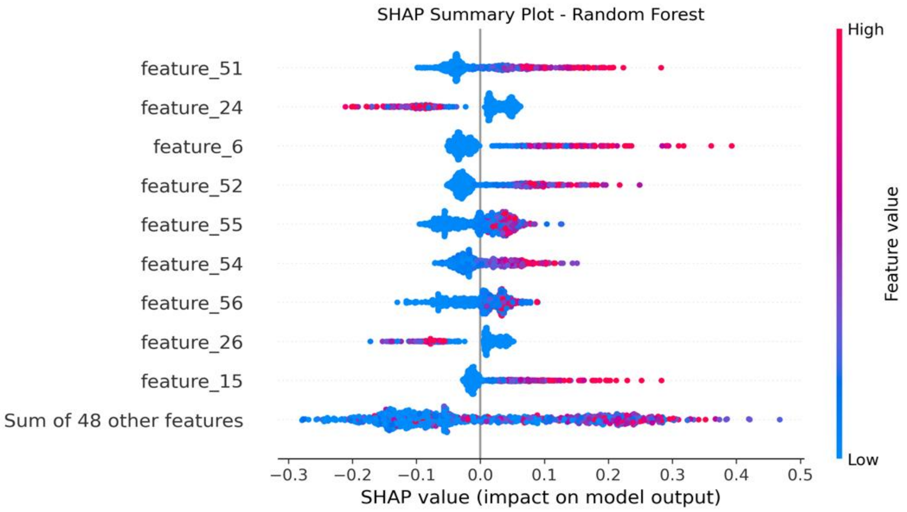

Fig. 2. SHAP Summary Plot Showing Feature Contributions in Spam Classification

Conclusion and Future Work

This study presented a comparative evaluation of classical machine learning algorithms for email spam detection using the UCI Spambase dataset. Multiple classifiers were analysed to assess their effectiveness in distinguishing spam from legitimate emails. Among the evaluated models, the Random Forest classifier demonstrated superior performance in terms of accuracy, precision, recall, and F1-score, indicating its robustness and reliability for spam filtering tasks. Furthermore, the application of SHAP provided meaningful insights into feature importance and model behaviour, improving transparency and interpretability of the classification decisions. The results confirm that classical machine learning approaches, when supported by explainable AI techniques, remain effective for real-world spam detection systems. Future work may explore hybrid models or deep learning approaches to further enhance detection performance on more diverse and evolving datasets.

References

1. Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository*. University of California, Irvine.
2. Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., & Spyropoulos, C. D. (2000). An evaluation of naive Bayesian anti-spam filtering. *Proceedings of the Workshop on Machine Learning in the New Information Age*.
3. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
4. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
5. Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.
6. Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), 832–844.
7. Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.
8. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
9. Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning* (3rd ed.). Packt Publishing.
10. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.
11. Alzahrani, S. M., & Alotaibi, S. S. (2021). Spam email detection using machine learning techniques. *Journal of Information Security and Applications*, 58, 102712.
12. Kumar, A., & Sharma, R. (2020). Email spam classification using ensemble learning methods. *International Journal of Computer Applications*, 176(20), 1–6.
13. Singh, P., & Kaur, G. (2022). A comparative study of machine learning algorithms for spam detection. *International Journal of Computer Science and Information Security*, 20(2), 45–52.
14. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
15. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD Conference*.
16. Zhang, Y., & Wallace, B. (2017). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *Proceedings of IJCNLP*.
17. Saini, R., & Kaur, P. (2023). Explainable artificial intelligence techniques in text classification: A review. *Applied Artificial Intelligence*, 37(1), 1–20.
18. Bansal, S., & Gupta, N. (2021). Machine learning-based approaches for spam detection: A survey. *International Journal of Information Technology*, 13(4), 1375–1385.