International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Comparative Study of Machine Learning Algorithms for Email Spam Detection

Mr. Dushyant Kaushik

Assistant Professor, Computer Science & Engineering, Meri College Of Engineering & Technology

Abstract

The conflict between spam emails and user inboxes has recently gained attention from cybercriminals, making the identification of spam a critical process for both users and businesses. In this regard, we analyze the performance of three widely used machine learning techniques for classifying email spam using the UCI Spambase dataset—Naive Bayes, Support Vector Machines (SVM) and Random Forest (RF). Each model will be evaluated based on achieved accuracy, precision, recall, F1 score computed value alongside training time. Although Random Forest Classifier performed best with greater accuracy measurement than comparative models, Naive Bayes classifier excelled at fast processing speeds.

1. Introduction

While spam messages may be unwelcome in any form of electronic communication, their relentless intrusion into Email – one of the most common mediums of communication among individuals– makes Email filtering an appealing target for both phishing strategies as well as more complex malicious attacks. A significant challenge caused due to Banking Scams is that almost 50% accounts for these types of communicational related services costs as dry decreases in revenue along with losing hurdles everyday. While adhering limitations introduced by traditional filtration systems built upon fixed stratagies have low adaptability problem solving can use futernistic approach through computation intelligence such as Machine Learning (ML).

This study analyzes the performance of Naive Bayes, Support Vector Machine (SVM), and Random Forest – three popular machine learning techniques used for spam email classification. The algorithms were tested using a publicly available dataset to evaluate their relative merits and demerits with an aim towards building effective spam filtering technologies.

2. Related Work

The problem of spam has generated a great deal of interest over the years, resulting in numerous proposed solutions. Most early solutions relied heavily on hand-crafted filtering rules tailored around specific domains. While these provided some initial level of detection or identification, they tend to be static and are no longer applicable in today's dynamic environments. With advances in artificial intelligence, it is now possible to create adaptive filters which learn intelligently from changing patterns.

Almeida et al. (2011) emphasized the use of large labeled datasets for training ML models based on content-based spamfiltering strategies. One of the earliest works analyzing comparative performance across several statistical spam filtering methodologies was conducted by Zhang et al. (2004). Their work showed that Naïve Bayes performed much better than expected given its simplicity. More recent studies have attempted to utilize advanced deep learning and recurrent neural network(RNN) transformers

models; however, their complexity coupled with heavy resource demands makes real-time implementation difficult.

This study looks into the application of classical machine learning (ML) approaches due to their lower computational cost and simpler implementation in resource constrained settings like mobile devices or small organizations.

3. Dataset and Methodology

3.1 Dataset

In this analysis, we utilize the UCI Spambase dataset comprising 4,601 email messages categorized as spam or non-spam (ham). Each message from the collection is represented with 57 features like word frequencies and other linguistic patterns. Roughly 39.4% of emails are labeled as spam. Given its size, diversity of features, class balance, amd suitability for training and evaluating ML systems, it served as a suitable dataset for this experiment.

3.2 Data Preprocessing

Preprocessing is an integral stage of every ML workflow. In the case of Spambase dataset the following steps were performed: -Normalization: Feature scaling helps control imbalance across multiple input parameters using Min-Max normalization bounds all inputs within a range. -Data Split: The dataset is split into a training set that makes up 80% of the data and a testing set that contains 20%. -Feature Selection: Full setof features was used to achieve meaningful models that can be compared during evaluation by preserving inter-model comparability which therefore increased reliability.

3.3 Algorithms Used - **Naïve Bayes (NB)**: A probabilistic classifier based on Bayes' Theorem with the assumption of feature independence. It is highly scalable and efficient for text classification. - **Support Vector Machine (SVM)**: A powerful classification algorithm that constructs an optimal hyperplane to separate data points of different classes. It performs well in high-dimensional spaces. - **Random Forest (RF)**: An ensemble learning method that builds multiple decision trees and combines their outputs. It tends to provide high accuracy and robustness against overfitting.

4. Experimental Setup

The models were implemented using the Python programming language with libraries such as scikit-learn and pandas. Each algorithm was trained and evaluated using the same training and testing datasets to ensure fair comparison. The evaluation metrics include: - **Accuracy**: The ratio of correctly predicted observations to the total observations. - **Precision**: The ratio of correctly predicted spam messages to the total predicted spam messages. - **Recall**: The ratio of correctly predicted spam messages to all actual spam messages. - **F1 Score**: The harmonic mean of precision and recall. - **Training Time**: Time taken to train each model.

5. Results and Discussion

Metric	Naïve Bayes	SVM	Random Forest
Accuracy	89.2%	94.7%	**96.3%**
Precision	88.1%	93.5%	**96.0%**
Recall	90.3%	94.9%	**97.2%**
F1 Score	89.2%	94.2%	**96.6%**
Training Time	**1.2 sec**	6.5 sec	8.3 sec

From the results, we observe that Random Forest consistently outperforms the other algorithms across all classification metrics. Its ensemble nature allows it to handle a variety of feature types and provides robustness against noise. However, this comes at the cost of increased training time and model complexity.

SVM also performs well, particularly in achieving a balance between accuracy and generalization. It is suitable for applications requiring higher accuracy but where computational resources are not significantly limited.

Naïve Bayes, though less accurate than the other two, is extremely fast and requires minimal memory. This makes it ideal for real-time spam filtering on devices with limited computational capabilities, such as smartphones and embedded systems.

6. Comparative Analysis

In real-world scenarios, the choice of a spam detection algorithm must consider trade-offs among accuracy, speed, and resource usage. The results show that while Random Forest is the most accurate, it is not the most efficient in terms of computational resources. Naïve Bayes is the best choice for environments where speed is critical, and SVM offers a middle ground.

Factor	Naïve Bayes	SVM	Random Forest
Accuracy	Low	Medium	High
Speed	High	Medium	Low
Resource Usage	Low	Medium	High
Scalability	High	Medium	High
Ease of Use	High	Medium	Medium

7. Limitations and Future Work

One limitation of this study is the reliance on a single dataset. While the UCI Spambase dataset is widely used, it may not represent the diversity of spam emails in real-world environments. Additionally, more advanced techniques such as feature engineering and hyperparameter tuning were not extensively explored.

Future work can extend this study by: - Testing on diverse datasets from different domains and languages. - Incorporating deep learning models for improved performance. - Using real-time deployment scenarios to evaluate practical feasibility. - Evaluating robustness against adversarial attacks.

8. Conclusion

This study provides a comparative analysis of three classical machine learning algorithms for email spam detection. Random Forest achieves the best overall performance but requires more computational resources. Naïve Bayes, though less accurate, excels in speed and simplicity, making it suitable for real-time or constrained environments. SVM offers a balance between the two. The findings suggest that the choice of algorithm should align with the specific requirements of the application environment.

References

1. T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, “Contributions to the study of SMS spam filtering: new collection and results,” in Proceedings of the 11th ACM Symposium on Document

Engineering, 2011, pp. 259–262.
2. L. Zhang, J. Zhu, and T. Yao, “An evaluation of statistical spam filtering techniques,” ACM Trans. Asian Lang. Inf. Process., vol. 3, no. 4, pp. 243–269, 2004.
3. UCI Machine Learning Repository, “Spambase Data Set.” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/spambase
4. C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
5. L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
6. I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Morgan Kaufmann, 2016.