Society for Multidisciplinary & Advanced Research Techniques (SMART)
West Midlands Open University – Projects, Research, Innovations, Strategies & Multimedia (PRISM) Centre
SMART Scientific Projects & Research Consortium (SMART SPaRC)
Sekinah-Hope Foundation for Female STEM Education
ICT University Foundations USA
Harmath Global Educational Services

38th International Science Technology Education Arts Management
& Social Sciences (iSTEAMS) Bespoke Conference - Accra Ghana 2024

# A Comparative Study of the Effectiveness of Selected Machine Learning Algorithms for Scalability and Accuracy in Filtering Email Spams

***[1]Fatimah Adamu-Fika, [2]Blessing Ashedze, [3]Kamaludeen Shehu Bature, [4]Dauda Sule, [5]Suleiman Abu Usman, [6]Henry Onyeoma Mafua, & [7]Onyinye Vivian Okpoko**
[1]Department of Cyber Security, Air Force Institute of Technology Kaduna
[2,3,4,5,6,7]Department of Computer Science, Air Force Institute of Technology Kaduna
*Corresponding Author: f.adamu-fika@afit.edu.ng

## ABSTRACT

The rise in email spam necessitates effective filtering systems. This study examines three machine learning algorithms—Naive Bayes, Random Forest, and Support Vector Machines (SVM)—for spam filtering. Using a diverse dataset, we evaluated each model on accuracy, precision, recall, F1-score, and scalability. SVM outperformed others with a recall of 93.4% and an F1-score of 95.5%, making it suitable for high-security needs. Naive Bayes was efficient, achieving consistent accuracy of 98.3% regardless of dataset size. Random Forest excelled in handling noisy data, boasting a precision of 99.3%. The results emphasise the trade-offs between scalability and accuracy, aiding in the selection of algorithms based on specific requirements. This study connects theoretical advancements to practical applications and suggests ways to enhance spam filtering systems while considering ensemble methods for future research.

**Keywords:** Email Spam Filtering, Machine Learning Algorithms, Naive Bayes, Random Forest, Support Vector Machines (SVM).

**Proceedings Citation Format**
Fatimah Adamu-Fika, Blessing Ashedze, Kamaludeen Shehu Bature, Dauda Sule, Suleiman Abu Usman, Henry Onyeoma Mafua, & Onyinye Vivian Okpoko (2024): A Comparative Study of the Effectiveness of Selected Machine Learning Algorithms for Scalability and Accuracy in Filtering Email Spams. Proceedings of the 38th iSTEAMS Multidisciplinary Bespoke Conference. 17th – 19th July, 2024. University of Ghana, Accra, Ghana. Pp 330-343. dx.doi.org/10.22624/AIMS/ACCRABESPOKE2024P33xx

## 1. INTRODUCTION

Email continues to be one of the most widely used communication tools around the world, facilitating both personal and professional interactions. Yet, despite its widespread use, the persistent issue of spam emails presents substantial challenges. Spam not only consumes bandwidth and storage resources but also often acts as a vehicle for phishing, malware, and fraud.

In fact, reports indicate that nearly half of all global email traffic consists of spam, highlighting the urgent need for effective detection systems (Radicati Group, 2022). In

response to this challenge, machine learning has become a promising method for detecting spam, overcoming the limitations of traditional rule-based systems.
These models dynamically learn from data patterns, allowing them to identify even the most sophisticated spam tactics. However, when it comes to implementing machine learning models in real-world scenarios, it is essential to consider factors like scalability and computational efficiency, especially given the diverse and ever-changing nature of datasets.

Although machine learning models have demonstrated significant potential for spam filtering, existing research predominantly focuses on improving accuracy in controlled experimental settings. This narrow emphasis often overlooks the critical issue of scalability. Spam filtering systems must efficiently handle large and heterogeneous datasets, ensuring consistent performance across diverse email types and evolving spam strategies. Ignoring scalability compromises the practical applicability of these systems, especially in large-scale, resource-constrained environments.

This research examines the scalability challenges of machine learning models in spam filtering, contributing to the discourse on their practical deployment. It evaluates Naive Bayes, Random Forest, and SVM using a dataset representative of real-world email traffic. The findings offer a comparative analysis highlighting each model's strengths and limitations, equipping practitioners with clear, evidence-based recommendations. This study advances the design of spam filtering systems capable of handling diverse datasets and evolving spam tactics, addressing gaps in existing literature.

### 1.1 Research Objectives

This study aims to assess and compare the scalability and accuracy of three widely used machine learning algorithms—Naive Bayes, Random Forest, and SVM—in the context of email spam filtering. The objectives of the study include:

1. Use key metrics like accuracy, precision, recall, and F1-score to evaluate model performance.
2. Analyse the trade-offs between scalability and accuracy as the dataset size increases.
3. Provide practical insights to help practitioners choose and optimise models tailored to various operational contexts.

## 2. RELATED WORKS

The rapid evolution of spam tactics has led to the creation of advanced detection systems that prioritise adaptability and efficiency. While early spam detection relied on rule-based systems, machine learning approaches have since taken precedence for their ability to learn patterns and adapt. This section reviews literature on spam filtering techniques, particularly Naive Bayes, Random Forest, and SVM, while noting gaps in research regarding scalability and ethical issues, positioning this study within the larger context of spam filtering research.

### 2.1 Introduction to Spam Filtering Research

Spam filtering is crucial in cybersecurity and machine learning due to its impact on user security and system efficiency. Early systems used rule-based approaches, which required frequent manual updates and struggled with evolving spam tactics (Patan et al., 2020). Machine learning improved this by enabling systems to learn from data, enhancing accuracy and robustness. Effective algorithms like Naive Bayes, Random Forest, and Support Vector Machines (SVM) are commonly used for spam detection.

## 2.2 Naive Bayes in Spam Filtering

Naive Bayes is a probabilistic classifier widely used for text classification due to its simplicity and computational efficiency. It assumes feature independence, which, although unrealistic in some cases, often produces satisfactory results (Reddy et al., 2022). It performs well with high-dimensional data, making it effective for email spam filtering, but its independence assumption may limit its ability to capture complex relationships in nuanced spam emails (Ghantasala et al., 2020). Nonetheless, its speed and low computational cost are advantageous for large-scale applications, especially in resource-constrained environments (Krishna et al., 2019).

## 2.3 Random Forest in Spam Filtering

Random Forest is an ensemble learning method known for its robustness against overfitting and effectiveness with noisy and unbalanced datasets. It enhances classification accuracy by aggregating predictions from multiple decision trees (Faris et al., 2016) and is particularly good at identifying key features, making it suitable for spam email classification. However, its higher computational demands compared to simpler models like Naive Bayes can pose scalability challenges for real-time applications (Kontsewaye et al., 2020).

## 2.4 Support Vector Machines (SVM) in Spam Filtering

SVM are widely regarded for their strong theoretical foundation and effectiveness in binary classification tasks. SVM is particularly well-suited for high-dimensional feature spaces, where it constructs optimal hyperplanes to separate data points from different classes (Amayri & Bouguila, 2010). The flexibility of SVM can be further enhanced by kernel functions, allowing it to model non-linear relationships. Despite these strengths, SVM often requires extensive parameter tuning and computational resources, which can limit its scalability for large-scale or dynamic spam filtering systems (Sculley & Wachman, 2007).

## 2.5 Gaps in Current Research

Most existing studies focus on improving algorithmic accuracy in controlled experimental settings, often using homogeneous datasets lacking the diversity of real-world email traffic (Peng et al., 2018). This narrow scope leaves critical issues such as scalability and computational efficiency underexplored. Moreover, while ethical considerations, including false positives and user privacy, are occasionally mentioned, they are rarely discussed comprehensively. These gaps limit the practical applicability of machine learning-based spam filters in operational contexts where large-scale, dynamic datasets are the norm.

## 2.6 Contribution of this Study

This study addresses the identified gaps by:

1. Evaluating the scalability and accuracy of Naive Bayes, Random Forest, and SVM using a diverse dataset that reflects real-world email traffic.
2. Providing a comparative analysis across key metrics, including accuracy, precision, recall, and F1-score, while highlighting the trade-offs between scalability and accuracy.
3. Integrating ethical considerations, such as the impact of false positives and the adaptability of models to evolving spam tactics, to provide a holistic evaluation.

## 3. METHODOLOGY

This section describes the methodology used to evaluate the performance and scalability of Naive Bayes, Random Forest, and SVM in email spam filtering. It includes details on dataset characteristics, preprocessing techniques, feature extraction methods, algorithm selection, experimental design, evaluation metrics, and the study workflow.

### 3.1 Dataset Description

The study utilised a publicly available dataset comprising 5,572 email messages, sourced from an established machine learning repository (NStugard, 2023). The dataset included both spam and legitimate emails (ham) with a balanced class distribution. Key characteristics of the dataset are:

1. **Size**: A total of 5,572 emails, including 3,359 ham and 2,213 spam messages.
2. **Diversity**: Emails covered a variety of topics, writing styles, and structures, reflecting real-world scenarios.

### 3.2 Data Preprocessing

The raw dataset underwent cleaning to make it suitable for learning:

1. **Text Normalisation**: Removed duplicates, irrelevant metadata (e.g., timestamps, headers), and HTML tag. Converted all text to lowercase and removed punctuation, special characters, and stopwords to simplify analysis.
2. **Tokenisation**: Split email content into individual words or tokens for feature extraction.
3. **Stemming and Lemmatisation**: Reduced words to their root forms to improve generalisability and minimise redundancy in the feature set.
4. **Feature Representation**:
    a. **Bag of Words (BoW)**: BoW was used for Naive Bayes to represent text as binary or frequency-based vectors. BoW is text representation technique that converts text into a numerical vector by counting the occurrence of each word in a document. While simple, BoW does not consider word order or context, making it effective for capturing word frequency but limited in capturing semantics.
    b. **Term Frequency-Inverse Document Frequency (TF-IDF)**: TF-IDF was used with Random Forest and SVM to assess word importance in relation to the dataset. Unlike BoW, TF-IDF measures a word's significance in a document compared to the entire dataset, assigning higher weights to frequent words that are rare in others, enhancing the capture of distinguishing features.

### 3.3 Workflow

The workflow for this study is summarised in figure 1, illustrating the key steps from dataset collection to performance evaluation. The process begins with sourcing and preprocessing a diverse dataset of emails, followed by feature extraction to prepare the data for machine learning models. The selected models—Naive Bayes, Random Forest, and SVM—are then trained and tested using cross-validation and hyperparameter tuning to optimise performance. Finally, the models are evaluated on key metrics, including accuracy, precision, recall, and F1-score, to assess their scalability and effectiveness.

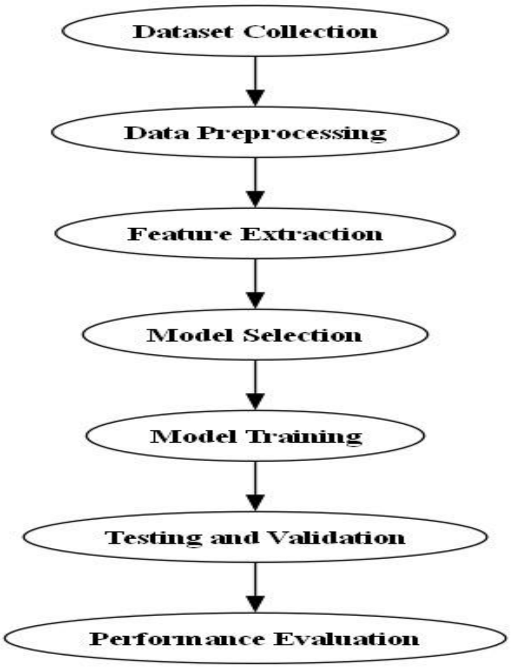

Figure 1: Workflow diagram illustrating the methodology for email spam filtering

### *3.4 Algorithm Selection and Justification*

The following machine learning models were selected for their complementary strengths in handling text classification tasks.

*Naive Bayes*

Naive Bayes simplicity makes it an excellent choice for real-time applications and systems with limited computational resources. However, while simplifying computations, the assumption of feature independence may oversimplify relationships in email text, potentially limiting its ability to capture nuanced patterns in spam.

*Random Forest*

Random Forest's robustness makes it effective for noisy and unbalanced datasets, such as those in spam filtering. Despite these advantages, the model's relatively higher computational requirements may present scalability challenges, particularly in resource-constrained or real-time environments.

*Support Vector Machines (SVM)*

SVM's ability to utilise kernel functions allows it to handle non-linear relationships, making it particularly suited for complex and high-dimensional datasets. However, the model requires significant computational resources and careful hyperparameter tuning, which may limit its applicability in large-scale or dynamic scenarios.

*Experimental Design*

The experimental design aimed to evaluate model performance under realistic conditions:

1. **Data Splitting**: The dataset was divided into training (75%) and testing (25%) subsets, ensuring a balanced distribution of spam and ham classes.
2. **Baseline Model**: A rule-based classifier was implemented as a baseline for comparison.

3. **Model Training**:
    a. Hyperparameter tuning was conducted using grid search to optimise model performance.
    b. Cross-validation ensured robust results by minimising overfitting.
4. **Testing**: Models were evaluated on the testing subset to assess their generalisability and performance on unseen data.

### 3.4 Evaluation Metrics

The models were evaluated based on four key metrics, with implications for their values described below:

1. **Accuracy**: Measures the proportion of correctly identifying spam. Higher accuracy indicates the overall reliability of the model, while lower accuracy could signify issues with generalisability or imbalanced training data.
2. **Precision**: Reflects the ratio of true positives (correctly identified spam) to all predicted positives (all spam classifications spam, either actual true or not). High precision is critical in reducing false positives, which can disrupt user experience, especially in environments with a low tolerance for misclassification (e.g., corporate email systems).
3. **Recall**: Indicates the ratio of correctly identified spam to all actual spam, highlighting the model's ability to detect spam. High recall is crucial in contexts where missing spam (false negatives) could have significant consequences, such as phishing detection.
4. **F1-Score**: Represents the harmonic mean of precision and recall, providing a balanced measure of performance. A high F1-score is particularly important when precision and recall are equally critical, ensuring the model maintains both detection capability and accuracy in classification.

### *3.5 Tools and Environment*

The implementation utilised the following tools and computing environment:

1. **Programming Language**: Python.
    a. Libraries:
        i. Scikit-learn for model implementation and evaluation.
        ii. Pandas and NumPy for data preprocessing and manipulation.
        iii. Matplotlib for visualisation.
2. **System Specifications**:
    a. Processor: Intel(R) Core (TM) i7.
    b. RAM: 8 GB.
    c. Operating System: Windows 10 (64-bit).

## 4. Results

This study compares the performance and scalability of Naive Bayes, Random Forest, and SVM in email spam filtering. Models were evaluated on accuracy, precision, recall, and F1-score using a diverse dataset. Scalability was assessed by analysing performance and training time as dataset size increased. Findings are presented with visual aids to showcase the strengths and limitations of each algorithm.

### 4.1 Performance Metrics

The models were evaluated on accuracy, precision, recall, and F1-score, with the results summarised in Table 1.

Table 1: Performance Metrics for Random Forest, Naive Bayes and SVM

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|---|
| Naive Bayes | 98.3 | 99.3 | 87.4 | 93.0 |
| Random Forest | 97.5 | 99.3 | 81.9 | 89.8 |
| SVM | 98.8 | 97.7 | 93.4 | 95.5 |

Figure 2 illustrates the precision, recall, and F1-score for all three models. As shown, SVM achieved the highest F1-score (95.5%), balancing precision (97.7%) and recall (93.4%). Naive Bayes and Random Forest both attained high precision (99.3%), effectively minimising false positives. However, Random Forest showed a lower F1-score (89.8%) due to its relatively lower recall (81.9%).

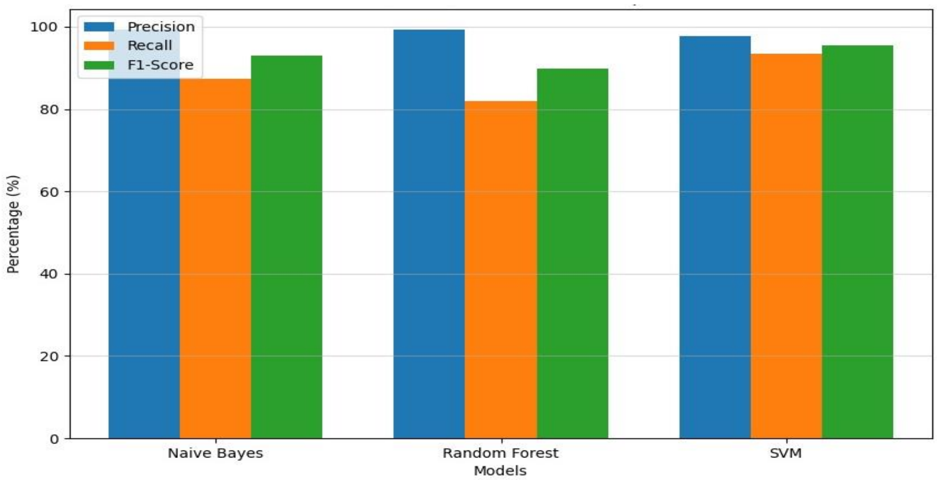

Figure 2: Comparing precision, recall, and F1-score for Naive Bayes, Random Forest, and SVM

### 4.2 Scalability Analysis

Scalability is a critical factor for spam filtering systems, as real-world applications often involve processing large and diverse datasets. This section examines how the performance and training time of Naive Bayes, Random Forest, and SVM are affected by increasing dataset sizes. Figures 3 and 4 provide visual representations of these trends, illustrating the trade-offs between computational efficiency and accuracy as the dataset grows.

### 4.3 Accuracy Trends

The accuracy trends across dataset sizes are presented in Figure 3. Naive Bayes maintained consistent accuracy across all dataset sizes, highlighting its computational efficiency and scalability. Conversely, Random Forest exhibited a slight decline in accuracy as dataset size increased, suggesting that it struggles with large-scale data. SVM consistently outperformed both models in terms of accuracy, achieving the highest value (98.8%) for the largest dataset, although its variability with smaller datasets highlights sensitivity to training size.

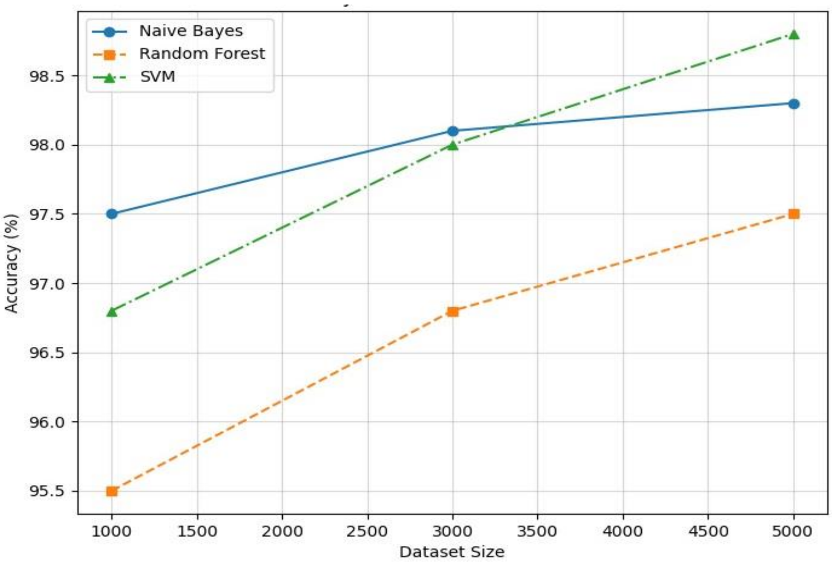

Figure 3: Line chart showing accuracy trends across dataset sizes for Naive Bayes, Random Forest, and SVM

### *4.4 Training Time Growth*

Training time growth is shown in Figure 4. Naive Bayes demonstrated minimal increases in training time as dataset size grew, making it the most computationally efficient model. Random Forest exhibited moderate growth, reflecting its ensemble-based complexity. While achieving the best overall performance, SVM displayed significant growth in training time as dataset size increased, indicating potential challenges for scalability in resource-constrained environments.

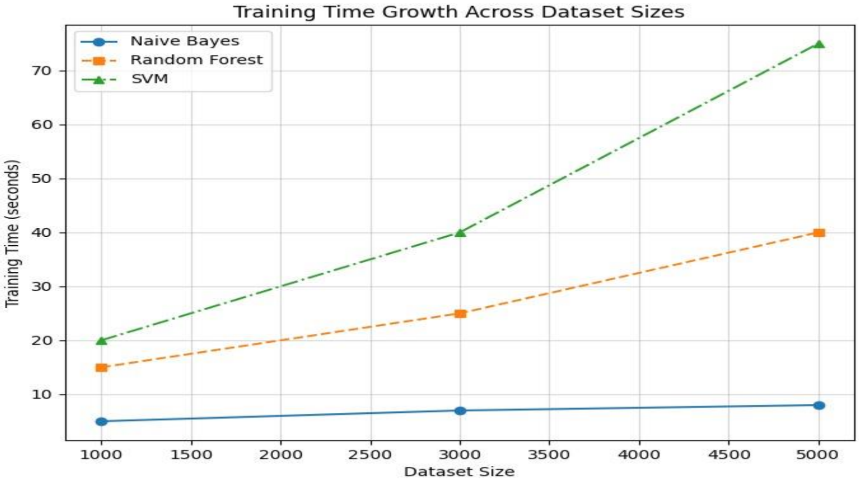

Figure 4: Line chart showing training time growth across dataset sizes for Naive Bayes, Random Forest, and SVM

### 4.5 Ethical and Practical Considerations

Spam filtering systems must balance accuracy with considerations such as user trust, privacy, and adaptability to evolving spam tactics. Ethical and practical factors play a crucial role in ensuring the effective deployment of these systems in real-world environments. This section discusses key considerations, including the importance of minimising false positives, ensuring adaptability to dynamic spam patterns, and addressing scalability challenges highlighted in Figures 3 and 4.

1. **False Positives**: High precision in Naive Bayes and Random Forest ensures minimal false positives, reducing user frustration and maintaining trust in email systems.
2. **Adaptability**: SVM's high recall makes it effective in security-critical environments, such as phishing detection, where missing spam could have serious consequences.
3. **Scalability**: The trends shown in Figures 3 and 4 highlight Naive Bayes 'suitability for real-time systems and Random Forest's moderate scalability. However, SVM's higher computational costs necessitate optimisation techniques, such as distributed computing, for large-scale applications.

## 5. DISCUSSION OF RESULTS

The discussion interprets the results of the comparative analysis, contextualises the findings within existing research, and explores their implications for email spam filtering. It highlights the key strengths and limitations of the models, their applicability in real-world scenarios, and potential challenges related to scalability and ethical considerations.

### 5.1 Interpretation of Results

The results demonstrate that each model offers distinct advantages and limitations, making them suitable for different operational contexts:

*Naive Bayes*

The Naive Bayes model maintained consistent accuracy across increasing dataset sizes, demonstrating its scalability and suitability for resource-constrained, real-time systems. Its high precision ensures minimal false positives, making it ideal for environments like corporate email systems. However, the feature independence assumption oversimplifies relationships in email content, limiting its ability to capture complex patterns. This reduction in recall may allow certain spam emails to bypass detection.

1. **Strengths**:
   a. Computationally efficient, suitable for real-time and resource-constrained applications.
   b. High precision reduces false positives.
   c. Scalable, maintaining consistent accuracy with larger datasets.
2. **Limitations**:
   a. Relies on the independence assumption, which limits its ability to capture nuanced relationships.
   b. Lower recall compared to other models, allowing some spam emails to evade detection.

Random Forest

The Random Forest model demonstrated strong robustness against noisy and unbalanced datasets, reflected in its high precision. These characteristics make it effective for applications where minimising false positives is critical. However, the model exhibited scalability issues, with growing computational demands as dataset size increased. Additionally, its lower recall than SVM highlights challenges in comprehensively identifying spam.

1. **Strengths**:
   a. Effective for noisy or unbalanced datasets.
   b. High precision minimises false positives, reducing user disruption.
2. **Limitations**:
   a. Scalability issues due to higher computational costs.
   b. Lower recall compared to SVM, reducing its ability to detect all spam.

### Support Vector Machines (SVM)

Support Vector Machines achieved the best overall performance in this study, excelling in recall and F1-score. These results suggest its suitability for high-security applications, where spam detection accuracy is paramount. Its ability to capture complex, non-linear relationships using kernel functions is a key advantage. However, the model's significant computational costs and long training times limit its scalability, particularly in resource-constrained environments or with vast datasets.

1. **Strengths**:
   a. Best overall performance, with the highest recall and F1-score.
   b. Captures non-linear relationships effectively using kernel functions.
2. **Limitations**:
   a. High computational cost and long training times.
   b. Scalability concerns for large-scale deployments.

### 5.2 Implications for Real-World Applications

These findings provide actionable insights for deploying machine learning models in email spam filtering systems:

1. **Naive Bayes**: Best suited for organisations prioritising speed and efficiency, such as small-scale or resource-constrained systems.
2. **Random Forest**: Ideal for use cases requiring high precision, such as financial or corporate environments where false positives can be particularly disruptive.
3. **SVM**: Recommended for high-security applications where recall is critical, provided sufficient computational resources are available.

### 5.3 Scalability Challenges

The trade-offs between scalability and accuracy remain a critical consideration for spam filtering systems.

1. **Naive Bayes**: Its lightweight nature makes it highly scalable and practical for large datasets.
2. **Random Forest**: While effective for noisy datasets, its computational requirements limit scalability, especially for real-time applications.
3. **SVM**: Though achieving superior accuracy and recall, its scalability is hindered by high computational demands. Techniques such as distributed computing or approximation algorithms may be necessary to enhance feasibility for large-scale deployments.

### 5.4 Ethical Considerations

Spam filtering systems must balance detection accuracy with fairness, privacy, and user experience. Ethical considerations ensure that deployed models align with organisational values and minimise unintended consequences. This study identifies key ethical issues that arise in spam filtering, particularly concerning false positives, adaptability, and deployment practices.

1. **False Positives**: Ensuring low false positive rates is critical for user trust. Naive Bayes and Random Forest demonstrated strong performance in this regard.
2. **Evolving Spam Patterns**: Continuous retraining and model updates are essential to adapt to new spam tactics. Transparent mechanisms must be established to ensure ethical alignment with privacy standards.
3. **Deployment Contexts**: Tailoring models to the operational environment can mitigate unintended consequences, such as over-filtering legitimate emails.

### 5.5 Comparison with Existing Literature

The study aligns with previous research emphasising the effectiveness of SVM for spam filtering (Amayri & Bouguila, 2010; Sculley & Wachman, 2007). However, it extends the discourse by:

1. Highlighting scalability as a critical factor for real-world deployment.
2. Demonstrating the impact of diverse datasets on model performance.
3. Addressing ethical considerations such as false positives and adaptability to dynamic spam trends.

## 6. Conclusion and RECOMMENDATIONS

The rapid growth of email traffic, coupled with increasingly sophisticated spam tactics, underscores the importance of scalable and accurate spam filtering systems. This study compared the performance of three machine learning models—Naive Bayes, Random Forest, and SVM—in addressing this challenge. Using a diverse dataset representative of real-world conditions, the models were evaluated on key metrics, including accuracy, precision, recall, and F1-score, as well as scalability.

The findings demonstrate that each model offers unique strengths and trade-offs:

1. **Naive Bayes**: Its computational efficiency and high precision make it ideal for real-time, resource-constrained systems. However, its reliance on the independence assumption limits its recall, reducing its ability to detect certain spam emails.
2. **Random Forest**: With high precision and robustness to noisy data, Random Forest is effective in scenarios where false positives must be minimised. Nevertheless, its scalability is constrained by computational demands and lower recall.
3. **SVM**: Achieving the best overall performance, particularly in recall and F1-score, SVM excels in detecting spam comprehensively. However, its high computational cost limits its applicability for large-scale or real-time systems without further optimisation.

These results highlight the trade-offs between scalability and accuracy, providing actionable insights for selecting machine learning models based on specific operational requirements. By addressing these considerations, this study contributes to developing adaptable and effective spam filtering systems.

The following recommendations stem from the findings:

1. **Algorithm Selection**:
    a. Use **Naive Bayes** for real-time applications requiring high speed and low computational resources.
    b. Opt for **Random Forest** in environments prioritising precision, such as financial or corporate email systems.
    c. Deploy **SVM** in high-security applications that demand balanced performance and comprehensive spam detection, provided adequate computational resources are available.
2. **Feature Engineering**: To improve model performance, enhance text representation by integrating advanced natural language processing (NLP) techniques, such as word embeddings or deep contextual representations.
3. **System Optimisation**: Explore hybrid approaches, such as ensemble methods, to leverage the complementary strengths of different algorithms and achieve optimal performance.
4. **Ethical Deployment**: Ensure ethical considerations, such as minimising false positives and maintaining user privacy, are prioritised in deploying spam filtering systems.

This study opens several avenues for further research:

1. **Exploration of Deep Learning Models**: Investigate the use of deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), for handling complex and dynamic spam patterns.
2. **Real-Time Spam Filtering**: Develop dynamic systems that adapt to evolving spam tactics in real time.
3. **Multilingual Datasets**: Incorporate multilingual datasets to improve model generalisability across diverse linguistic contexts.
4. **Ensemble Methods**: Explore ensemble techniques that combine the strengths of Naive Bayes, Random Forest, and SVM to achieve enhanced performance and scalability.

## REFERENCES

Amayri, O., & Bouguila, N. (2010). A study of spam filtering using support vector machines. Artificial Intelligence Review, 34(1), 73–108. https://doi.org/10.1007/s10462-010-9166-x

Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: Review, approaches and open research problems. Heliyon, 5(6), e01802. https://doi.org/10.1016/j.heliyon.2019.e01802

Faris, H., Aljarah, I., & Al-Shboul, B. (2016). A hybrid approach based on particle swarm optimization and random forests for e-mail spam filtering. In Computational Collective Intelligence (pp. 424–434). Springer. https://doi.org/10.1007/978-3-319-45243-2_46

Ghantasala, S., Gupta, S., & Gupta, R. (2020). Spam detection using naive Bayes classifier and feature selection. International Journal of Engineering Research and Technology, 9(6), 590–596. https://doi.org/10.17577/IJERTV9IS060457

Krishna, S. G., Gaddam, V. R., & Kumar, A. R. (2019). A comprehensive review on spam email detection using naive Bayes classifier. Journal of Computational and

Theoretical Nanoscience, 16(3), 1168–1173. https://doi.org/10.1166/jctn.2019.8210

Kontsewaye, Y., Antonov, E., & Artamonov, A. (2020). Evaluating the effectiveness of machine learning methods for spam detection. Procedia Computer Science, 190, 479–486. https://doi.org/10.1016/j.procs.2021.06.056

NStugard. (2023). Spam dataset. GitHub repository. https://github.com/NStugard/Intro-to-Machine-Learning/blob/main/spam.csv

Patan, A. K., Naresh, B., & Rao, S. S. (2020). A critical analysis of email spam detection methods based on machine learning algorithms. Journal of Emerging Technologies and Innovative Research, 7(6), 34–41.

Peng, W., Huang, L., Jia, J., & Ingram, E. (2018). Enhancing the naive Bayes spam filter through intelligent text modification detection. Proceedings of the 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 431–439. https://doi.org/10.1109/TrustCom/BigDataSE.2018

Reddy, A., Maheswari, M. U., Viswanathan, A., & Vikram, G. (2022). Using support vector machine for classification and feature extraction of spam. International Journal of Advanced Research in Computer Science, 13(2), 112–118.

Sculley, D., & Wachman, G. (2007). Relaxed online SVMs for spam filtering. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 415–422). https://doi.org/10.1145/1277741.1277817