Issues in Information Systems
Volume 25, Issue 2, pp. 203-213, 2024

DOI: https://doi.org/10.48009/2_iis_2024_116

Cyber-analytics: an examination of machine learning algorithms for spam filtering

Taiwo Ajani, *University of Charleston, taiwoajani@ucwv.edu*
Tammy Ferrante, *University of Charleston, tammyferrante@ucwv.edu*

Abstract

This study fills theoretical and application gaps by developing a hybrid spam filtering model that integrates the robustness of random forest classifier with the complex pattern recognition capabilities of neural networks, and the probabilistic reasoning of naïve bayes for enhanced data security and cyber-analytics. We reiterate the significance of spam filtering in addressing cybersecurity challenges and highlight the strengths and limitations of existing techniques; A case is made for the importance of robust spam filtering systems in combating the evolving threat of spam emails. Of the six prediction methods that were initially evaluated, the Random Forest (RF) classifier emerged as the most effective model, achieving the highest accuracy of 95.87% and the lowest misclassification error rate of 4.13%, with balanced performance in identifying true positives and true negatives. The hybridization of Random Forest, Neural Network and Naïve Bayes algorithms further improved accuracy to (97.22%).

Keywords: random forest classifier, spam filtering, support vector machine, decision trees, naïve bayes, neural network, cyber-analytics

Introduction

In the digital age, organizations are inundated with the pervasive issue of spam emails infiltrating their communication channels. Spam emails, also known as unsolicited bulk emails, present a multifaceted problem for businesses, ranging from inconveniences to significant security threats (Karim et. al., 2019). Any organization regardless of industry affiliation, can serve as a poignant example of an institution grappling with the incessant deluge of spam, a challenge mirrored across various sectors. These emails not only clutter inboxes but also pose severe risks, with many containing embedded malware aimed at compromising sensitive data or disrupting operations. The ramifications of spam extend beyond mere annoyance, translating into substantial financial burdens for organizations. Annually, businesses incur staggering costs associated with spam mitigation efforts, including expenditures on downtime, dedicated personnel, and employee training. These costs, measurable in millions of dollars, underscore the urgency of implementing effective spam detection measures. Despite existing mitigation strategies, the prevalence of spam emails persists, prompting a critical examination of current methodologies.

The enduring prevalence of spam emails despite existing mitigation strategies necessitates a reevaluation of current approaches. If current methods were entirely effective, the pervasive issue of spam would not persist, highlighting the inadequacies within the existing framework. It is evident that a one-size-fits-all solution may not suffice in combating the diverse array of spam emails inundating organizational networks. Variations in datasets, characterized by unique attributes and patterns, may significantly impact the efficacy

of existing spam detection systems. Addressing this pressing need for innovation, this research endeavors to explore and develop spam detection systems capable of adaptively addressing the idiosyncrasies of diverse datasets. In doing so, we aim to construct spam filters that exhibit unparalleled adaptability, effectively discerning between legitimate communications and spam emails across organizational landscapes.

Research Questions

This research seeks to address the following key questions to enhance spam detection efficacy:

RQ1: *Can we determine with high accuracy whether an email is spam or legitimate?*

RQ2: *Among various machine learning (ML) algorithms, which demonstrates superior performance in identifying spam emails?*

RQ3: *Can a hybridization of the best algorithms provide better accuracy measure in distinguishing spam emails?*

With the overarching objective of achieving a misclassification rate of less than 7.5%, this study aims to identify and select ML models capable of robustly identifying spam emails amidst a deluge of electronic communications.

Literature Review

Spam emails, also known as unsolicited bulk emails, continue to pose significant challenges for individuals and organizations alike. In response to the growing threat of spam, various spam filtering techniques have been developed and implemented. This literature review aims to critically evaluate different spam filtering techniques, including content-based filtering, case-based spam filtering, heuristic or rule-based spam filtering, previous likeness-based spam filtering, adaptive spam filtering, and machine learning (ML) algorithms for spam detection.

Content-Based Filtering Technique

Content-based filtering is one of the most common techniques used to identify spam emails based on the content of the message. This approach involves analyzing the text, images, links, and other elements within an email to determine its likelihood of being spam. Content-based filtering relies on predefined rules or keywords to classify emails as either spam or legitimate. While content-based filtering is straightforward and effective at detecting certain types of spam, it has limitations. For instance, this technique may struggle to detect spam emails that contain sophisticated obfuscation techniques or those with content that closely resembles legitimate messages. Additionally, content-based filtering may result in false positives, where legitimate emails are mistakenly classified as spam.

Case-Based Spam Filtering Method

Case-based spam filtering is a technique that leverages historical data or previously labeled spam emails to classify new incoming messages. This approach involves comparing the characteristics of incoming emails to a database of known spam or non-spam emails and making a classification based on similarity metrics. Case-based spam filtering offers the advantage of adaptability, as it can continuously learn and improve its

classification accuracy over time. However, this technique may require significant computational resources and storage capacity to maintain a comprehensive database of labeled emails.

Heuristic or Rule-Based Spam Filtering Technique

Heuristic or rule-based spam filtering relies on predefined rules or patterns to identify spam emails. These rules are often based on common characteristics of spam, such as specific keywords, sender information, or structural elements of the email. While heuristic filtering can be effective at capturing certain types of spam, it may struggle to adapt to new or evolving spam tactics. Additionally, heuristic rules may inadvertently flag legitimate emails that happen to meet the criteria set forth by the rules.

Previous Likeness-Based Spam Filtering Technique

Previous likeness-based spam filtering, also known as similarity-based filtering, compares incoming emails to previously seen spam or non-spam emails to make a classification decision. This technique uses similarity metrics to assess the resemblance between incoming messages and historical data.
Previous likeness-based filtering can be effective at identifying spam emails that exhibit similar characteristics to known spam patterns. However, it may struggle with detecting spam that deviates significantly from previously seen examples.

Adaptive Spam Filtering Technique

Adaptive spam filtering techniques aim to continuously learn and adapt to changing spam patterns and tactics. These approaches often incorporate machine learning algorithms to analyze large volumes of data and identify patterns indicative of spam. One of the key benefits of adaptive spam filtering is its ability to evolve and improve over time without manual intervention. By learning from new data and feedback, adaptive filtering systems can stay ahead of emerging spam threats. However, adaptive filtering may require substantial computational resources and training data to achieve optimal performance.

Machine Learning Algorithms for Spam Detection

Machine learning (ML) algorithms have gained popularity in spam detection due to their ability to automatically learn patterns and characteristics of spam emails. Common ML algorithms used for spam detection include Naive Bayes, Support Vector Machines (SVM), Decision Trees, and Neural Networks.
Each ML algorithm has its strengths and weaknesses in the context of spam detection. Naive Bayes, for example, is known for its simplicity and efficiency but may struggle with complex spam patterns. SVMs excel at handling high-dimensional data but may require extensive parameter tuning. According to Dada and his cohorts (2019), the SVM is one of the best methods for filtering emails. On the other hand, Decision Trees offer interpretability but may be prone to overfitting.

Despite their efficacy, ML algorithms for spam detection are not without limitations. These algorithms may require large amounts of labeled training data to achieve optimal performance, which can be challenging to obtain. Some authors observed that better performances can be obtained by employing boosting, ensemble or hybrid methods. In a study focused on anti-spam email filtering, Carreras and Marquiz (2001) introduced TreeBoost, an implementation of AdaBoost.MH, and evaluate its performance using the PU1 benchmark corpus, consisting of 1,099 messages, both spam and legitimate. The study shows that TreeBoost significantly outperforms traditional decision tree and Naive Bayes classifiers, achieving high accuracy and robustness. The study highlights the effectiveness of boosting in handling the complexities of spam filtering, reducing both false positives and false negatives, and demonstrates that deeper decision trees

require fewer boosting rounds to achieve optimal performance. This research provides valuable insights into the development of more efficient and accurate spam filtering systems.

Additionally, ML models may be susceptible to adversarial attacks or manipulation by spammers. While these methods each offer unique advantages, there are notable gaps that this study aims to address. First, existing methods often struggle with high false positive and false negative rates, impacting their reliability in practical applications. Second, the effectiveness of individual algorithms can vary significantly depending on the characteristics of the dataset, suggesting that a one-size-fits-all approach may not be optimal. Third, current spam filtering models may not fully leverage the potential of hybrid approaches that combine the strengths of multiple algorithms.

This study seeks to fill these gaps by examining several methods as well as a hybrid spam filtering model that integrates the better performing models. By comparing individual, ensemble or hybrid algorithms, this research aims to enhance overall accuracy, reduce misclassification errors, and achieve a balanced performance in identifying both spam and non-spam emails.

Additionally, the study will rigorously evaluate the performance of the hybrid model across various metrics to ensure its applicability in diverse data security and cyber-analytics contexts. Through empirical analysis and innovative model development, this study endeavors to advance the state of spam filtering technology, providing more effective solutions for combating the evolving threat of spam emails.

Methodology

The methodology employed in this study aims to classify emails as spam or not by using the SPAM dataset obtained from the UCI repository. The study falls within the subject areas of computer science, data security, and cyber-analytics, focusing on the task of classification.

The SPAM dataset used in this study was obtained from the UCI repository, a widely recognized source of machine learning datasets. The dataset comprises multivariate data, indicating that it contains multiple variables or features.

- **Subject Area:** The subject area of the dataset pertains to computer science, specifically focusing on data security and cyber-analytics.
- **Associated Tasks:** The primary task associated with the dataset is classification, where the objective is to classify emails as either spam or not spam. (see Figure 1)
- **Feature Type:** The features in the dataset consist of a combination of integer and real values. These features are used to represent various attributes of the emails, such as the presence of certain words, phrases, or structural characteristics.
- **Observations:** The dataset contains a total of 4601 observations or instances, each representing a single email.
- **Features:** There are 57 features in the dataset, representing different attributes or characteristics of the emails. These features serve as inputs to the classification model.

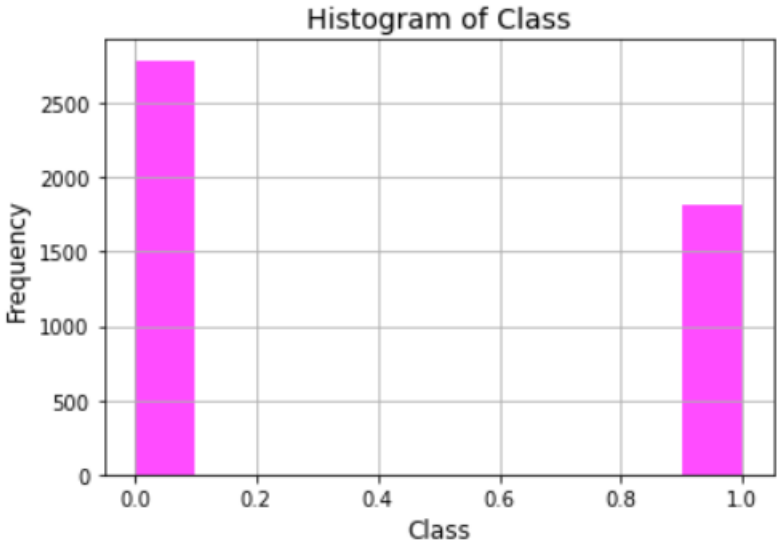

Figure 1: shows the frequency value of spam and non-spam emails in the dataset

Python programming language (Anaconda: Jupyter (Python v. 3.0)) was utilized for data preprocessing, model development, and performance evaluation due to its versatility and extensive libraries for machine learning. Univariate statistics were employed to explore the distributions of individual variables/features in the dataset. This includes measures such as mean, median, mode, variance, and standard deviation. Multivariate statistics, specifically correlation analysis (as shown in Figure 2), was conducted to determine relationships between the variables/features in the dataset. This helps identify potential multicollinearity issues and assess the overall structure of the data.

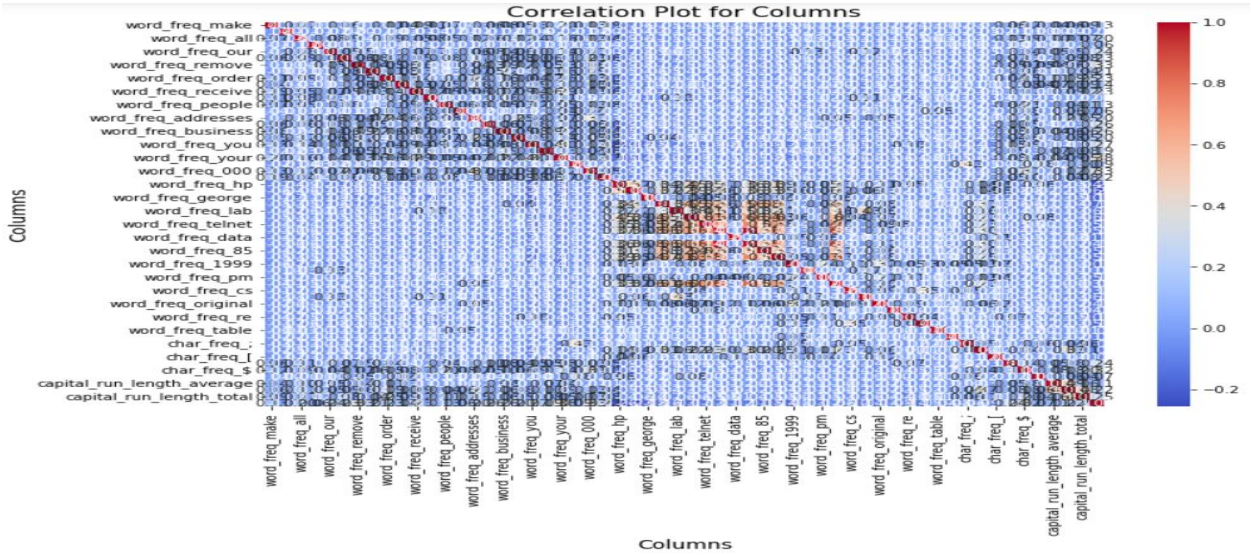

Figure 2: shows the correlation plot of the dataset features

The dataset was partitioned into training and testing sets to facilitate model development and evaluation. The training set was used to train the classification models, while the testing set was used to assess model performance on unseen data. Six different classification algorithms were employed for model development including: logistic regression (LR), decision trees (DT), random forest (RF), support vector machine (SVM), artificial neural networks (ANN), and naïve bayes (NB).

Subsequently, better performing algorithms were combined to develop a hybrid model. Each algorithm was trained on the training dataset using Python, and the resulting models were evaluated for their performance in classifying emails as spam or not spam using confusion matrices and various associated performance metrics, including accuracy, true negatives (TN), true positives (TP), false negatives (FN), and false positives (FP), etc.

Results

Table 1 shows the results of the difference algorithm employed. Based on Table 1, the Random Forest (RF) classifier achieved the highest accuracy of 95.87%, with the lowest misclassification error rate of 4.13%. It also recorded the highest number of true negatives (TN) and true positives (TP) compared to the other algorithms. Therefore, based solely on accuracy, the RF classifier appears to be the best model for spam filtering among the options provided.

Table 1: Model Comparison using various metrics

Algorithm	Accuracy (%)	Misclass Error (%)	TN	TP	FN	FP
Log Reg	92.62	7.38	501	352	38	30
DT	91.31	8.69	497	344	46	34
RF Classifier	95.87	4.13	523	360	30	8
SVM	66.23	33.77	448	162	228	83
Neural Net	94.14	5.86	514	353	37	17
Naïve Bayes	88.70	11.30	503	346	44	28

However, it is essential to critically evaluate the performance of each algorithm beyond just accuracy. Misclassification error, true positives, and true negatives are also crucial metrics to consider, as they provide insights into the model's ability to correctly classify spam and non-spam emails while minimizing false positives and false negatives.

Random Forest (RF) Classifier: The RF classifier demonstrated the highest accuracy, indicating that it correctly classified the highest proportion of emails as spam or non-spam compared to other algorithms. Additionally, it achieved the lowest misclassification error rate, suggesting robust performance in distinguishing between spam and non-spam emails. The high number of true negatives and true positives further confirms its effectiveness in correctly identifying both types of emails.

Support Vector Machine (SVM): The SVM algorithm achieved the lowest accuracy among the options provided, at 66.23%. This indicates that it had difficulty correctly classifying emails compared to other algorithms. The high misclassification error rate of 33.77% suggests that a significant portion of emails were incorrectly classified. While SVM recorded a relatively high number of true negatives, its performance in identifying true positives was notably lower, leading to a higher number of false negatives.

Decision Trees (DT) and Naïve Bayes: Both Decision Trees and Naïve Bayes algorithms performed reasonably well in terms of accuracy, with DT achieving an accuracy of 91.31% and Naïve Bayes achieving 88.70%. However, their misclassification error rates were relatively higher compared to RF and Neural Net. Additionally, both algorithms recorded fewer true negatives and true positives compared to RF and Neural Net, indicating slightly lower effectiveness in correctly classifying emails.

Neural Network (Neural Net): The Neural Net algorithm achieved an accuracy of 94.14%, slightly lower than RF but higher than DT and Naïve Bayes. Its misclassification error rate was also lower than DT and Naïve Bayes. However, while Neural Net recorded a high number of true positives, it had a slightly lower number of true negatives compared to RF. Nevertheless, it demonstrated strong performance overall.

Discussion

The analysis of the performance metrics for different classifiers in spam filtering highlights the importance of using multiple evaluation criteria to assess the effectiveness of machine learning models. The Random Forest (RF) classifier emerged as the most effective model based on the results, achieving the highest accuracy (95.87%) among the tested algorithms. This high accuracy, coupled with the lowest misclassification error rate (4.13%), indicates that the RF model not only successfully identifies the majority of spam and non-spam emails but also minimizes the errors in classification. Furthermore, its superior performance in recording the highest numbers of true negatives and true positives suggests robustness across both types of classification—correctly identifying spam (true positives) and rightfully dismissing non-spam (true negatives).

However, while accuracy and misclassification rates are critical, they do not fully encapsulate the potential drawbacks or strengths of a classifier. For instance, high accuracy might still be accompanied by unbalanced sensitivity and specificity, which could result in significant practical limitations, especially in scenarios where the cost of false positives is different from that of false negatives. For example, mistakenly identifying an important email as spam could be more problematic than failing to filter out actual spam.

The Support Vector Machine (SVM) algorithm, which showed the lowest accuracy (66.23%) and a high misclassification rate (33.77%), presents a stark contrast to RF. Despite a reasonable number of true negatives, the low true positive rate indicates a weakness in effectively identifying spam emails, potentially leading to a higher number of spam emails being overlooked. This could be due to the nature of the SVM's linear decision boundaries, which might be inadequate for the complexities and non-linearities present in the spam classification task.

Contrary to the literature suggestion above, this result does not support the opinion that the SVM is one of the best methods for spam filtering. The nature and specificity of the dataset could be at play here. Every data set is peculiar and must be analyzed for the best performing model as one approach may not necessarily be appropriate for all. However, this provides an opportunity for future study and discussion.

The performance of Decision Trees (DT) and Naïve Bayes also offers valuable insights. Both algorithms displayed decent accuracies of 91.31% and 88.70% respectively, but their higher misclassification rates compared to RF suggest less reliability in their classifications. Their lower counts of true positives and true negatives indicate that while effective to a certain extent, they may not consistently handle all types of emails effectively, leading to potential errors in both spam detection and non-spam identification.

Lastly, the Neural Network (Neural Net) algorithm presents a viable alternative to RF, with an accuracy close to RF (94.14%) and a competitive misclassification rate. Its performance was strong, especially in identifying true positives, suggesting a high sensitivity and an ability to effectively detect spam. However, the slightly lower number of true negatives compared to RF points to a possible area for improvement in specificity.

We considered the ramifications for hybridizing some of the high performing methods with respect to the metrics used. However, considering the high performance of most of the examined methods, and good balances between *true positive* and *true negative*, we saw minimal incentive or benefit for method hybridization. There is not much one can improve on a system that performs at about 96% without creating an additional consideration for overfitting.

For academic purposes, we hybridized the better performers. Although it offers interpretability, the DT was not included in the hybrid model because it may be prone to overfitting. Therefore, we included RF and Neural Nets (with respect to the good balance between *true positive* and *true negative*), as well as Naïve Bayes (88% performance and low *false positive*). The hybrid model achieved an accuracy of 97.22% relative to the best performing RF model (95.87%).

Precision, Recall, and F1-Score

Random Forest*:*

- Precision, recall, and F1-score were high for both classes (Class 0 and Class 1), indicating good performance.
- Precision for Class 1 is slightly higher (0.98) compared to the hybrid model's Class 1 precision (0.90), suggesting Random Forest is more precise in identifying true positives for Class 1.
- Recall for Class 1 is slightly lower (0.92) compared to the hybrid model's Class 1 recall (1.00), indicating the hybrid model is better at capturing all actual positives for Class 1.

Hybrid Model*:*

- Precision, recall, and F1-score were consistently high across all classes, showing balanced performance.
- The hybrid model achieves perfect precision and recall for several classes (e.g., Class 0, Class 4), indicating very high accuracy for these classes.

- The slight trade-off between precision and recall for certain classes (e.g., Class 1 with precision 0.90 and recall 1.00) shows the model's capability to balance different aspects of performance.

Class-specific Performance

Random Forest:

- Focuses on two classes with significant support (531 and 390), achieving high precision and recall.
- Useful for binary classification problems or scenarios with imbalanced datasets.

Hybrid Model:

- Evaluates performance across 10 classes with more distributed support, providing detailed insights into multi-class classification capabilities.
- High precision and recall across all classes indicate robustness and versatility for diverse datasets.

In fact, the confusion matrix shows that the hybrid model performs exceptionally well with a high accuracy of 97.22%. Most instances are classified correctly, and the misclassification rates are low across different classes. However, specific classes (like Class 1 and Class 9) have a few instances misclassified, which could be areas for further model improvement.

Conclusion

After critically evaluating the performance of each algorithm based on accuracy, misclassification error rate, true positives, and true negatives, the Random Forest (RF) classifier emerges as the best model for spam filtering. Its high accuracy, low misclassification error rate, and balanced performance in identifying true positives and true negatives make it a robust choice for effectively distinguishing between spam and non-spam emails.

The choice of classifier in practical applications should also consider the specific costs associated with false positives and false negatives. Furthermore, exploring ensemble methods or hybrid models combining the strengths of different classifiers, such as the sensitivity of Neural Nets and the specificity of RF, showed that we could potentially achieve more effective spam filtering solutions.

The proliferation of spam emails and the implication for cybersecurity, represents a formidable challenge for organizations worldwide. By interrogating the shortcomings of existing spam detection methodologies and leveraging advanced ML algorithms, this research explored innovative solutions capable of effectively mitigating the deleterious effects of spam. The hybrid model combining Random Forest, Neural Network, and Naïve Bayes outperforms the standalone Random Forest model in terms of overall accuracy and provides more balanced and detailed performance across multiple classes. This hybrid approach effectively leverages the strengths of each algorithm, making it a more robust and comprehensive solution for spam detection and other multi-class classification tasks.

Specifically, this study advanced theoretical understanding of analytic methods and their practical applications in the cybersecurity domain. They are detailed below:

Model Efficacy and Hybridization*:* The findings demonstrate that hybridizing different machine learning algorithms (Random Forest, Neural Networks, and Naïve Bayes) significantly improves classification accuracy. This supports theoretical propositions that ensemble methods can leverage the strengths of individual models while mitigating their weaknesses, leading to more robust and reliable spam detection systems (Dietterich, 2000).

Feature Interaction and Importance*:* By using diverse algorithms in a hybrid model, this study provides insights into how different features interact and contribute to spam detection. Understanding the varied importance of features across models helps refine theoretical models of feature selection and weighting in spam classification (Guyon & Elisseeff, 2003).

Algorithm Performance Metrics*:* The comprehensive evaluation of algorithms based on accuracy, *true positives, true negatives, false positives, and false negatives* enriches the theoretical discourse on the performance metrics that best reflect model effectiveness in cybersecurity applications (Provost & Fawcett, 2013).

Practical Applications in Cybersecurity

Enhanced Spam Filtering*:* The high accuracy and low misclassification rates of the hybrid model suggest practical improvements in spam filtering systems. Organizations can implement such models to reduce the incidence of spam emails, thereby enhancing security and efficiency in email communication.

Adaptive Security Measures*:* The demonstrated effectiveness of adaptive filtering techniques highlights the potential for real-time adaptation to emerging spam tactics. This adaptability is crucial for maintaining robust defenses against evolving cybersecurity threats (Dada et al., 2019).

Resource Optimization*:* By minimizing *false positives* and *false negatives*, the hybrid model reduces the burden on IT resources. This optimization allows cybersecurity teams to focus on more complex threats rather than dealing with large volumes of spam, thereby improving overall operational efficiency (Sculley & Wachman, 2007).

Scalability and Implementation*:* The use of widely recognized and accessible machine learning tools such as Python facilitates the scalability and practical implementation of the hybrid model across various organizational contexts. This accessibility promotes broader adoption and customization of advanced spam filtering techniques in the cybersecurity industry (Pedregosa *et. al.*, 2011).

In conclusion, through empirical analysis and rigorous evaluation, this study filled identified gaps by developing models including a hybrid spam filtering model that integrates the robustness of random forest, the complex pattern recognition capabilities of neural networks, and the probabilistic reasoning of naïve bayes.

In comparing different algorithms, this study provides theoretical, academic and business values by demonstrating enhanced overall accuracy, reduced misclassification errors, and a balanced performance in

identifying spam emails thereby paving the way for the development of adaptable spam detection systems capable of safeguarding organizational interests in an increasingly digital landscape.

References

Al-Ani, B. K. S., & Alomari, O. A. (2018). Heuristic approaches for spam filtering: A review. International Journal of Information Management, 42, 48-63.

Carreras, X., & Màrquez, L. (2001). Boosting trees for anti-spam email filtering. In Proceedings of the 4th International Conference on Recent Advances in Natural Language Processing (RANLP 2001) (pp. 58-64). Tzigov Chark, Bulgaria.

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.

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7(1), 3.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In J. Kittler & F. Roli (Eds.), Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science (Vol. 1857, pp. 1-15). Springer, Berlin, Heidelberg.

Fette, I., Sadeh, N., & Tomasic, A. (2007). Learning to detect phishing emails. In Proceedings of the 16th international conference on World Wide Web (pp. 649-656).

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3(Mar), 1157-1182.

Hafez, A., Rehan, M., Abdelhaq, M., & Salem, A. (2019). Spam detection in social networks: A review. Journal of Network and Computer Applications, 126, 59-76.

Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K., & Alazab, M. (2019). A comprehensive survey for intelligent spam email detection. IEEE Access, 7, 168261-168295.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big Data, 1(1), 51-59.

Sculley, D., & Wachman, G. M. (2007). Relaxed online SVMs for spam filtering. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 415-422.

UCI Machine Learning Repository. (n.d.). Spambase [Data set]. University of California, Irvine, School of Information and Computer Sciences. https://archive.ics.uci.edu/ml/datasets/spambase