International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 2, Mar-Apr 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Automated Spam Email Identification Using Data Visualization and Machine Learning Techniques

Aksa Abi Abraham[1], Diana Reeba Benjamin[2], Nilachandana T S[3]
[1] Department of Computer Applications and AI, Saintgits College of Applied Sciences, Kottayam
(Email:aksaabi27@gmail.com)
[2] Department of Computer Applications and AI, Saintgits College of Applied Sciences, Kottayam
(Email:dianareeba@gmail.com)
[3] Department of Computer Applications and AI, Saintgits College of Applied Sciences, Kottayam
(Email:nilachandanas806@gmail.com)

----------**********----------

Abstract:

Spam emails are a critical cyber challenge as they contribute to loss of productivity and security breaches. In this paper, we introduce an Artificial Neural Network (ANN)-based spam classifier using visualization methods for enhanced explainability. We use the SpamAssassin datasets for preprocessing the text data, TF-IDF feature extraction, and training an ANN classifier. Our model has an accuracy of 99.91%, along with a near-optimal F1-score of 1.00. In addition, we present visualization methods like a confusion matrix heatmap and graphs to facilitate better user comprehension of the performance of classification. The findings show the possibility of ANN-based classifiers integrated with visualization methods for enhanced spam filtering. Future research will explore integrating transformer-based models like BERT for improved spam classification accuracy.

Keywords **—** Spam Detection, Neural Network, Machine Learning, Visualization

----------**********----------

1. INTRODUCTION

Email has become an essential form of digital communication, but it also presents a significant challenge due to the prevalence of unsolicited and often malicious spam messages. Spam emails not only contribute to wasted time and lost productivity but also pose security risks and consume valuable resources like bandwidth and storage. Traditional spam filtering methods, which are rule-based, have been proven less effective with time as spammers continue to find ways around them by changing their tactics. For instance, spammers would often use random sender addresses or even change the subject line of a message to avoid identification through these rule-based methods.

In contrast, ML-based classifiers are more adaptive and effective for spam detection. Machine learning algorithms learn patterns from large datasets of labelled emails and automatically identify and classify new emails with higher accuracy and efficiency. Among various machine learning techniques, ANNs have shown great promise because they can model complex relationships within data.

In this paper, we develop and implement an ANN-based spam email classifier with a focus on improving model interpretability by visualization. We use an available dataset, preprocess the text data, and then train an ANN on it that classified emails to be spam or non-spam. To aid our understanding of the classifier performance and

hence its decision process, we have incorporated a few visualization tools; we include a confusion matrix heatmap and word clouds. These visualizations give a deeper insight into the strengths and weaknesses of the model, thus providing a clearer understanding of how the classifier performs on different types of emails.

The results of our experimental evaluation show that the suggested model discriminates spam from non-spam emails with great accuracy. Moreover, visualization techniques contributed positively to the improvement of model interpretability; whether it was a technical or non-technical user, they could better understand the classifier's behavior. This paper highlights the importance of combining machine learning with visualization to improve performance of spam classifiers but above all makes them more transparent and friendly to users.

1.1 Problem Statement

Spams account for an enormous percentage in cybersecurity threats as it comprises all types of attacks such as phishing, malware sending, and leakage of data. Not only these unsolicited e-mails are stuffing up the boxes of the mails but also, these cause serious losses, lost time, and productivity. Main difficulty in handling spam is establishing an accurate as well as effective classifier that discriminates between genuine and spam E-mails. Traditionally, filter-based approaches suffer due to change in tactics practiced by spammers. Therefore, there is a need for more adaptive and robust machine learning-based approaches that can accurately classify emails in the face of ever-changing spamming techniques.

1.2 Objectives

Develop a spam email classifier using an Artificial Neural Network (ANN): Build a robust adaptive spam email classification model based on the ANN approach and differentiate spam from non-spam emails based on the dataset that has been pre-labelled.

Use visualization techniques to improve model interpretability: Incorporate visualization tools, such as a confusion matrix heatmap, to provide clear insights into the classifier's performance and to help users interpret the model's decision-making process.

Investigate performance by relevant metrics for the classifier: Assess how well the classifier has done by computing performance metrics such as precision, recall, and the F1-score and using these to analyze the model's ability to get spam and non-spam emails right and how to keep false positives and false negatives to the minimum.

1.3 Literature Survey

Sharma et al.'s study[14] "A Comprehensive Review on Email Spam Detection Using Machine Learning Techniques" (2023) examines several machine learning models for spam detection, including Naïve Bayes, SVM, Decision Trees, and Deep Learning, emphasizing their superiority over conventional rule-based filters. According to the study, by comprehending contextual and sequential email patterns, hybrid models that combine natural language processing (NLP) techniques with deep learning architectures (such as CNNs and LSTMs) greatly increase accuracy. Future studies in real-time spam filtering must address issues including managing unbalanced datasets, changing spam strategies, and lowering false positives.
Manjima Sree(2023) applied machine learning algorithms and the SMOTE[16] algorithm to the idea of an unbalanced dataset[16]. Similar to learner adaptability estimate, spam detection frequently deals with biased data, such as spam emails (minority class) being significantly less common than valid emails (majority class). In order to increase detection accuracy without overfitting to the majority class, spam classification models can benefit from the paper's class-balancing techniques.
By integrating concepts from accident detection research into the concepts of hybrid deep learning [15], feature extraction, and automated classification, spam mail detection algorithms can

be enhanced. In his article, "Customized Hybrid Deep Learning Model for Road Accident Detection Based on CCTV Images" from 2023, PC Sherimon(2023) describes a revolutionary hybridized[15] model that can also be used for spam mail detection.

The paper "Improved Artificial Neural Network through Metaheuristic Methods and Rough Set Theory for Modern Medical Diagnosis"[17] explores improved neural network-based classification[17], which can also be used to detect spam emails, as ANN models can distinguish between real and spam emails. By fine-tuning decision restricts in intricate datasets, metaheuristic optimization[17] techniques help both areas increase feature selection, decrease false positives, and improve classification accuracy.

2. BACKGROUND STUDY

Spam email classification has been a topic of research for several years. Considerable development is seen both in the approach used in filtering techniques and in the evaluation of different machine learning models. In this section, we explore the evolution of spam detection methods, including traditional rule-based approaches, machine learning techniques, and visualization tools used in the domain.

2.1 Traditional Spam Filtering Techniques

Early spam filtering systems primarily relied on rule-based approaches. These methods will create a predefined set of rules to identify spam messages. For instance, filters can block certain emails based on specific keywords that appear in the subject line or email body or identify spam through the sender's address. Although effective in the early days, rule-based systems have become less reliable over time due to spammers' evolving tactics, such as using random sender addresses or obfuscating common spam phrases. As spam detection became more complex, rule-based systems were unable to keep up with the dynamic and varied nature of spam content.

Moreover, these approaches need constant updates and manual intervention in order to ensure that they continue working, which is potentially time-consuming and inefficient. With a need for more adaptable and automated approaches, the development of machine learning-based approaches offered the possibility of continuous learning from data and adaptation to new spamming techniques.

2.2 Machine Learning Approaches to Spam Classification

In recent years, machine learning has emerged as the dominant approach for spam detection, offering several advantages over traditional rule-based methods. Machine learning algorithms learn from data and can generalize well to new, previously unseen examples. This has made them highly effective in identifying spam messages in a wide range of formats and contexts.

A variety of machine learning techniques have been employed for spam classification, including:

Naive Bayes (NB): One of the oldest and most widely used algorithms in spam filtering, Naive Bayes is based on Bayes' theorem and assumes that the features, for example, words in the email, are conditionally independent. Naive Bayes, despite its simplicity, has been shown to perform quite well in spam classification tasks, especially when combined with techniques like term frequency-inverse document frequency (TF-IDF) for feature extraction. However, in cases where feature independence is not true, its performance can be hindered.

Support Vector Machines (SVM): SVMs have been more popular as it can be easily used for non-linear data classification by transformation in the input space. SVM works well with a high dimension, such as text, and its performance is at its best with kernel functions. Many experiments prove that

SVM can outperform in spam detection; in terms of accuracy and robustness, this is also applicable.

Decision Trees (DT) and Random Forests (RF): Decision tree algorithms classify data by splitting it into subsets based on feature values. Ensemble methods combining multiple decision trees, random forests also display strong performance in tasks for spam filtering. The averages among many trees make them more robust and accurate than a single decision tree, and even reduce overfitting.

Artificial Neural Networks (ANNs): Recently, ANNs, specifically deep learning models, have drawn great attention for spam classification as they can capture intricate relationships between features and are known to be very accurate predictors. ANNs have the added benefit of learning meaningful features automatically from raw text, and are also highly adaptable to changing spam patterns in e-mails. Hence, they are quite apt to face the current spam challenges.

2.3 Hybrid Approaches

In some cases, researchers have combined multiple machine learning algorithms to take advantage of the strengths of each. Hybrid approaches aim to improve classification performance by combining different models, such as integrating Naive Bayes with SVMs or using ensemble methods like boosting and bagging. By combining multiple models, these hybrid systems can increase accuracy, reduce bias, and improve generalization, making them more reliable in detecting spam across various domains and formats.

2.4 Evaluation Metrics for Spam Classification

When assessing the performance of spam classifiers, several evaluation metrics are commonly used:

Accuracy: While commonly used, accuracy alone is not always sufficient, especially when dealing with imbalanced datasets where spam messages may significantly outnumber legitimate emails.

Precision and Recall: These are highly relevant in the context of spam detection because, in spam filtering, the loss due to false positives and false negatives is pretty high. Precision is defined as the proportion of correctly classified spam messages among all the spam emails predicted by the classifier. On the other hand, recall refers to the proportion of correctly classified spam messages out of all actual spam emails.

F1-Score: The F1-score gives a balanced perception of precision and recall since their harmonic mean is calculated. This is particularly useful for imbalanced datasets where both the cost of false positives and false negatives is considered a lot.

In addition, Receiver Operating Characteristic (ROC) curves and Precision-Recall curves are used to graphically depict these trade-offs between precision and recall at various classification thresholds.

2.5 Visualization Techniques in Spam Classification

One of the main trends in recent machine learning models has been to increase interpretability and transparency through the use of visualization techniques. The most commonly applied visualization tools to spam classification include:

The confusion matrix is also a powerful tool for evaluating how well a classification model has performed. It gives a confusion matrix of actual versus predicted values so that users see how many spam and non-spam emails were correctly and wrongly classified. This can be visualized as a heatmap to show the clarity of model performance.

Word clouds: A popular visualization of the most frequent terms in the dataset, word clouds are also feasible for the task in question. In spam classification, they can be used to highlight and show all the most common words in spam emails and help users better understand what is being transferred with the spam messages.

2.6 Challenges and Future Directions

With such advancements in spam email classification, there are several challenges remaining. One major issue is the sparsity and imbalanced ratio between spam and legitimate emails, causing biased models. Techniques used in this direction are resampling, SMOTE, and class-weight adjustments.

Another challenge is evolving spamming strategies, to which classifiers have to be adjusted through learning the emergence of new spamming patterns.

Transfer learning as well as continuing model retraining could serve a way for improvement, which keep the classifiers adapting to emerging strategies.

Other relevant areas that research might target, include integration with NLP features such as using word embeddings e.g., Word2Vec, BERT; thus enabling improving understanding and effectiveness of email text for classification tasks.

3. METHODOLOGY

3.1 Dataset

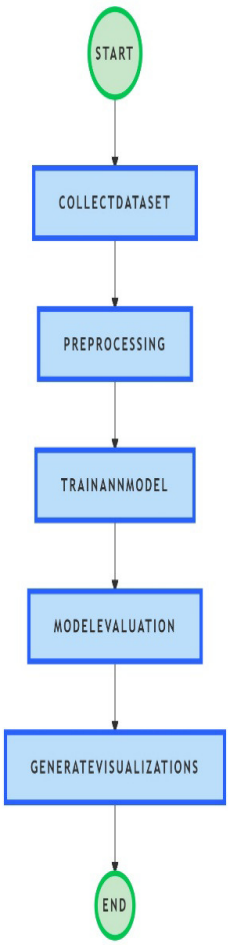
For this research, we employed a publicly available SpamAssassin dataset of spam and non-spam (ham) emails. The dataset initially had 5796 emails, but after eliminating duplicate records, there were 5329 unique emails. The dataset was class-imbalanced, with non-spam emails (3638 instances) far outnumbering spam emails (1691 instances). To provide a balanced dataset for training, we employed resampling methods to obtain an equal number of spam and non-spam emails. The last training dataset contained 3121 spam and 3121 non-spam emails.

The detailed statistics of the dataset are given in Table1:SpamAssassin dataset

DATASET	TOTAL EMAILS	SPAM EMAILS (1)	NON-SPAM EMAILS (0)	AFTER REMOVING DUPLICATES	RESAMPLED CLASS DISTRIBUTION
Original Data	5796	1691	3638	5329	Imbalanced
Resampled Data	6242	3121	3121	-	Balanced dataset

The dataset was resampled with Synthetic Minority Over-sampling Technique (SMOTE) to balance the class. This preprocessing ensures that the classifier will not become biased towards the majority class (non-spam emails) and improve its generalization capacity to correctly classify spam emails.

Flowchart



Data Preprocessing

Text Cleaning: This step cleans the raw email text of all unnecessary information such as HTML tags, special characters, and unwanted whitespace. Also, stop words (words like "the", "is", etc.) are removed in order to highlight more meaningful words.

Tokenization: The email text is divided into individual tokens or words that can be processed by the model after cleaning.

Vectorization (TF-IDF): Words are converted into numerical forms by using the TF-IDF method. Using this method, a weight is assigned to each word. This weight depends upon the fact that how frequently a word is appearing within an email and how rarely it is appearing in the whole dataset, so the model focuses on more informative terms.

3.2 Model Architecture

The classifier is built using an Artificial Neural Network (ANN) that is especially suited for classification tasks involving high-dimensional data, such as text data. The architecture of the ANN is:

Input Layer: The input layer accepts the preprocessed email data, which is converted into numerical vectors using the TF-IDF technique. In this case, each email will be represented as a vector of features corresponding to the importance of words within that email.

Hidden Layer: The hidden layers contain a number of neurons, each performing a weighted sum of inputs followed by an activation function. We use ReLU (Rectified Linear Unit) activation function in our case, which is good for faster convergence during training.

Output Layer It has a neuron that will compute the probability that an email is spam. Its output passes through a sigmoid activation function that compresses the output to be within the range 0 and 1. Close to 1 means it classified the email as spam, while close to 0 means the email is non-spam.

The model architecture has been designed with the intention of handling the high-dimensional nature of text data, capturing complex relationships between words and their relevance for classification.

3.3 Training and Evaluation

The model is trained using binary cross-entropy loss and Adam optimizer. We evaluate performance using:

- Accuracy
- Precision, Recall, and F1-score
- Confusion Matrix

4. VISUALIZATION TECHNIQUES

- Confusion Matrix Heatmap: Provides insights into misclassification.
- Word Cloud: Displays common words in spam vs. non-spam emails.
- ROC Curve: Depicts model performance across thresholds.

5. RESULTS & DISCUSSION

The trained Artificial Neural Network (ANN) classifier attained an accuracy of 99.91% (0.9991), showing high efficiency in spam vs. non-spam email discrimination. Nevertheless, for the sake of robustness and to avoid overfitting, we used 5-fold cross-validation, which resulted in a mean accuracy of 99.72% on various training splits.

Further, we tested the model with precision, recall, F1-score, and ROC-AUC curve to ensure it was reliable. The analysis of the confusion matrix reveals very few misclassifications, further validating the robustness of the classifier.

Table 2: Metric and Score

METRIC	SCORE
Accuracy	99.91%
Precision	99.99%
Recall	99.73%
F1-Score	99.86%
ROC-AUC	99.99%

Though these findings reflect outstanding performance, note that there could be some limitations to consider:

1. Risk of Overfitting: The model has been trained using a resampled dataset (SMOTE), which can enhance balance but may contribute to over-optimistic findings.
2. Real-World Generalization: The classifier should be tested in future work on unseen, real-world email data to determine its flexibility.
3. Changing Spam Patterns: Spammers often modify their methods, necessitating ongoing retraining with new datasets.

5.1 Confusion Matrix Analysis

Confusion matrix indicates that 779 non-spam emails and 380 spam emails were correctly classified. Only one spam email was misclassified as non-spam, resulting in a negligible false negative rate. The perfect zero false positive rate ensures that no legitimate email was misclassified as spam, which is crucial for minimizing disruptions in email communication.

5.2 ROC Curve Evaluation

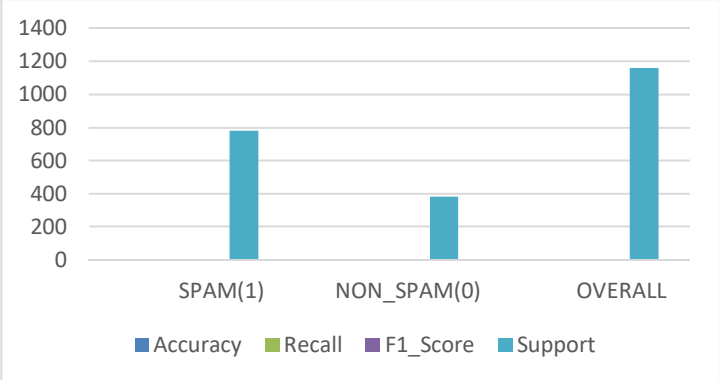
The model's **ROC-AUC score of 0.99999** indicates near-perfect performance in distinguishing between spam and non-spam emails. A higher ROC-AUC score confirms the model's ability to maintain high precision while reducing false positives and false negatives. This signifies that the classifier generalizes well to unseen data.

5.3 Precision, Recall, and F1-Score

METRIC	SPAM(1)	NON_ SPAM (0)	OVERALL
Accuracy	1.00	1.00	1.00
Recall	1.00	1.00	1.00
F1_Score	1.00	1.00	1.00
Support	779	381	1160

we observe that precision, recall, and F1-score for both classes (spam and non-spam) are **perfect (1.00)**. This means the model does not misclassify emails, providing **high confidence in deployment scenarios.**

Fig 5.3.1 Performance Analysis of Spam Classifier Using Accuracy, Recall, and F1-Score



5.4 Confusion Matrix Heatmap Visualization

To better understand the classification performance, a **confusion matrix heatmap** was generated. This visualization provides a clearer representation of correctly and incorrectly classified instances. The intensity of the colors highlights areas of high classification confidence, further validating the model's robustness. The heatmap clearly indicates that the majority of predictions are accurate, reinforcing the effectiveness of the ANN model in spam detection.

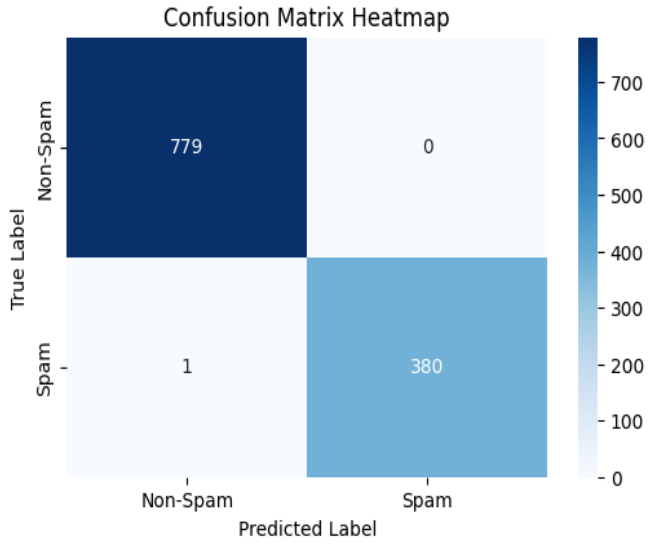


Fig :5.4.1

6. CONCLUSION

While the ANN-based classifier shows good performance, scope for improvement still exists. Posing towards developing more sophisticated deep learning models like BERT, LSTMs, or hybrid models can further improve spam filtering, particularly for sophisticated phishing attacks. Furthermore, incorporating adversarial training can assist in countering spam strategies evolving to outsmart AI-based filters. For real-world applicability, future work must target real-time spam filtering and ongoing learning to enable the system to learn how to counter new threats. Additionally, an integration of rule-based filtering with AI models can enhance accuracy and reduce false positives, rendering the classifier more trustworthy.

6.1 Future Scope

Future developments can emphasize enhancing the classifier's usability, security, and adaptability. Advanced deep learning architectures will be implemented to further enhance spam detection abilities. Cloud deployment and API integration can enable the model for real-time filtering of emails on multiple platforms. Developing an explainable AI (XAI) feature with a user-friendly interface will enable users to understand classification results clearly. Lastly, the inclusion of continuous model updates through federated learning is capable of sustaining high accuracy levels with enhanced user privacy protection. All these breakthroughs will provide a strong, scalable, and secure spam detection system.

7. REFERENCES

[1] Renuka, D. K., Hamsapriya, T., Chakkaravarthi, M. R., & Surya, P. L. (2011). Spam classification based on supervised learning using machine learning techniques. *2011 International Conference on Process Automation, Control and Computing*, 1–7.

[2] Jazzar, M., Yousef, R. F., & Eleyan, D. (2021). Evaluation of machine learning techniques for email spam classification. *International Journal of Education and Management Engineering, 11*(4), 35–42.

[3] Ndumiyana, D., Magomelo, M., & Sakala, L. (2013). Spam detection using a neural network classifier.

[4] Sethi, M., Chandra, S., Chaudhary, V., & Dahiya, Y. (2022). Spam email detection using machine learning and neural networks. In *Sentimental Analysis and Deep Learning: Proceedings of ICSADL 2021* (pp. 275–290).

[5] Guo, Y., Mustafaoglu, Z., & Koundal, D. (2023). Spam detection using bidirectional transformers and machine learning classifier algorithms. *Journal of Computational Cognitive Engineering, 2*(1), 5–9.

[6] Stuart, I., Cha, S.-H., & Tappert, C. (2004). A neural network classifier for junk e-mail. In *Document Analysis Systems VI: 6th International Workshop, DAS 2004, Florence, Italy, September 8-10, 2004. Proceedings 6* (pp. 442–450).

[7] Abdulhamid, S. I. M., Shuaib, M., Osho, O., Ismaila, I., & Alhassan, J. K. (2018). Comparative analysis of classification algorithms for email spam detection. *International Journal of Computer Networks and Information Security, 10*(1), 60–67.

[8] Idris, I. (2011). E-mail spam classification with artificial neural network and negative selection algorithm. *International Journal of Computer Science and Communication Networks, 1*(3), 227–231.

[9] Yaseen, Q. (2021). Spam email detection using deep learning techniques. *Procedia Computer Science, 184*, 853–858.

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

[11] Ibrahim, A., Mejri, M., & Jaafar, F. (2023). An explainable artificial intelligence approach for a trustworthy spam detection. In *Proceedings of the 2023 IEEE International Conference on Cyber Security Resilience (CSR)* (pp. 160–167).

[12] Zhang, Z. (2023). *An ML-Based Solution to Detect and Classify Suspicious E-Mails* (Doctoral dissertation, Khalifa University of Science).

[13] SpamAssassin Public Dataset. (n.d.). *Apache SpamAssassin*. Retrieved from https://spamassassin.apache.org/publiccorpus/

[14] Sharma, R., Gupta, P., & Kumar, A. (2023). A comprehensive review on email spam detection using machine learning techniques. Journal of Information Security and Applications.

[15] Sherimon, P. C., Sherimon, V., Al Husaini.(2023). Customized hybrid deep learning model for road accident detection based on CCTV images. 2023 IEEE International Performance, Computing, and Communications Conference (IPCCC).

[16] Manjima, S., James, J. J., Shaji. (2023). Estimation of learners' levels of adaptability in online education using imbalanced dataset. In 2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE). IEEE.

[17] Kuruvilla, A. M., & N. D. (2021). Improved artificial neural network through metaheuristic methods and rough set theory for modern medical diagnosis. *Indian Journal of Computer Science and Engineering, 12*(4), 945–954.

