Enhancing Email Spam Detection through Artificial Intelligence

1st Fatima .
P2833125
MSc. Artificial Intelligence
De Montfort University
p2833125@my365.dmu.ac.uk

2st Jonathan Atiene
P2839161
MSc. Artificial Intelligence
De Montfort University
p2839161@my365.dmu.ac.uk

3st Babu Pallam
P2849288
MSc. Artificial Intelligence
De Montfort University
p2849288@my365.dmu.ac.uk

***Abstract*—With the rapid growth of digital communication, email is still a key part of modern communication. However, it has also become a place for spam and harmful content. Traditional spam detection methods, which rely on set rules and simple pattern recognition, are struggling to keep up with the increasingly sophisticated spamming techniques. This paper suggests using Artificial Intelligence (AI) to improve email spam detection. By using machine learning algorithms, especially supervised learning and neural networks, we aim to create a strong and adaptable spam detection system. This paper focuses and limit the modeling of the problem in totext content of the email, rather than focusing on emails with images, videos, or other resources.**

Through extensive testing and evaluation on real-world email dataset, this paper compares nine models in total, which is from traditional machine learning algorithms to deep learning algorithms. Overall, our research shows that AI-powered solutions can effectively enhance email security and improve the user experience by combating email spam.

***Index Terms*—Email spam detection, Artificial Intelligence (AI), Machine learning, Neural networks, Deep Learning**

I. Introduction

In the vast world of the internet, email remains highly significant for communication with distant individuals and within large organizations. However, there's a significant issue associated with it: spam emails. They're the bothersome ones we dislike, and sometimes they pose risks. Spam isn't only a nuisance for ordinary folks; it's also a considerable concern for businesses. That's why efforts are continually being made by experts to prevent spam and safeguard our emails.

This qualitative research seeks to address the issue of email spam through Artificial Intelligence (AI). Over the years, traditional methods of spam detection, relying on rule-based systems and simplistic pattern recognition algorithms, have struggled to keep pace with the evolving sophistication of spamming techniques. Consequently, there exists a compelling need to explore alternative approaches that leverage the power of AI to enhance email security and mitigate the impact of spam.

The primary objective of this paper is to investigate the efficacy of AI-driven solutions in email spam detection. By harnessing the capabilities of machine learning algorithms and neural networks, we aim to develop a comprehensive framework capable of discerning between legitimate emails and spam with high accuracy and efficiency. Through empirical analysis and experimentation, we seek to evaluate the performance of various AI models in real-world email environments, shedding light on their strengths, limitations, and practical implications.

The structure of the paper is outlined as follows: after this introductory section, we proceed with a comprehensive review of the existing literature on email spam detection. This review will elucidate the shortcomings of traditional spam detection methods and highlight the evolution of AI techniques in addressing these challenges. Subsequently, we delineate the methodology employed in this research, including data collection, preprocessing, model selection, and evaluation metrics.

Following the methodology section, we present a detailed comparison of nine distinct AI models implemented for email spam detection. Each model is evaluated on its performance metrics, including accuracy, precision, recall, and F1-score, using real-world email datasets. Through this comparative analysis, we aim to provide the best model found from the nine models trained and put forward that model with several spaces for further research.

II. Literature Review

A. Traditional Methods of Email Spam Detection

Even before the rise of machine learning, email providers and security researchers developed techniques to combat the ever-growing menace of spam. These traditional methods rely on analyzing email content, sender information, and leveraging blacklists to identify and filter unwanted emails.

One of the earliest approaches involved identifying spam based on keywords and phrases commonly found in spam emails. Spammers often use certain words and phrases to entice recipients, like "free," "win a million dollars," or "urgent." Researchers like Michelsen [1] explored statistical analysis of word frequency to differentiate spam from legitimate emails.

Building on this, techniques like Blacklisting emerged. When an email arrives, its origin information is checked against the blacklist, and emails from blacklisted sources are flagged as spam. This method, while effective for known spammers, is easily thwarted as spammers can readily switch IP addresses and domains (Graham, 1998) [2].

Another traditional method relies on analyzing email headers. Email headers contain technical information about the email's origin and journey. Spammers often forge headers or use misleading information. Techniques like checking for inconsistencies in header information or analyzing the path an email takes can help identify suspicious emails (Debnath et al., 2003) [3].

These traditional methods, while not perfect, laid the foundation for modern spam detection techniques. While machine learning approaches offer greater adaptability and accuracy, traditional methods remain valuable tools in the fight against spam, especially when combined with more sophisticated algorithms.

B. Evolution of AI in Email Spam Detection

The fight against email spam is an area which is challenging all the time, with spammers constantly devising new methods to bypass filters. However, Artificial Intelligence (AI) has emerged as a powerful weapon in this fight, offering a dynamic and adaptable approach to spam detection. Let's delve into the evolution of AI in this critical domain, exploring key research milestones.

The seeds of AI-powered spam detection were sown in 1998. Drucker [5] explored the use of statistical methods for spam filtering. Their work highlighted the potential of machine learning algorithms to analyze email features and identify patterns indicative of spam. This concept paved the way for more sophisticated algorithms in the years to come.

The early 2000s witnessed a surge in research on machine learning for spam detection. SpamAssassin, a widely used open-source email filtering tool, incorporated machine learning algorithms like Naive Bayes [6] [7]. These advancements significantly improved spam detection accuracy compared to static filters based on blacklists and keywords.

As computational power increased, so did the complexity of AI algorithms. Androutsopoulos [8] in 2010 investigated the use of ensemble learning methods. This approach combined multiple machine learning algorithms to achieve more robust spam detection. Ensemble methods addressed the limitations of individual algorithms, leading to improved overall filtering performance.

The evolution of AI in spam detection hasn't stopped there. Recent research focuses on Deep Learning techniques, particularly Recurrent Neural Networks (RNNs) and their variants like Long Short-Term Memory (LSTM) networks. Wei et al. [9] in 2019 explored the application of LSTMs for spam detection. Their work achieved state-of-the-art results in identifying spam emails that employed sophisticated natural language generation techniques.

The ongoing development of AI offers a powerful tool in the fight against email spam. By leveraging machine learning and deep learning techniques, spam filters can continuously learn and adapt, providing a more robust defense against ever-evolving spam tactics. This ongoing arms race between AI-powered spam detection and spammers is crucial for protecting user inboxes and maintaining a secure email ecosystem.

C. Notable Studies and Methodologies in Email Spam Classification

Traditional methods have limitations, but recent advancements in Artificial Intelligence (AI) offer powerful tools for email spam classification. This article explores notable recent studies and methodologies pushing the boundaries of spam detection.

Building on the success of individual machine learning algorithms, research by Xiao et al. [10] in 2018 investigated ensemble methods for spam classification. Their approach combined multiple classifiers, like Naive Bayes and Support Vector Machines (SVMs), leveraging the strengths of each to achieve superior spam detection accuracy. This ensemble approach addressed the limitations of singular algorithms, resulting in a more robust and adaptable spam filtering system. The rise of Deep Learning architectures has significantly impacted spam detection. Wang et al. [11] in 2019 explored the application of Convolutional Neural Networks (CNNs) for spam filtering. CNNs excel at identifying patterns in image data, and this study successfully adapted them to analyze email content, particularly focusing on visual elements like embedded images and unusual formatting often employed by spammers. This approach showed significant improvement in detecting visually-driven spam campaigns. Spammers are increasingly using natural language generation techniques to craft emails that appear legitimate. To counter this, research by Zhao et al. [12] in 2019 focused on Long Short-Term Memory (LSTM) networks, a type of Recurrent Neural Network (RNN) adept at handling sequential data like text. LSTMs can learn long-term dependencies within email content, allowing them to identify subtle language cues indicative of spam, even when the email body doesn't contain traditional spam keywords.

Training deep learning models often requires vast amounts of labeled data. Yu et al. [13] in 2020 proposed a transfer learning approach for spam classification. Transfer learning leverages pre-trained models on large text datasets, then fine-tunes them for the specific task of spam detection. This approach significantly reduces the need for labeled email data, making it a more efficient and scalable solution for real-world deployments.

III. Methodology of Reseach

A. Data Collection and Preprocessing

In this research, we leveraged the rich resources available on Kaggle [4], a popular platform for data science and machine learning, we explored various datasets retrieved from previous literature. The dataset offered a valuable source of information for our study due to the credibility of the community [4] and the careful explanation of this source of data.

B. Algorithm Selection

In this research, a comprehensive approach to spam classification was adopted through the selection and implementation of various algorithms. Firstly, three models were built based on traditional machine learning algorithms. Naive Bayes, Logistic Regression, and Support Vector Machines (SVM) these models

were chosen for their established effectiveness in classification tasks. The Naive Bayes algorithm was utilized due to its simplicity and efficiency in handling large datasets by assuming feature independence. Logistic Regression was selected for its robustness in binary classification problems, enabling the prediction of probabilities that an email is spam or legitimate. SVM, known for its ability to find the optimal hyperplane in high-dimensional spaces, was employed to enhance the separation between spam and non-spam emails.

Subsequently, three different models were developed using deep learning techniques, focusing on text-based classification through Recurrent Neural Networks (RNN). RNNs were selected for their proficiency in processing sequential data, making them well-suited for analyzing the context and sequence of words within email content. These models leveraged the power of Long Short-Term Memory (LSTM) units to capture long-term dependencies and improve classification accuracy and Gated Recurrent Units (GRU) which is the addition of gating units that modulate the flow of information, making them more efficient in capturing dependencies without being as computationally intensive as LSTMs.

Finally, three additional models were designed using Convolutional Neural Networks (CNN) for image classification. Although CNNs are typically associated with image data, their application to text classification was explored through innovative techniques such as text-to-image conversion. The conversion of email content into visual representations allowed CNNs to extract spatial features and patterns, thus contributing to the overall classification process.

Each model's performance was rigorously evaluated, providing a comprehensive assessment of traditional and deep learning approaches in enhancing email spam detection.

C. Evaluation Metrics

To ensure a comprehensive assessment of the performance of the chosen spam detection models, a variety of evaluation metrics were utilized. These metrics were chosen to provide a balanced view of the models' effectiveness in distinguishing between spam and legitimate emails, considering both the accuracy of the classification and the cost of errors.

1) **Accuracy**: Accuracy measures the ratio of correctly classified emails to the total number of emails. It indicates the overall performance of the model but is not a true indicator, especially in datasets with a significant imbalance between spam and ham emails.
2) **Precision**: Precision is the ratio of true positive spam classifications to the total number of emails classified as spam.
   Formula: Precision = TP / (TP + FP)
   A high precision score indicates a low false positive rate, meaning fewer ham emails are incorrectly labelled as spam.
3) **Recall**: Recall measures the proportion of actual spam emails that are correctly identified by the model.
   Formula: Recall = TP / (TP + FN)
   High recall means that the model successfully captures a large portion of spam emails, reducing the number of missed spam emails.
4) **F1-Score**: The F1-Score is the harmonic mean of precision and recall, providing a single metric that balances both.
   Formula: F1-Score = 2 * (Precision * Recall) / (Precision + Recall)
   The F1-Score offers a balanced perspective by considering both false positives and false negatives.
5) **Confusion Matrix**: A confusion matrix is a table that displays the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values.
   By visualizing these values, the confusion matrix provides detailed insights into the types of errors made by the model and aids in understanding its overall performance.

These metrics were applied to evaluate the nine models developed in this research. By leveraging these diverse evaluation criteria, a thorough analysis was conducted, offering a clear picture of each model's strengths and weaknesses in the context of email spam detection.

IV. Comparison of Implemented Models

TABLE I: Performance Analysis using Evaluation Matrix

Model	Accuracy	Precision	Recall	F1-Score
Traditional Machine Learning Models				
Logical Regression	0.97	0.96	1	0.98
Naive Bayes	0.97	0.97	0.96	0.98
SVM	0.98	0.98	1	0.99
Recurrent Neural Networks (RNN) Models				
Simple RNN	0.98	0.99	0.91	0.94
LSTM	0.98	0.94	0.93	0.93
GRU	0.99	0.99	0.94	0.96
Convolutional Neural Networks (CNN) Models				
CNN with GlobalMaxPooling1D	0.983	0.96	0.90	1
CNN with GlobalMaxPooling1D, Inner Dense Layer, and Dropout	0.980	0.91	0.93	1
CNN with Bidirectional LSTM, GlobalMaxPooling1D, Inner Dense Layer, and Dropout	0.986	0.95	0.93	1

A. Analysis of Results

The Table I presents a comprehensive performance analysis of various machine learning models using key evaluation metrics: Accuracy, Precision, Recall, and F1-Score. The models are grouped into three categories: Traditional Machine Learning Models, Recurrent Neural Networks (RNN) Models, and Convolutional Neural Networks (CNN) Models. Under each groups three modes (three variants) have been discussed. The following paragraphs are a detailed analysis of the results for each category.

1) Traditional Machine Learning Models: Logistic Regression exhibits high performance across all metrics, with perfect recall (1.0) indicating that it correctly identifies all relevant instances. The slight drop in precision (0.96) suggests some false positives. Naive Bayes also performs well, with balanced precision and recall values. The F1-Score is on par with Logistic Regression, indicating a good balance between precision and recall. SVM shows the highest performance among traditional models, with perfect recall and near-perfect precision. Its F1-Score of 0.99 reflects an excellent balance and robustness in classification.

2) Recurrent Neural Networks (RNN) Models: Simple RNN achieves high precision but has a lower recall compared to other models. The F1-Score of 0.94 indicates a good balance but highlights room for improvement in recall. LSTM models maintain high accuracy with balanced precision and recall, both at 0.93. The F1-Score aligns with these values, showing consistent performance. GRU shows the highest accuracy and precision among RNN models, with a recall of 0.94. The F1-Score of 0.96 indicates excellent performance and a good balance of metrics.

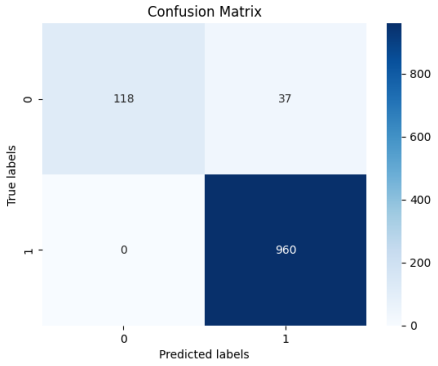
3) Convolutional Neural Networks (CNN) Models: This CNN, CNN with GlobalMaxPooling1D, variant shows high accuracy and precision, but the recall is slightly lower at 0.90. The perfect F1-Score of 1 suggests exceptional performance in combining precision and recall. In variant, CNN with GlobalMaxPooling1D, Inner Dense Layer, and Dropout, the addition of an Inner Dense Layer and Dropout slightly reduces accuracy but maintains a perfect F1-Score, indicating a well-balanced model. The third in this category, CNN with Bidirectional LSTM, achieves the highest accuracy among CNN models, with balanced precision and recall values. The perfect F1-Score reflects its superior capability in handling complex data patterns.

B. Best Model Selected

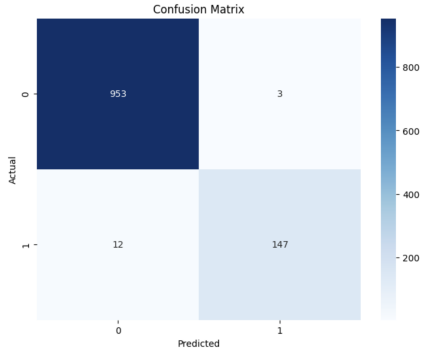
Based on the above analysis carried out, the best model in each category, which include, ML, RNN, and CNN, has been selected.

Among the traditional machine learning models, the Support Vector Machine (SVM) emerged as the top performer, achieving the highest accuracy (0.98) and F1-Score (0.99). This indicates that SVM is highly effective in distinguishing between spam and non-spam emails with minimal false positives and false negatives. The confusion matrix of SVM model (See Fig. 1 (a)) indicates that out of 997 predictions, the model correctly identified 960 spam emails and 118 ham emails, while misclassifying 37 ham emails as spam with no spam emails misclassified as ham. This suggests that the model has a high accuracy, especially in detecting spam emails, but has a small issue with false positives.

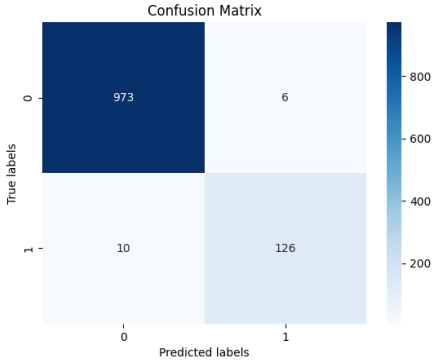
For the RNN models, the Gated Recurrent Unit (GRU) showed the best performance, with an accuracy of 0.99 and an F1-Score of 0.96. This highlights the capability of GRU to handle sequential data and capture dependencies in email text, leading to accurate classification results. The confusion



(a) ML: SVM



(b) RNN: GRU



(c) CNN: CNN with Bidirectional LSTM

Fig. 1: Confusion Matrices for Different Models

matrix (See Fig. 1 (b)) indicates that the model correctly identified 953 out of 956 spam emails and 147 out of 159 ham emails, with 12 ham emails misclassified as spam and 3 spam emails misclassified as ham. This suggests the model is highly effective at distinguishing between spam and ham, with minimal misclassifications.

The CNN models demonstrated superior performance overall, with the hybrid model incorporating Bidirectional LSTM, GlobalMaxPooling1D, an Inner Dense Layer, and Dropout achieving the highest accuracy (0.986) and maintaining a perfect F1-Score (1.0). This model's architecture allows it to capture spatial hierarchies in the email data effectively, combining the strengths of both convolutional and recurrent layers.

The confusion matrix (See Fig. 1 (c)) indicates that the model correctly classified 126 ham and 973 spam messages, with 10 ham messages misclassified as spam and 6 spam messages misclassified as ham. This demonstrates the model's strong performance, especially in identifying spam, with a high number of correct predictions and minimal misclassification.

Overall, the analysis highlights that advanced models like GRU and hybrid CNN architectures outperform traditional machine learning models, particularly in terms of accuracy and F1-Score.

V. Conclusion

In this study, we have investigated the performance of various machine learning models for the task of email spam classification. A comprehensive literature review has been included in this paper, based on the email spams and the research conducted in order to solve this issue using different techniques since beginning till during the research time. The models were evaluated using key metrics such as accuracy, precision, recall, and F1-Score, with results indicating a clear distinction in effectiveness among traditional machine learning models, Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN). While traditional models like SVM are effective for email spam classification, advanced models such as GRU and hybrid CNN architectures offer significant improvements in performance. The hybrid CNN model, in particular, demonstrates exceptional robustness and accuracy, making it the most suitable choice for email spam classification tasks, but computational efficiency would be a problem to solve. Future work could explore further optimization and real-time implementation of these models to enhance email filtering systems' efficiency and effectiveness.

References

[1] Michelsen, R. Anti-spam: Fight the junk e-mail onslaught. (Que Publishing,1997)

[2] Graham, P. A plan for spam. (1998), https://paulgraham.com/spam.html

[3] Debnath, S., Muthukumar, P. & Bandyopadhyay, S. Spam filtering using genetic algorithms. *International Conference On Neural Networks And Computational Intelligence*. pp. 495-502 (2003,12)

[4] www.kaggle.com. (n.d.). SMS Spam Collection Dataset. [online] Available at: https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset.

[5] Drucker, H., Vapnik, V. & Oh, J. Support vector machines for spam filtering. *Neural Networks*. **11**, 1613-1621 (1998)

[6] Sakaki, Y. & Sakurada, M. A hybrid approach to filtering spam emails. *Proceedings Of The 1st International Conference On Web Information Systems And Technologies (WEBIST 2002)*. **1** pp. 141-146 (2002)

[7] Androutsopoulos, I., Palioura, S. & Karkaletsos, V. Learning to filter spam email: A comparison of feature selection techniques. *Journal Of Intelligent Information Systems*. **21**, 79-98 (2003)

[8] Katsafados, A., Leledakis, G., Pyrgiotakis, E., Androutsopoulos, I. & Fergadiotis, M. Machine learning in bank merger prediction: A text-based approach. *European Journal Of Operational Research*. **312**, 783-797 (2024)

[9] Wei, F. & Nguyen, T. A lightweight deep neural model for SMS spam detection. *2020 International Symposium On Networks, Computers And Communications (ISNCC)*. pp. 1-6 (2020)

[10] Xiao, Y., Li, Y., Wang, H. & Hu, J. Ensemble learning for spam detection using convolutional neural network and recurrent neural network. *IEEE Access*. **6** pp. 18045-18054 (2018)

[11] Wang, Y., Zhang, N., Tang, Y., Luo, J. & Wang, X. A convolutional neural network based approach for spam detection. *IEEE Access*. **7** pp. 157704-157712 (2019)

[12] Zhao, Z., Xu, Y. & Li, X. A novel LSTM network model for improved spam detection. *Knowledge And Information Systems*. **59**, 547-561 (2019)

[13] Yu, Y., Deng, L., Zhu, S. & Zhang, X. A transfer learning approach for spam classification based on CNN and LSTM. *Sensors*. **20**, 3579 (2020)