International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 07 July 2024
Page No: 1882- 1890
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0258

Email Spam Detection Based on Exceptional Precision

Madhav Aggarwal[1]*, Manik Thakur*[2]*, Sahil Nagpal*[3]*, Anup Singh Kushwaha*[4]

[1,2,3,4]*Department of Computer Science, MRU, Faridabad-Haryana, India*

Email ID: *madhavaggarwal286@gmail.com*[1]*, thakurmanikst10@gmail.com*[2]*, snagpal2003@gmail.com*[3]*, anup@mru.edu.in*[4]

Abstract

Email remains a crucial means of communication in personal and professional spheres; however, its efficiency is often compromised by the widespread presence of unwanted messages. The increase in unsolicited emails not only inundates email inboxes but also poses significant threats such as phishing, malware distribution, and financial fraud. To tackle these issues and enhance the effectiveness of email exchanges, there has been a notable emphasis on utilizing machine learning techniques for identifying spam. This paper will explore various machine learning algorithms and apply them to our datasets. The most optimal algorithm will be selected for email spam detection based on its exceptional precision and accuracy.

Keywords: *Phishing; Machine Learning; Spam Detection*

1. Introduction

Most individuals opt for utilizing the email system in their daily activities to engage in communication with others. However, the presence of spam messages poses a significant threat to both individuals and society at large. Essentially, spam messages entail the transmission of unwanted messages to the inboxes of specific target users. This results in the inundation of users' inboxes with undesirable content, causing annoyance to the recipient. Additionally, there is a risk that target users may inadvertently succumb to the schemes orchestrated by the attacker. Consequently, the dissemination of spam messages represents a clear menace to email users and the broader Internet community. Furthermore, upon reading the entire content of a spam that has been sent to their inbox, users may realize its nature and subsequently choose to disregard it. As a result, users end up squandering their valuable time. Hence, it is imperative to implement measures aimed at either preventing or filtering out spam messages from legitimate correspondence. Machine learning techniques have shown enhanced efficacy through the utilization of a dataset comprising pre-categorized email examples for training purposes. Various algorithms of machine learning can be utilized within machine learning approaches for the purpose of email filtering. There are two separate classifications of misleading review spam that occur within the domain of manipulating online consumer feedback. The first category is known as hyper spam, where fabricated positive reviews are strategically assigned to products in order to enhance their visibility and promote sales. This form of spam is designed to mislead potential customers by dishonestly boosting the status of a product with deceptive endorsements. Conversely, the second category, defaming spam, involves the dissemination of unjustifiably negative reviews targeting rival products with the intention of undermining their credibility among consumers. Deceptive practices have the potential to negatively impact the reputation and market performance of rival brands. Positive deceptive review spam are characterized by reviews that are crafted to elevate the image or reputation of a specific brand or product, thereby conveying a favorable sentiment towards the item in question. On the other hand, negative deceptive review spam are manifested through critiques that are designed to tarnish the image or position of a competing product, expressing a disparaging viewpoint towards the target product in an effort to sway consumer perceptions. Figure 1 explains the spam detection using ML.

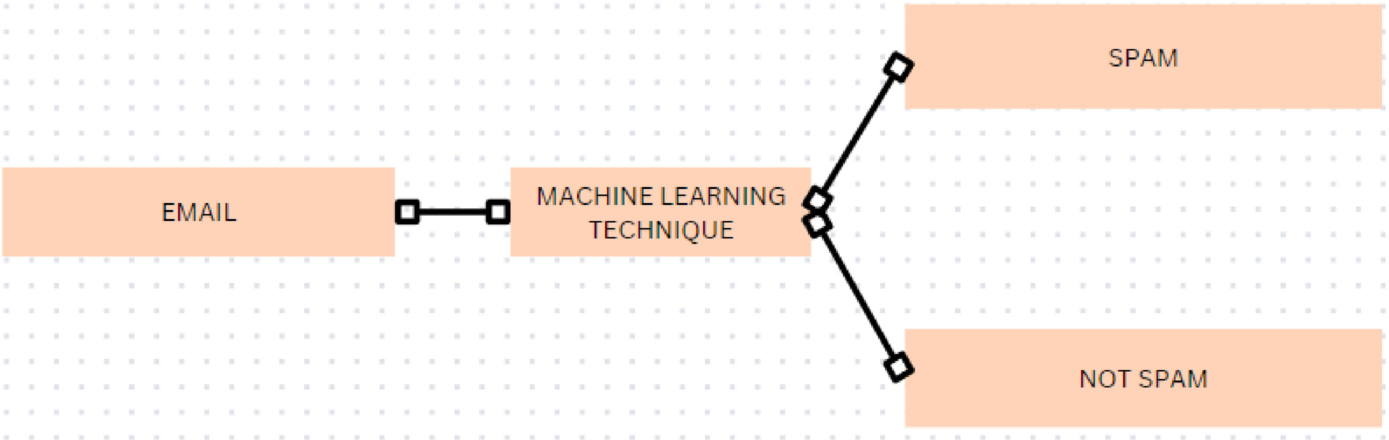

Figure 1 **Spam Detection using ML**

2. Literature Review

A) Email [1] spam has emerged as a significant issue in contemporary times, propelled by the rapid expansion of internet users. The prevalence of email spam is on the rise, with individuals resorting to them for illicit and unethical activities such as phishing and fraud. Malicious links are being distributed through spam emails, posing a threat to our systems and infiltrating them. The simplicity with which spammers can create fake profiles and email accounts allows them to masquerade as genuine individuals in their spam emails, targeting unsuspecting individuals who are unaware of such fraudulent schemes. Consequently, there is a pressing need to distinguish fraudulent spam emails. This research project aims to identify such spam emails through the application of machine learning techniques. The article explores different machine learning algorithms by applying them to datasets in order to identify the best efficient algorithm for recognizing email spam with the highest level of precision and accuracy. The findings suggest that Multinomial Naïve Bayes yields the best results, notwithstanding its limitation in misclassifying some tuples due to class-conditional independence. Ensemble techniques, conversely, have demonstrated their advantages through the utilization of a variety of classifiers to make predictions on class labels. Given the vast volume of emails exchanged daily, our project's testing capability is constrained by a limited corpus. The spam detection system focuses on filtering emails based on content rather than domain names or other criteria, thereby presenting room for enhancement. Potential improvements include filtering spam based on trusted domain names, emphasizing the significance of classifying spam emails to differentiate them from legitimate ones. This approach could be adopted by organizations to discern desired emails from unwanted ones.

B) Communication [2] via email in the current Internet age has gained popularity due to its cost-effectiveness and ease of use for sending messages and sharing crucial information. However, the presence of spam messages floods users' inboxes with unwanted content, leading to wastage of resources and users' valuable time. Thus, an efficient and precise method is essential to differentiate between spam and legitimate messages. This research presents an innovative framework for identifying spam through the analysis of email body text sentiment. The framework combines Word Embedding’s with a Bidirectional LSTM network to scrutinize both emotional and sequential dimensions of the text. Additionally, to expedite training and extract advanced text features for the Bi-LSTM network, a Convolutional Neural Network is employed. Two datasets, namely ling spam dataset and spam text message classification dataset, are utilized, with recall, precision, and f-score serving as metrics to assess the efficacy of the proposed approach. The model attains an enhanced accuracy performance of approximately 98-99%. Notably, it outperforms various machine learning classifiers and cutting-edge methods in spam detection, demonstrating its superior capability. This research introduces a novel methodology for detecting spam or legitimate emails, focusing on textual analysis. The model comprises three distinct networks - Word Embedding's, CNN, and Bi-LSTM. By incorporating a Convolutional layer post Word Embedding and prior to the LSTM network, the training process is expedited, and higher-level features are extracted for the Bidirectional LSTM network. The adoption of Bidirectional LSTM enhances the model's accuracy by effectively capturing contextual meaning and sequential patterns in sentences, resulting in an improved accuracy performance of around 98-99%.

C)In the current era, machine learning algorithms are being proficiently employed for the automated detection of spam emails. This research investigates the examination of different prevalent machine learning methods and clarifies their efficacy in recognizing spam emails. Electronic mail, or email, stands out as a prevalent form of communication due to its ease of access, facilitation of swift message exchanges, and minimal transmission expenses. Email is esteemed as the swiftest and most cost-effective means of correspondence. The issue of spam in emails poses a significant challenge for email platforms. Spam emails are essentially unsolicited messages sent for promotional, deceptive, or other purposes without targeting a specific recipient. Various detection and filtration methods are employed to manage the influx of spam. Notably, the KNN algorithm, known for its content-based approach, emerges as one of the most pragmatic and uncomplicated techniques. To classify folders and subjects in the research, a substantial collection of personal emails was initially utilised. The KNN algorithm was subsequently enhanced and broadened to I-KNN to enhance its efficiency. The refined algorithm was then applied holistically to yield superior results in Email Spam Detection. Implementation was carried out using Java, focusing on parameters such as calculation time and similarity score. Results indicate that the proposed approach attains higher efficiency compared to conventional methods. While technology offers numerous advantages, email usage also presents challenges such as junk mail, commonly referred to as spam, which can be sent to any email address by third parties. This research recommends five machine learning methods for anti-spam filtering, thereby providing an overview of spam filtering techniques. Although the term "spam" lacks a precise definition, it generally refers to unsolicited communications causing ethical and economic dilemmas. Efforts have been made to define and prohibit spam through legislation. The KNN-classifier based filtering method is strongly recommended and widely utilized for anti-spam purposes. Various innovative techniques are currently employed to categorize emails based on diverse criteria across different regions. These techniques often involve the automated processing of incoming messages. Additionally, human intervention in both current and archived messages is also encompassed within this concept. The conventional N-gram strategy can potentially be substituted with the proposed efficient approach. Utilizing the recommended method on actual-time data sets may prove beneficial for future studies. Furthermore, Intrusion Detection System (IDS) techniques can enhance strategic approaches. Utilizing the recommended method on actual-time data sets may prove beneficial for future studies. Furthermore, Intrusion Detection System (IDS) techniques can enhance strategic approaches.

D) Email [4] spam remains a significant issue, particularly in the realm of the Internet. It involves harmful malwares that target users' devices to steal data, damage the device, and deceive users into purchasing their products. Despite the development of spam detection and email filtering systems, there is a growing number of spam-containing emails. This study aims to identify valuable email header features for spam detection through the analysis of two email datasets, namely the Anomaly Detection Challenges and Cyber Security Data Mining datasets. The primary goal is to extract suitable email header features and assess them using Support Vector Machine (SVM) in tools like Rapid Miner Studio and Weka 3.9.2. The research methodology comprises five phases: Data Collection, Data Pre-Processing, Feature Selection, Classification, and Detection. The classification of email headers using SVM for CSDM2010 dataset outperforms the Anomaly Detection Challenges dataset, achieving accuracy rates of 88.80% and 87.20% respectively. Thus, SVM emerges as an effective classifier with an accuracy rate exceeding 80% for both datasets. Several experiments were conducted to validate an email spam detection framework using header emails, revealing that Support Vector Machine (SVM) yielded the best performance with two datasets. The extracted features, including a list of ten features, were found to be suitable for the framework and resulted in over 80% accuracy in detecting spam emails, further reinforcing the effectiveness of SVM as a classifier for this purpose.

E) The proliferation of a large quantity of unsolicited emails presents a threat to user security. Despite the presence of diverse security protocols, spammers introduce significant weaknesses to the internet. [5] This research delves into the efficient strategies for employing well-known algorithms to develop a machine learning model that can differentiate between spam and authentic emails. The UCI Machine Learning Repository Spam Base Data Set is utilized as the foundation for this investigation. The study evaluates the efficacy of five fundamental machine learning classification algorithms - Logistic Regression, Decision Tree, Naïve Bayes, KNN, and SVM - in creating a robust model for identifying email spam. The training and testing of the dataset are carried out using the Weka tool. The results indicate that the random tree algorithm surpasses other classification approaches in all performance metrics. Although KNN produces comparable outcomes, it necessitates more time for model development than the random tree. Subsequent research will concentrate on enhancing the model by integrating additional assessment parameters to enhance spam detection effectiveness.

F) In [6], Researchers compared different advanced models to see how well they could identify spam emails. The BERT base-cased transformer model performed best, with an accuracy of nearly 98.7% and an F1 score close to that value. This success likely stems from the model's use of attention layers, which allow it to fully grasp the context of the email text. BERT's contextual word embedding’s also significantly improved spam detection compared to simpler methods like Keras word embedding’s, which assign a unique number to each word. The BiLSTM model, which uses Keras embedding’s, achieved an accuracy of around 96.4%. Testing the models on unseen data confirmed their effectiveness and generalizability. To potentially improve future results, researchers suggest extending the maximum email length considered (currently limited to 300 words due to hardware constraints) and exploring spam detection in other languages like Arabic.

G) This paper [7] explores the concept of leveraging network-level characteristics to identify spam emails. The argument centers around the limitations of content-based filtering, citing its computational cost and susceptibility to manipulation by spammers. The authors propose a multi-layered filtering system. The first stage utilizes DNS blacklists to weed out known spam sources. Subsequent stages employ filters built upon features extracted from the initial email packet and network data associated with the email transfer. Content-based filtering serves as the final stage. This design aims to eliminate blatant spam emails at each level, minimizing the burden on the content filter. The study reveals that network-level features can substantially decrease spam filtering processing time without compromising accuracy. However, the effectiveness of these features can diminish over time due to network-related events. The research demonstrates that combining network-level filtering with DNS blacklists offers a robust spam detection mechanism, even during network disruptions. Overall, this work highlights the potential of network-level features in enhancing both the efficiency and effectiveness of spam filtering.

H) This research explores the potential of feed forward neural networks (FFNNs) for spam email detection. Spam emails pose a significant problem, consuming bandwidth and potentially tricking users into revealing personal information. [8] While traditional methods may struggle, machine learning offers a promising solution. The authors propose using an FFNN to classify emails as spam or legitimate. They train their FFNN on a well-known email dataset, the Enron corpus, and compare its performance to another machine learning model, BERT. Their findings suggest that the FFNN achieves good results, with the number of layers and neurons impacting its effectiveness. Future work aims to evaluate the FFNN against other machine learning models.

I) This paper [9] provides an extensive evaluation of the most efficient content-based methods for filtering spam emails. Our focus is predominantly on spam filters powered by Machine Learning and their variations, while also discussing a wide range of topics including relevant concepts, initiatives, efficacy, and current advancements. The initial discussion delves into the fundamentals of email spam filtering, the changing landscape of spam, the ongoing battle between spammers and email service providers (ESPs), as well as the role of Machine Learning in combating spam. Our analysis concludes by assessing the effectiveness of Machine Learning-driven filters and investigating the promising developments stemming from recent innovations. Future research should focus on the co-evolutionary nature of e-mail spam filtering, where the filter's accuracy is challenged by spammers evolving their tactics. Effective strategies need to detect changes in spam characteristics, with content-based filtering showing significant success. Machine learning systems allow for adaptation to new spam threats, but a multi-faceted approach combining legal and technical measures may be necessary for a comprehensive solution. Despite advancements, challenges in e-mail spam filtering persist, making further research crucial for improving anti-spam measures.

J) Email, [10] a vital communication tool, is plagued by spam. These unsolicited emails not only waste time but also pose security risks. Automatic email filtering is the primary defense, with content-based filtering being a key technique. This approach analyses email content, like word frequency and usage patterns, to identify spam. Machine learning algorithms power this analysis, flagging emails with suspicious characteristics. This paper explores a methodology for improving spam classification accuracy. The methodology involves:

- Experimental Evaluation: Testing various filtering techniques on different emails.
- Parameter Tuning: Fine-tuning the algorithms for better performance.
- Test Train Splitting: Dividing email data for training and evaluating the algorithms.

The study emphasizes the importance of data pre-processing, which involves cleaning and organizing the data before analysis. This improves filtering accuracy. Two algorithms, Support Vector Machine (SVM) and Naive Bayes, were tested. SVM outperformed Naive Bayes in accuracy, achieving 99.01% compared to 97.67%. The confusion matrix also confirmed SVM's effectiveness in classifying spam emails. This research highlights the potential of machine learning for highly accurate spam filtering. Future work could involve incorporating data from multiple sources and exploring optimization techniques for faster filtering.

K) The paper [11] offers a meticulous examination of the evolving landscape of spam detection, focusing on AI and machine learning methodologies. Here's a succinct review: The paper delves into the intricacies of spam email detection, dissecting emails into distinct parts such as headers, SMTP envelope, and content. It meticulously evaluates non-AI based anti-spam systems like cryptographic hashing, fuzzy hashing, DCC, grey listing, and DNS blacklisting/white listing, highlighting their strengths and limitations in the face of increasingly sophisticated spam tactics. Emphasizing the dominance of supervised machine learning in spam detection, the paper showcases algorithms like Support Vector Machines and Naive Bayes as primary tools. It outlines potential areas for future research, advocating for hybrid and multi-algorithm systems to enhance detection robustness. Additionally, it proposes advanced email header analysis and strategies to combat concept drift and phishing emails. The paper advocates for stronger anti-spam regulations as part of the future framework, recognizing the need for comprehensive measures beyond technological advancements alone.

L) The [12] battle against spam emails has reached a new frontier with the introduction of Ensemble-based Lifelong Classification with Adjustable Dataset Partitioning (ELCADP). In the face of constantly evolving spammer tactics, traditional classification models grapple with the challenge of concept drift, where the characteristics of spam emails undergo continuous transformation. ELCADP addresses this challenge head-on by harnessing the power of ensemble learning and an innovative Early Drift Detection Module (EDDM) to dynamically adapt to shifting data distributions. By partitioning incoming data streams and incorporating new classifiers while retaining previously learned knowledge, ELCADP effectively mitigates concept drift and catastrophic forgetting. The efficacy of ELCADP is demonstrated through rigorous evaluations using the Enron-Spam dataset, showcasing its superiority over various stream mining methods in terms of accuracy, precision, recall, and F1-score. This breakthrough not only highlights the limitations of traditional approaches but also positions ensemble-based methods as robust solutions for lifelong spam email classification. The implications of ELCADP extend beyond spam filtering, emphasizing the importance of adaptive machine learning algorithms in addressing evolving threats in the digital realm. As such, this paper not only marks a significant advancement in spam classification but also sets the stage for further research in adaptive learning methodologies to safeguard against emerging cyber security challenges.

M) The [13] provided context is a research article titled "Machine Learning-Based Detection of Spam Emails" written by Zeeshan Bin Siddique, Mudassar Ali Khan, Ikram Ud Din, Ahmad Almogren, Irfan Mohiuddin, and Shah Nazir. The article discusses the use of machine learning algorithms to detect spam emails, specifically focusing on the Urdu language. The authors gathered a dataset of 5573 emails from online sources and used it to train seven machine learning models. The highest accuracy achieved was 98.5% using Multinomial Naive Bayes. However, the authors note that Naive Bayes has limitations due to class-conditional dependency. They suggest using ensemble approaches, which use multiple learners to predict categories, as a potential solution. The authors also compare their results to previous work in the field of spam email detection. They note that there is notable work done in roman Urdu script, but it is different from Urdu writing script. They write Roman Urdu using English alphabets, but Urdu script is adapted from Persian language and is based on Arabic alphabets. The authors also encountered only one approach in their literature review that used Urdu script for spam email detection, and it had lower results than their approach. Overall, the article presents a machine learning-based approach to detecting spam emails in the Urdu language and compares the results to previous work in the field. The authors suggest using ensemble approaches as a potential solution to the limitations of Naive Bayes.

N) The paper [14] explores a recent study that utilizes the Harris Hawks Optimizer (HHO) algorithm for spam classification in emails. The authors aimed to develop a method effective for high-dimensional data while achieving superior spam detection accuracy compared to existing techniques. The paper is well-structured, providing a clear introduction to email spam and the rationale behind using HHO. A comprehensive literature review explores current content-based spam detection methods and supervised machine learning algorithms employed for email classification. The core concepts of the HHO algorithm and k-Nearest Neighbours (k-NN) theory are explained, which are fundamental to understanding the proposed approach. The study introduces a novel HHO-KNN algorithm, detailing the pre-processing stage and feature selection processes that leverage both HHO and k-NN. The experimental setup and results are presented transparently, comparing the proposed method's performance against other algorithms including Binary Dragonfly, Equilibrium Optimizer, Teaching-Learning-based Optimization, Seagull Optimization Algorithm, and Marine Predators Algorithm. The findings indicate that the HHO-KNN approach outperforms these algorithms in terms of classification accuracy, achieving a significant rate of 94.3% on the Spam base dataset. Additionally, convergence curves demonstrate the HHO-KNN algorithm's superior performance during initial iterations. However, there are areas for potential improvement. The authors could address limitations associated with their approach, such as the computational complexity of the HHO algorithm and its generalizability to other datasets. Furthermore, incorporating comparisons with more advanced spam detection techniques would strengthen the paper's overall contribution. In conclusion, the study presented in "Detecting Spam Email with Machine Learning Optimized with Harris Hawks optimizer

(HHO) Algorithm" proposes a novel spam classification technique that integrates the HHO algorithm.

O) This study [15] explored how machine learning can sort spam emails. The researchers tested J48 decision tree, SVM, ANN, and Naive Bayes on a dataset from a trusted public repository (UCI machine learning repository). SVM proved most adept at this task. The study delved into the consequences of spam and methods to curb it. They recommend SVM as a powerful spam filter due to its high accuracy and low misidentification rate. Future studies could focus on applying these techniques to unearth spam within social media platforms. The paper provides a well-structured examination of email spam and the various tools used to tackle it. Utilizing a reputable dataset and referencing other research strengthens the study's validity. Table 1 shows the comparison of algorithms used for detection of spamming.

Table 1 Data Comparison of Algorithms used for detection of spamming.

ALGORITHMS	FEATURES	EFFICIENCY
Support vector machine (SVM)	The primary classifier in the experiment is used to separate data points of different classes using the hyper plane with the maximum margin.	90.7629%
K-Nearest Neighbour (K-NN)	This approach categorizes data points by assessing their resemblance to the k nearest neighbours within the training dataset.	99.9348%
Naïve Bayes	This classifier operates by assessing the likelihood of occurrences and is well-suited for predicting spam, although it may not be optimal for non-spam emails.	79.5262%
Harris Hawks optimizer (HHO) Algorithm	In email spam detection, HHO can optimize machine learning model parameters or create better spam detection algorithms by imitating Harris hawks' collaborative hunting strategy to reach an optimal solution.	HHO algorithm is relatively new, there might not be extensive literature or case studies available specifically on its performance in this domain.
Random Forest (RF)	An ensemble classifier integrates several decision tree models to enhance performance.	99.9348%

Conclusion

In this research paper, we have examined the crucial issue of detecting email spam and investigated diverse strategies and methods to address this prevalent concern. The utilization of email as a primary means of correspondence encounters continuous obstacles from unwanted messages flooding mailboxes, presenting risks like phishing, dissemination of malware, and financial deception. Our study highlights the significance of integrating advanced technologies, specifically machine learning and artificial intelligence, to improve the efficiency of spam detection systems. Various methodologies have been scrutinized, such as Bidirectional Long Short-Term Memory with Convolutional Neural Networks, Support Vector Machines, and deep learning techniques, demonstrating their effectiveness in distinguishing between legitimate and spam emails. Our investigation emphasizes the importance of precision and accuracy in algorithms for spam detection, as these aspects directly influence the dependability of email filtration systems. By thoroughly evaluating the performance of different machine learning algorithms, we endeavor to pinpoint the most optimal solution for detecting email spam, with the objective of reducing false positives and false negatives. Furthermore, our exploration transcends traditional methods, as we have delved into innovative frameworks like header-based detection mechanisms and analysis of network characteristics for identifying spam. These strategies present promising opportunities for enhancing email security and countering sophisticated spamming tactics. As we wrap up this analysis, it becomes apparent that combating email spam necessitates a multifaceted strategy that integrates technological advancements with robust algorithmic structures. While machine learning and AI offer formidable tools in this pursuit, continuous adaptation and enhancement are essential to outpace evolving spamming tactics. Ultimately, our study contributes to the ongoing dialogue on email security by offering insights into effective spam detection strategies and paving the way for future progress in this field. By leveraging the capabilities of state-of-the-art technologies, we can strengthen email communication channels, ensuring a safer and more efficient digital environment for users worldwide.

References

[1]. N. Kumar, S. Sonowal, and Nishant, “Email Spam Detection Using Machine Learning Algorithms,” Jul. 2020,pp.108–113.doi: 10.1109/ICIRCA48905.2020.9183098.

[2]. S. Rahman and S. Ullah, “Email Spam Detection using Bidirectional Long Short Term Memory with Convolutional Neural Network,” Jan. 2020, pp. 1307–1311.doi: 10.1109/TENSYMP50017.2020.9230769.

[3]. M. K. Sen and P. Richhariya, “DETECTING SPAM BY APPLYING MACHINE LEARNING APPROACH OVER EMAIL,” vol. 03, no. 12, 2023.

[4]. S. Khamis, C. F. Mohd Foozy, M. Aziz, and N. Rahim, “Header Based Email Spam Detection Framework Using Support Vector Machine (SVM) Technique,” 2020, pp. 57–65. doi: 10.1007/978-3-030-36056-6_6.

[5]. N. Sundaresan and J. Marseline, “Performance Evaluation of Machine Learning Algorithms for Email Spam Detection,” Feb. 2020, pp. 1–4. doi: 10.1109/ic-ETITE47903.2020.312.

[6]. I. AbdulNabi and Q. Yaseen, “Spam Email Detection Using Deep Learning Techniques,” Procedia Computer Science, vol. 184, pp. 853–858, Jan. 2021, doi: 10.1016/j.procs.2021.03.107.

[7]. T. Ouyang, S. Ray, M. Allman, and M. Rabinovich, “A large-scale empirical analysis of email spam detection through network characteristics in a stand-alone enterprise,” Computer Networks, vol. 59, pp. 101–121, Feb. 2014, doi: 10.1016/j.comnet.2013.08.031.

[8]. S. Kaddoura, O. Alfandi, and N. Dahmani, “A Spam Email Detection Mechanism for English Language Text Emails Using Deep Learning Approach,” in 2020 IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Sep. 2020, pp. 193–198. doi: 10.1109/WETICE49692.2020.00045.

[9]. A. Bhowmick and S. M. Hazarika, “Machine Learning for E-mail Spam Filtering: Review,Techniques and Trends.” arXiv, Jun. 03, 2016. Accessed: Apr. 21, 2024. [Online]. Available: http://arxiv.org/abs/1606.01042

[10]. T. Mehrotra, G. K. Rajput, M. Verma, B. Lakhani, and N. Singh, "Email Spam Filtering Technique from Various Perspectives Using Machine Learning Algorithms," in Data Driven Approach Towards Disruptive Technologies, T. P. Singh, R. Tomar, T. Choudhury, T. Perumal, and H. F. Mahdi, Eds., Singapore: Springer, 2021, pp. 423–432. doi: 10.1007/978-981-15-9873-9_33.

[11]. A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, "A Comprehensive Survey for Intelligent Spam Email Detection," IEEE Access, vol. 7, pp. 168261–168295, 2019, doi: 10.1109/ACCESS.2019.2954791.

[12]. R. Mohammad and A. Mohammad, "A lifelong spam emails classification model," Applied Computing and Informatics, vol. ahead-of-print, Jan. 2020, doi: 10.1016/j.aci.2020.01.002.

[13]. Z. B. Siddique, M. A. Khan, I. U. Din, A. Almogren, I. Mohiuddin, and S. Nazir, "Machine Learning-Based Detection of Spam Emails," Scientific Programming, vol. 2021, p. e6508784, Dec. 2021, doi: 10.1155/2021/6508784.

[14]. A. S. Mashaleh, N. F. Binti Ibrahim, M. A. Al-Betar, H. M. J. Mustafa, and Q. M. Yaseen, "Detecting Spam Email with Machine Learning Optimized with Harris Hawks optimizer (HHO) Algorithm," Procedia Computer Science, vol. 201, pp. 659–664, Jan. 2022, doi: 10.1016/j.procs.2022.03.087.

[15]. M. Jazzar, R. Yousef, and D. Eleyan, "Evaluation of Machine Learning Techniques for Email Spam Classification," International Journal of Education and Management Engineering, vol. 11, pp. 35–42, Aug. 2021, doi: 10.5815/ijeme.2021.04.04.